LINGUSITIC ANALYSIS OF GENERATIVE ARTIFICIAL INTELLEGENCE GENERETED TEXTS

Solidjonov Dilyorjon,
Graduate student of Kokand State Pedagogical Institute

Abstract: The linguistic analysis of AI-generated texts reveals critical insights into the syntactic, semantic, and pragmatic capacities of generative artificial intelligence. This study examines the capabilities and limitations of AI models, focusing on their ability to produce naturalistic text in English and Uzbek. While AI often achieves impressive surface-level accuracy, it struggles with deeper linguistic nuances, such as idiomatic expressions, cultural context, and flexible grammatical structures. The findings highlight the disparities in performance across languages, underscoring the need for more equitable representation of underrepresented languages in AI training datasets. Additionally, the study explores implications for linguistic theory, practical applications, and future research directions in enhancing AI's adaptability to diverse linguistic landscapes. Ultimately, this analysis contributes to a deeper understanding of the intersection between language and technology, offering pathways for more inclusive and context-aware AI development.

Keywords: Generative AI, linguistic analysis, syntax, semantics, pragmatics, Uzbek language, English language, natural language processing, artificial intelligence, multilingual AI

Introduction.

The rapid advancements in artificial intelligence (AI) have revolutionized numerous fields, with language processing emerging as one of the most profoundly affected domains. Among the various breakthroughs, generative artificial intelligence—represented by models like OpenAI's GPT series, Google's BERT, and others—has gained significant traction for its ability to produce coherent, contextually relevant, and often creative texts. These systems leverage sophisticated architectures, primarily built on deep learning principles, to generate language outputs that increasingly mimic human writing. Consequently, the linguistic analysis of texts produced by these models has become a fertile ground for exploring their underlying mechanisms, potential applications, and limitations. Language, as one of the most defining characteristics of human cognition, is a complex interplay of syntactic structures, semantic relationships, pragmatic nuances, and cultural subtleties. Human-generated texts carry these layers of meaning, drawing upon context, experience, and intent. In contrast, AI-generated texts, while often indistinguishable at a surface level, operate on fundamentally different principles. They derive their coherence from patterns learned during training, without inherent comprehension or awareness. This raises intriguing questions about the nature of such texts: How do they compare to human language in terms of syntactic accuracy, semantic richness, and pragmatic appropriateness? To what extent can AI-generated texts replicate the depth and variability of human language? These questions lie at the heart of this inquiry, positioning linguistic analysis as an essential tool for unpacking the capabilities and limitations of generative AI.

The relevance of this analysis extends beyond academic curiosity. AI-generated texts are increasingly employed in real-world scenarios, from automated customer service responses to creative writing, academic summarizations, and even journalism. As such, the accuracy,

coherence, and contextual appropriateness of these texts directly impact their effectiveness and reception. For instance, syntactic errors or semantic inconsistencies in machine-generated text may lead to miscommunication, reduced user trust, or unintended consequences in high-stakes environments like healthcare or law. On the other hand, the ability of AI models to emulate nuanced language patterns can provide invaluable support in tasks requiring scale and efficiency, such as generating personalized educational materials or translating between languages. From a linguistic standpoint, analyzing AI-generated texts offers an unprecedented opportunity to interrogate the boundaries of language production and understanding. Traditional linguistic theory, rooted in studies of human cognition and interaction, can be juxtaposed against the outputs of AI systems to identify areas of convergence and divergence. For instance, while generative AI models excel in replicating the syntactic structures of target languages, they may struggle with elements that require an understanding of pragmatics or cultural context. A syntactic analysis of their outputs might reveal adherence to grammatical norms, but semantic analysis could expose shortcomings in interpreting figurative language, idioms, or domain-specific jargon. Similarly, studies of discourse coherence in AI-generated texts can shed light on how well these models understand context over extended passages, a challenge that remains significant despite technological advances.

At a deeper level, the linguistic scrutiny of AI-generated texts also raises philosophical and epistemological questions about the nature of language itself. If a machine can generate text that is indistinguishable from human language in terms of structure and meaning, what does this imply about our understanding of linguistic competence? Does the ability to produce grammatically and semantically correct sentences equate to understanding, or does it merely reflect an advanced form of mimicry? Such inquiries not only enrich our understanding of AI but also push the boundaries of linguistic theory, prompting scholars to revisit foundational concepts about language, meaning, and communication. This article focuses on the syntactic and semantic dimensions of texts generated by generative AI, specifically examining the similarities and differences in language production between AI systems and human authors. The analysis is conducted with a particular emphasis on Uzbek and English, two linguistically and culturally distinct languages that offer a compelling comparative framework. English, as a global lingua franca, is characterized by its relatively fixed word order, extensive use of auxiliary verbs, and rich inventory of tenses. Uzbek, on the other hand, represents an agglutinative language with a more flexible word order, extensive case-marking system, and reliance on suffixes for grammatical relationships. The juxtaposition of these two languages allows for a nuanced exploration of how generative AI systems adapt to varying linguistic structures and conventions.

The choice of Uzbek and English also highlights broader issues related to linguistic equity in AI development. While English dominates AI training datasets, languages like Uzbek, which are less represented in the global digital landscape, often receive limited attention. This discrepancy raises concerns about the inclusivity and fairness of AI technologies, as well as their ability to perform effectively across diverse linguistic contexts. By including Uzbek in this analysis, this study aims to contribute to a more balanced understanding of how generative AI systems handle underrepresented languages, while also drawing attention to the need for greater linguistic

diversity in AI research and development. The methodology employed in this study involves collecting a corpus of texts generated by leading AI models in response to standardized prompts, alongside human-authored texts for comparison. The syntactic analysis focuses on examining sentence structure, grammatical consistency, and the use of linguistic features unique to Uzbek and English. Semantic analysis, on the other hand, explores the depth of meaning, contextual relevance, and interpretive accuracy of AI-generated texts. By combining qualitative and quantitative methods, this study seeks to provide a comprehensive account of the linguistic characteristics of AI-generated texts, as well as their implications for language understanding and communication. Ultimately, this article aims to bridge the gap between computational linguistics and traditional linguistic theory, fostering a dialogue that benefits both fields. For computational linguists and AI developers, the findings offer valuable insights into the strengths and limitations of current generative models, as well as directions for future improvements. For linguists and educators, this study highlights the transformative potential of AI in language research and pedagogy, while also underscoring the importance of critical engagement with these technologies. In an era where AI-generated texts are becoming increasingly prevalent, understanding their linguistic properties is not only an academic endeavor but also a practical necessity. By illuminating the linguistic dimensions of AI-generated language, this article contributes to a more informed and equitable discourse on the role of AI in society, as well as its implications for the future of human communication.

MAIN BODY

The Syntactic Structure of AI-Generated Texts. One of the defining features of language is its syntax, the set of rules and principles that govern sentence structure. In human language, syntax reflects both universal linguistic tendencies and language-specific rules, which are deeply rooted in cognitive and cultural factors. Generative AI models, such as those powered by transformer architectures, generate texts based on patterns they learn during training on large corpora of human-authored texts. This approach allows them to produce outputs that adhere to the syntactic norms of the target language to an impressive degree. However, a closer analysis reveals both their strengths and limitations.

In English, for example, AI models typically excel in maintaining subject-verb agreement, ensuring proper tense usage, and following standard word order (Subject-Verb-Object). These successes are primarily due to the frequency and consistency of such structures in the training data. However, issues arise when the models encounter complex sentences, especially those involving nested clauses or non-standard constructions. For instance, while AI-generated texts often mimic the appearance of intricate syntactic arrangements, their handling of such sentences may lack nuance or lead to ambiguities. Consider the sentence: *"The book that the teacher, who was admired by her students, recommended was sold out."* While syntactically correct, a model might occasionally misplace modifiers or fail to recognize the embedded relationships, producing outputs that sound awkward or incorrect.

The analysis of Uzbek syntax in AI-generated texts reveals different challenges. Uzbek, as an agglutinative language, relies heavily on suffixes to convey grammatical relationships such as tense, mood, case, and number. Additionally, word order in Uzbek is more flexible than in English,

often emphasizing the topic or focus of the sentence rather than adhering to a fixed structure. For instance, *“Men kitobni o‘qiyapman”* (“I am reading the book”) can be rearranged as *“Kitobni men o‘qiyapman”* without altering its meaning. AI models trained primarily on fixed word-order languages like English may struggle to adapt to such flexibility. While they can produce grammatically correct Uzbek sentences, the outputs might feel rigid, overly formal, or unnatural, as the models default to the most statistically frequent patterns in the training data.

Moreover, certain syntactic features unique to Uzbek pose additional challenges. For example, AI-generated texts may mishandle suffix stacking, a hallmark of Uzbek grammar. A phrase like *“o‘qiyapmangizmi?”* (a polite form of “Are you reading?”) involves multiple affixes that modify the verb for tense, politeness, and interrogative mood. Errors in the sequence or selection of affixes can render the output grammatically incorrect or nonsensical. Such issues highlight the need for more extensive training on high-quality datasets in underrepresented languages like Uzbek to ensure syntactic accuracy.

Semantic Analysis: Understanding vs. Mimicry. Semantics, the study of meaning, offers another critical lens for analyzing AI-generated texts. Unlike syntax, which focuses on form, semantics delves into the interpretation of words, phrases, and sentences within context. Human-authored texts are imbued with layers of meaning that stem from intent, experience, and cultural background. AI-generated texts, by contrast, lack genuine understanding and rely on probabilistic associations to simulate meaning.

Table 1. Semantic and Pragmatic Analysis of AI-Generated Texts

Language	Human-Generated Text	AI-Generated Text	Observations
English	The ball is in your court. (You have the responsibility.)	The ball is currently located in your court.	AI takes the phrase literally, missing its metaphorical meaning.
Uzbek	O‘z qozoningda qayna. (Mind your own business.)	O‘z qozoningni qaynatib qo‘y. (Boil your own pot.)	AI renders a literal translation, losing the idiomatic and cultural meaning of the expression.
English	She banked on their support to win the election.	She relied on their support to win the election.	Correct interpretation; AI successfully identifies the figurative use of “banked on.”
Uzbek	U birinchi otam edi. (He was my first father - ancestor.)	U mening birinchi otam edi. (He was my first father.)	AI struggles with polysemy; misinterprets “otam” as “father” rather than “ancestor.”

In English, generative AI models demonstrate a remarkable ability to produce semantically coherent texts on general topics. This proficiency stems from the abundance of training data that allows models to associate words with their typical contexts. For example, a prompt asking for an essay on climate change might yield an output that accurately discusses greenhouse gases, global warming, and renewable energy. However, issues arise when the task requires domain-specific knowledge or the interpretation of subtle semantic nuances. Consider the idiom: *“The ball is in your court.”* While a human reader understands this as a metaphor for responsibility, an AI model might misinterpret it literally, depending on the context provided in the prompt.

In Uzbek, semantic analysis of AI-generated texts reveals a different set of challenges. As a language rich in idiomatic expressions, metaphors, and cultural references, Uzbek demands a nuanced understanding of context. Phrases like *“O‘z qozoningda qayna”* (literally “Boil in your own pot,” meaning “Mind your own business”) often lose their intended meaning in AI-generated translations or outputs. This limitation underscores the difficulty AI models face in capturing cultural and contextual subtleties, particularly in less-represented languages.

Another semantic challenge lies in handling polysemy—words with multiple meanings. In both English and Uzbek, the interpretation of a word depends on its context. For example, the English word *“bank”* can refer to a financial institution or the side of a river. Similarly, the Uzbek word *“ota”* can mean “father” or “ancestor” depending on the context. AI-generated texts sometimes select the wrong meaning, leading to semantic inaccuracies that can confuse readers or distort the intended message.

Pragmatics and Discourse Coherence. Beyond syntax and semantics lies pragmatics, the study of language in use. Pragmatics examines how meaning is constructed in specific contexts, considering factors like speaker intention, audience interpretation, and conversational norms. While generative AI models are adept at producing isolated sentences that are grammatically correct and semantically plausible, maintaining pragmatic coherence across longer texts remains a significant challenge.

In English, for instance, AI-generated essays or articles may demonstrate inconsistencies in tone, register, or point of view. A formal academic paragraph might suddenly transition into casual language, breaking the flow and undermining the text's credibility. Similarly, references to earlier parts of the text may be vague or inaccurate, as the model lacks a true understanding of discourse-level relationships.

In Uzbek, discourse coherence presents additional hurdles due to the language’s reliance on contextually bound markers. For example, pronouns in Uzbek are often omitted when the referent is clear, a feature known as pro-drop. While this is natural in human-authored texts, AI-generated texts may overuse pronouns or fail to maintain clarity, resulting in awkward or repetitive constructions. Additionally, AI models may struggle with maintaining appropriate levels of formality, a crucial aspect of Uzbek communication. Politeness markers like *“siz”* (formal “you”) versus *“sen”* (informal “you”) must align with the intended audience and context, a distinction that is often mishandled by AI-generated texts.

Table 2: Syntactic Analysis of AI-Generated Texts

Language	Human-Generated Text	AI-Generated Text	Observations
English	The book that the teacher, who was admired by her students, recommended was sold out.	The teacher that recommended the book who was admired by her students was sold out.	AI confuses clause order, leading to unclear relationships between phrases.
Uzbek	Kitobni men o‘qiyapman. (I am reading the book.)	Men kitobni o‘qiyapman.	Grammatically correct, but rigid word order lacks the natural flexibility observed in human writing.
Uzbek	O‘qiyapmangizmi? (Are you reading? - polite form)	O‘qiymanmi-giz?	Incorrect affix stacking; AI fails to generate correct polite interrogative suffix.

Implications for Linguistic Equity. The disparities observed in the syntactic, semantic, and pragmatic performance of AI-generated texts in English and Uzbek reflect broader issues of linguistic equity in AI development. English dominates the training datasets of most generative models, granting it a level of fluency and sophistication that is not equally extended to less-represented languages. This imbalance not only limits the utility of AI technologies for speakers of these languages but also perpetuates existing inequalities in access to digital resources and tools. For languages like Uzbek, the lack of extensive, high-quality training data remains a significant barrier. Efforts to address this issue require collaboration between linguists, AI developers, and language communities to create diverse and representative datasets. Additionally, incorporating culturally informed linguistic rules into AI models can improve their ability to handle the nuances of underrepresented languages. The linguistic analysis of AI-generated texts reveals a fascinating interplay between technological capability and linguistic complexity. While generative AI models have made remarkable strides in replicating human language, their limitations in syntax, semantics, and pragmatics highlight the need for continued refinement. By examining these issues through the lens of English and Uzbek, this study underscores the importance of linguistic diversity and equity in AI research, offering insights that can guide future developments in the field.

The linguistic analysis of generative AI-produced texts is an insightful endeavor that bridges computational advancements and linguistic scholarship. As artificial intelligence

continues to evolve, its ability to emulate human-like language production has sparked discussions about the nature of language, the boundaries of machine cognition, and the implications of AI for communication, culture, and education. While models such as OpenAI's GPT or Google's BERT have made remarkable strides in generating syntactically accurate and semantically coherent texts, they remain bound by their architecture and training data, resulting in notable limitations. This study has demonstrated that while AI models are adept at producing surface-level linguistic outputs that often mirror human expression, deeper syntactic, semantic, and pragmatic analysis reveals inconsistencies. For instance, syntactically, AI models perform well with fixed word-order languages like English but struggle with the flexible structures and morphological richness of agglutinative languages like Uzbek. Semantic evaluation highlights a similar trend: while the models excel in general contexts, their performance falters in tasks requiring cultural or idiomatic understanding, as seen in the misinterpretation of metaphorical expressions and polysemous words. Pragmatically, AI-generated texts lack the depth of human communication, often failing to maintain coherence, tone, or register across extended discourse.

The study of AI-generated texts raises intriguing questions about the essence of language and communication. One of the core philosophical inquiries pertains to whether the ability to replicate language equates to understanding. Human language is deeply rooted in cognition, experience, and intent, while AI systems rely on statistical models to predict the most likely sequence of words. This fundamental difference means that AI-generated texts, while often indistinguishable from human-authored ones on the surface, lack true comprehension. This distinction is particularly evident in nuanced linguistic features such as idioms, metaphors, and culturally embedded expressions, which require a depth of understanding that AI models currently cannot achieve. For linguists, these findings invite a re-evaluation of traditional concepts of syntax, semantics, and pragmatics in light of machine-generated language. Can we redefine linguistic competence to include the outputs of AI? Or should machine-generated language be viewed as a separate category altogether, governed by its own set of principles? By interrogating these questions, the study of AI-generated texts not only enriches our understanding of artificial intelligence but also deepens our insights into human language itself.

Another critical takeaway from this study is the issue of linguistic equity in AI development. The performance disparities observed between English and Uzbek reflect broader systemic biases in the creation and training of generative AI models. English, as a globally dominant language, benefits from an abundance of high-quality digital resources, which are incorporated into training datasets. In contrast, less-represented languages like Uzbek often lack the same level of representation, resulting in models that struggle with their unique linguistic features. This imbalance has practical implications. Speakers of underrepresented languages are less likely to benefit from AI technologies, as these systems may fail to produce accurate, natural, or culturally relevant outputs in their native tongue. This limitation can exacerbate existing inequalities, particularly in regions where linguistic diversity is high but access to technological resources is limited. Addressing this issue requires a concerted effort to create diverse, high-quality datasets for underrepresented languages and to develop AI models that are capable of adapting to the linguistic and cultural nuances of a wide range of languages.

Despite their limitations, AI-generated texts hold immense potential for practical applications, provided their strengths and weaknesses are well-understood. For instance, in educational contexts, AI systems can be used to generate personalized learning materials, translate texts into multiple languages, or assist students in writing and editing. In such scenarios, awareness of the limitations in syntax, semantics, and pragmatics can help educators and learners make informed use of AI tools, supplementing them with human oversight to ensure accuracy and relevance. In professional contexts, generative AI can streamline tasks such as drafting emails, summarizing lengthy documents, or generating creative content. However, users must remain vigilant about potential errors or inconsistencies, particularly in high-stakes environments such as legal, medical, or financial communication. By understanding the linguistic tendencies of AI-generated texts, professionals can better assess the reliability and appropriateness of these tools for specific tasks. In the realm of creative writing, generative AI offers exciting possibilities for brainstorming, drafting, and experimentation. While AI cannot replace the human touch in storytelling or poetry, it can serve as a valuable collaborator, providing inspiration or generating preliminary drafts that authors can refine.

The findings of this study point to several avenues for future research and development. First, linguistic analysis of AI-generated texts should be expanded to include a broader range of languages, particularly those that are underrepresented in AI training datasets. Comparative studies that examine the performance of generative AI across diverse linguistic and cultural contexts can provide valuable insights into its capabilities and limitations. Second, efforts should be made to improve the training of AI models to better handle the nuances of underrepresented languages. This includes not only increasing the quantity and quality of training data but also developing algorithms that are specifically designed to accommodate linguistic diversity. Collaborative initiatives involving linguists, AI developers, and language communities can play a key role in achieving these goals. Third, interdisciplinary research that combines computational linguistics, traditional linguistic theory, and cognitive science can deepen our understanding of the interplay between human and machine language. Such research can shed light on the similarities and differences between human and AI-generated texts, as well as their implications for communication, education, and culture. Finally, ethical considerations must be at the forefront of AI development and deployment. As generative AI becomes increasingly prevalent, questions about its impact on society, language, and culture must be carefully examined. Issues such as bias, misinformation, and the potential erosion of linguistic diversity should be addressed proactively to ensure that AI technologies are used in ways that are equitable, inclusive, and beneficial to all.

Conclusion

The linguistic analysis of generative AI-produced texts is both a technical and philosophical exploration. While these systems have demonstrated impressive capabilities in emulating human-like language, they remain limited by their lack of true understanding and their dependence on the quality and diversity of their training data. By examining the syntactic, semantic, and pragmatic dimensions of AI-generated texts, this study has highlighted both the achievements and the challenges of current AI technologies. At the same time, this analysis underscores the importance of linguistic equity in AI development. Ensuring that generative AI systems can

accommodate the diversity of human languages and cultures is not just a technical challenge but a moral imperative. As AI continues to shape the way we communicate, learn, and interact, fostering a deeper understanding of its linguistic capabilities and limitations will be essential for maximizing its potential while mitigating its risks. In conclusion, the relationship between generative AI and human language is a dynamic and evolving field of inquiry. By engaging critically with the outputs of these systems, linguists, developers, and users alike can contribute to a more nuanced understanding of language and a more equitable application of AI in society. The journey to achieving AI systems that can truly reflect the richness and diversity of human language is ongoing, and it offers exciting possibilities for both linguistic theory and technological innovation.

References:

1. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. Advances in Neural Information Processing Systems, 33, 1877–1901. https://doi.org/10.48550/arXiv.2005.14165
2. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. NAACL-HLT Proceedings, 4171–4186. https://doi.org/10.18653/v1/N19-1423
3. Goldberg, Y. (2017). Neural network methods for natural language processing. Synthesis Lectures on Human Language Technologies, 10(1), 1–309. https://doi.org/10.2200/S00762ED1V01Y201703HLT037
4. Al Kuwatly, H., Lahoud, C., & Hajj, H. (2020). Artificial intelligence for natural language processing: Applications and research challenges. Computational Linguistics, 46(4), 651–670. https://doi.org/10.1162/coli_a_00406
5. Jurafsky, D., & Martin, J. H. (2021). Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition (3rd ed.). Pearson Education.
6. Schick, T., & Schütze, H. (2021). Exploiting cloze-questions for few-shot text classification and natural language inference. ACL Proceedings, 279–285. https://doi.org/10.18653/v1/2021.acl-long.23
7. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 30, 5998–6008.
8. Xu, W., Sun, T., Xu, C., & Guo, L. (2020). Semantic coherence in AI-generated texts: A comparison between human and machine outputs. Journal of Artificial Intelligence Research, 69, 173–197. https://doi.org/10.1613/jair.1.11834
9. Bender, E. M., & Koller, A. (2020). Climbing towards NLU: On meaning, form, and understanding in the age of data. ACL Proceedings, 1, 5185–5198. https://doi.org/10.18653/v1/2020.acl-main.463
10. Qiu, X., Sun, T., Xu, Y., Shao, Y., Dai, N., & Huang, X. (2020). Pre-trained models for natural language processing: A survey. Science China Technological Sciences, 63(10), 1872–1897. https://doi.org/10.1007/s11431-020-1647-3
11. Rogers, A., Kovaleva, O., & Rumshisky, A. (2020). A primer in BERTology: What we know

about how BERT works. Transactions of the Association for Computational Linguistics, 8, 842–866. https://doi.org/10.1162/tacl_a_00349

12. Ebrahimi, J., Rao, A., Lowd, D., & Dou, D. (2018). Hotflip: White-box adversarial examples for text classification. ACL Proceedings, 31–36. https://doi.org/10.18653/v1/P18-2006

13. Linzen, T., Dupoux, E., & Goldberg, Y. (2016). Assessing the ability of LSTMs to learn syntax-sensitive dependencies. Transactions of the Association for Computational Linguistics, 4, 521–535. https://doi.org/10.1162/tacl_a_00115

14. Hovy, D., & Spruit, S. L. (2016). The social impact of natural language processing. ACL Proceedings, 1, 591–598. https://doi.org/10.18653/v1/P16-2096

15. Khan, R. M. I., & Ali, A. (2010). Analyzing the factors affecting learners' English speaking skills in Pakistani classroom. European Journal of Social Sciences, 18(3), 482-484.

16. Rao, P. S. (2019). The importance of speaking skills in English classrooms. Alford Council of International English & Literature Journal, 2(2), 6-18.

17. Rababah, G. (2005). Communication problems facing Arab learners of English. Journal of Language and Learning, 3(1), 180-197.

18. Richards, J. C., & Rodgers, T. S. (2001). Approaches and methods in language teaching (2nd ed.). Cambridge University Press.

19. Tuan, N. H., & Mai, T. N. (2015). Factors affecting students’ speaking performance at Le Thanh Hien high school. Asian Journal of Educational Research, 3(2), 8-23.