SCIENCE CONNECTS PEOPLE

INTERNATIONAL SCIENTIFIC JOURNAL OF

2026 / No. 1(2)

Contemporary Pedagogical and Philological Research

A CORPUS-BASED ANALYSIS OF UZBEK EFL STUDENTS' WRITING BEFORE AND AFTER AI-ASSISTED LEARNING

Narzullayeva Dilshoda Oybek qizi

Senior student, SamSIFL, Samarkand

Abstract. The rapid development of artificial intelligence has introduced new opportunities for improving English as a Foreign Language (EFL) students' writing performance. AI-assisted learning tools, particularly ChatGPT, provide learners with opportunities to receive immediate feedback, improve grammatical accuracy, expand vocabulary, and develop more coherent written texts. However, the extent to which AI-assisted learning influences the linguistic characteristics of students' writing requires further investigation. The present study examines the potential changes in Uzbek EFL students' writing before and after AI-assisted learning from a corpus-based perspective. A qualitative, corpus-informed research design is adopted to explore changes in lexical diversity, grammatical complexity, syntactic structures, lexical bundles, cohesive devices, and academic vocabulary. The study draws on established principles of corpus linguistics and previous research on generative artificial intelligence in second language writing. Particular attention is given to the comparison of learner-generated texts before AI-assisted instruction and revised texts produced following guided interaction with AI tools. Previous research indicates that AI-supported writing activities may contribute to improvements in grammatical knowledge, lexical development, and students' willingness to write. At the same time, research has identified differences between independently produced student texts and AI-generated writing, demonstrating the importance of distinguishing genuine learner development from direct AI text production. The study proposes a systematic framework for analyzing linguistic changes through corpus-based measures and emphasizes the importance of learner authorship, responsible AI use, and critical evaluation of generated feedback. It concludes that corpus-based analysis can provide valuable insights into the linguistic

SCIENCE CONNECTS PEOPLE

development of Uzbek EFL students and help teachers evaluate the educational contribution of AI-assisted learning.

Keywords: corpus linguistics, Uzbek EFL students, artificial intelligence, AI-assisted learning, ChatGPT, lexical diversity, grammatical complexity, academic writing, learner corpus, writing development.

INTRODUCTION

Writing is one of the most complex skills in foreign language learning because it requires learners to coordinate grammatical knowledge, vocabulary, discourse organization, and communicative competence. Successful writing involves the ability to construct meaningful sentences, organize information logically, express ideas accurately, and maintain coherence throughout a text. For students learning English as a Foreign Language, developing these abilities can be challenging, particularly when opportunities for authentic language practice and individualized feedback are limited. In recent years, artificial intelligence has become an increasingly important component of language education. AI-powered applications offer new opportunities for students to receive language feedback, identify errors, improve vocabulary, and revise written assignments. Among these technologies, ChatGPT has attracted considerable attention because it allows learners to interact with an AI system through natural language instructions and receive immediate responses. The use of artificial intelligence in writing instruction has created opportunities for more flexible and individualized learning. Students can ask AI systems to explain grammatical structures, suggest alternative expressions, identify unclear sentences, and provide recommendations for improving paragraph organization. Such assistance may support learners throughout the writing process, from planning and drafting to editing and revision (Kohnke et al., 2023). However, the effectiveness of AI-assisted learning should not be evaluated only through students' perceptions or the apparent quality of their final texts. It is also necessary to examine whether their actual linguistic performance changes after

SCIENCE CONNECTS PEOPLE

receiving AI-supported instruction. A student may produce a more sophisticated essay with the assistance of ChatGPT without necessarily developing the independent language knowledge required to reproduce similar improvements in a new writing task. Corpus linguistics provides a useful methodological approach to investigating this issue. A corpus is a systematically collected body of authentic language data that can be examined to identify patterns in vocabulary, grammar, syntax, and discourse. Corpus-based research enables investigators to move beyond subjective impressions and analyze linguistic features through quantitative and qualitative methods (McEnery & Hardie, 2012). In language education, learner corpora are particularly valuable because they contain texts produced by language learners. By comparing texts written at different stages of learning, researchers can investigate changes in lexical choices, grammatical structures, sentence complexity, and discourse organization. Biber, Conrad, and Reppen (1998) explain that corpus-based analysis allows researchers to examine language patterns systematically and identify differences between texts and communicative contexts. This approach is relevant to EFL writing research because it provides measurable indicators of linguistic development. The combination of corpus linguistics and artificial intelligence offers additional opportunities for studying writing development. AI-assisted learning may influence the linguistic features of students' texts by introducing new vocabulary, correcting grammatical errors, suggesting alternative sentence structures, and providing examples of academic language. For instance, students who initially use simple vocabulary and repetitive sentence structures may encounter more varied lexical expressions and syntactic patterns through guided interaction with ChatGPT. If learners understand and subsequently use these features independently, such changes may indicate genuine language development. However, it is important to distinguish between improvements resulting from learning and changes resulting from direct AI rewriting. A text that has been extensively rewritten by an AI system may demonstrate improved grammatical accuracy and lexical sophistication

SCIENCE CONNECTS PEOPLE

without reflecting the student's actual writing ability. Mizumoto, Yasuda, and Tamura (2024) investigated linguistic differences between Japanese EFL students' essays and ChatGPT-generated texts. Their study examined 263 essays using automated linguistic analysis and machine learning techniques. The researchers identified differences between student-produced and AI-generated texts, demonstrating that the linguistic characteristics of AI-generated writing should not automatically be interpreted as evidence of learner development. Similarly, research by Lu (2025) examined the contribution of AI-generated corpus learning to grammatical structures, lexical bundles, and willingness to write among 80 EFL students. The study employed a repeated-measures quasi-experimental design and reported improvements in the experimental group's performance following AI-based corpus instruction. These findings indicate the potential value of combining corpus-informed learning with AI-supported language practice. Cheung and Crosthwaite (2025) also explored the integration of corpus linguistics and generative AI through the development of CorpusChat. Their pilot study demonstrated how corpus-informed AI tools could support university students' disciplinary writing practices and digital academic literacy. These studies provide an important foundation for investigating the relationship between AI-assisted learning and writing development. Nevertheless, the linguistic development of Uzbek EFL students requires particular attention because learners' first-language backgrounds, educational experiences, and levels of English proficiency may influence their writing patterns. Uzbek university students may experience difficulties with article usage, verb forms, prepositions, word order, academic vocabulary, collocations, and the organization of complex sentences. AI-assisted learning may provide additional opportunities to address these difficulties through individualized explanations and repeated writing practice. A corpus-based approach can help identify whether students' texts demonstrate changes in these linguistic features following AI-supported instruction. It can also reveal which aspects of writing improve more noticeably and

SCIENCE CONNECTS PEOPLE

which continue to require teacher support. The purpose of the present study is to examine the potential linguistic changes in Uzbek EFL students' writing before and after AI-assisted learning through corpus-based analysis. Particular attention is given to lexical diversity, lexical sophistication, grammatical accuracy, syntactic complexity, cohesive devices, and academic vocabulary. The study is guided by the following research questions: How can corpus-based methods be used to identify linguistic changes in Uzbek EFL students' writing before and after AI-assisted learning? Which lexical, grammatical, and syntactic features are particularly relevant when evaluating changes in students' writing? How can researchers distinguish genuine learner development from linguistic improvements introduced directly through AI-generated revisions?

The present study adopts a qualitative, corpus-informed research design to examine the potential relationship between AI-assisted learning and changes in Uzbek EFL students' written production. The methodological framework is informed by corpus linguistics, learner corpus research, and recent investigations into generative artificial intelligence in second language writing. Corpus-based analysis is appropriate for this research because it enables systematic examination of linguistic features across multiple texts. As McEnery and Hardie (2012) explain, corpus linguistics provides methodological tools for investigating language patterns through systematically collected and analyzed linguistic data. The proposed research framework involves two comparable collections of student writing. The first collection represents students' independently produced texts before AI-assisted learning, while the second represents their independently produced texts after a period of guided AI-supported instruction. To ensure comparability, both collections should contain texts written by the same students under similar conditions. Writing prompts should be equivalent in topic, complexity, and communicative purpose. Students should receive comparable time limits and writing instructions during both stages. The AI-assisted learning period may

involve structured activities using ChatGPT. These activities can include identifying grammatical problems, receiving explanations of vocabulary and collocations, examining sentence structures, improving paragraph coherence, and revising independently written drafts. An important methodological distinction should be maintained between AI-assisted student texts and fully AI-generated texts. The final assessment texts should be written independently without direct AI rewriting. This makes it possible to investigate whether learners have transferred the knowledge acquired during AI-assisted instruction into their own writing. The corpus analysis focuses on several dimensions of linguistic performance. Lexical diversity is examined to determine the variety of vocabulary used in student texts. Type-token ratio may provide an initial measure, although its results are influenced by text length. For more reliable comparisons, researchers may employ length-sensitive measures such as the measure of textual lexical diversity proposed by McCarthy and Jarvis (2010). Lexical sophistication is examined through the use of less frequent vocabulary, academic expressions, and appropriate lexical combinations. Particular attention is given to collocations and recurring multiword expressions because these features contribute to natural and effective academic writing. Grammatical accuracy is evaluated through the identification of grammatical errors involving articles, prepositions, verb forms, subject–verb agreement, and sentence structure. Errors can be examined in relation to the total number of words or clauses in each text. Syntactic complexity is considered through indicators such as mean sentence length, clause structures, and the use of subordinate clauses. These features may provide information about the complexity of students' sentence construction. Cohesion and discourse organization are examined through the use of transitions, reference expressions, and other cohesive devices. The analysis considers whether students establish clear relationships between ideas and organize information effectively. The interpretation of linguistic changes should combine quantitative corpus measures with qualitative examination of selected writing

SCIENCE CONNECTS PEOPLE

samples. This approach can help identify not only how frequently particular features occur but also whether they are used appropriately. The framework also incorporates ethical considerations. Students should be informed about the role of AI tools in the learning process, and their original writing should remain identifiable. AI-generated revisions should be documented separately from independently produced texts. Since the present manuscript develops a corpus-informed research framework and literature-based analysis rather than reporting an already completed Uzbek student corpus experiment, no original participant scores, corpus frequencies, or statistical significance values are claimed.

The analysis of previous corpus-based research and AI-assisted writing studies identifies several linguistic dimensions that are particularly important when examining changes in EFL students' writing before and after AI-assisted learning.

Lexical diversity represents the range of different vocabulary items used within a text. It is an important dimension of writing development because effective academic writing requires learners to select appropriate vocabulary rather than repeatedly relying on a limited number of familiar words. AI-assisted learning may support lexical development by introducing students to synonyms, academic expressions, collocations, and alternative ways of expressing ideas. For example, students may initially use general expressions such as "very important," "a lot of problems," or "good results." Through guided AI interaction, they may become familiar with alternatives such as "highly significant," "numerous challenges," and "positive outcomes." However, lexical diversity should not be interpreted as the simple appearance of more different words. Repetition may be necessary for maintaining coherence, and sophisticated vocabulary is not always appropriate McCarthy and Jarvis (2010) emphasize the importance of selecting appropriate measures when investigating lexical diversity. Consequently, comparisons between pre-instruction and post-instruction texts should consider text length and the limitations of individual lexical indices. The findings of Lu (2025)

SCIENCE CONNECTS PEOPLE

provide relevant evidence concerning AI-generated corpus learning and lexical bundle development. The study reported improvements in lexical bundle acquisition among EFL learners following AI-based instruction. This suggests that corpus-informed AI activities may help students become more familiar with recurring expressions and conventional combinations of words.

Grammatical accuracy is another important dimension of learner writing that can be examined through corpus-based methods. Uzbek EFL students may experience difficulties with English articles, prepositions, verb tense consistency, subject–verb agreement, and complex sentence construction. AI-assisted learning can provide explanations and corrections related to these features. When students understand the reasons behind corrections, they may become more aware of recurring grammatical problems. For example, an initial text may contain the sentence "The students was interested in learning English." Following guided feedback, learners may recognize the subject–verb agreement problem and produce "The students were interested in learning English." The important question is whether students can apply this knowledge in later writing tasks without AI assistance. Lu (2025) reported improvements in grammatical structure learning through AI-generated corpus instruction. This finding supports the relevance of investigating grammatical changes through repeated writing samples. Nevertheless, a reduction in grammatical errors within AI-revised texts should not automatically be interpreted as improved learner competence. Independent post-instruction writing is necessary to establish whether students have developed greater grammatical control.

Syntactic complexity concerns the range and sophistication of grammatical structures used in writing. Beginning and intermediate EFL learners may rely heavily on short, simple sentences. With increased language exposure and practice, they may develop the ability to construct more complex sentences using subordinate clauses, relative clauses, and varied sentence patterns. AI-assisted learning may support this

process by providing explanations and examples of alternative sentence structures. For instance, students may learn to combine two simple sentences into a more complex construction by using subordinating conjunctions or relative clauses.

However, syntactic complexity should not be considered an independent measure of writing quality. Longer sentences may contain grammatical problems or obscure the intended meaning. Therefore, syntactic analysis should be combined with grammatical accuracy and contextual evaluation. Mizumoto et al. (2024) demonstrate the importance of examining multiple linguistic features when comparing student writing with AI-generated texts. Their corpus-based findings indicate that linguistic differences extend beyond individual vocabulary choices and include broader characteristics of text production.

Academic vocabulary and lexical bundles are important resources in university-level writing. Lexical bundles are recurring sequences of words that perform conventional functions in academic discourse. Examples include "on the other hand," "the results of the," and "it is important to." Students who have limited knowledge of such expressions may experience difficulties organizing arguments and communicating academic information naturally. AI-assisted learning can expose students to examples of these expressions and explain their functions in different contexts. Lu (2025) investigated the relationship between AI-generated corpus learning and students' acquisition of lexical bundles. The findings indicated that AI-supported corpus instruction was associated with improvements in learners' knowledge of these recurring language patterns. Cheung and Crosthwaite (2025) also demonstrated the potential of combining generative AI with authentic academic corpora. Their CorpusChat project used corpus-informed chatbots to support discipline-specific academic writing development. These findings suggest that combining AI interaction with authentic language examples may provide a useful approach to developing students' academic vocabulary.

Cohesion and discourse organization are essential features of effective academic writing. Students need to establish clear relationships between sentences, paragraphs, and arguments. A text may contain grammatically correct sentences but remain difficult to understand if its ideas are not logically connected. AI-assisted learning may help students identify missing transitions, repetitive expressions, and unclear relationships between arguments. For example, learners may become more familiar with expressions such as "furthermore," "nevertheless," "in contrast," and "as a consequence." However, the appropriate use of discourse markers is more important than their frequency. Excessive use of connectors may make writing repetitive and unnatural. Corpus-based analysis can therefore examine both the frequency and contextual appropriateness of cohesive devices.

One of the most important findings from existing research concerns the distinction between student-authored writing and texts produced by generative AI. Mizumoto et al. (2024) compared Japanese EFL students' essays with ChatGPT-generated texts and identified linguistic differences between the two groups. Their findings demonstrate that AI-generated writing may display distinctive linguistic characteristics. This distinction is essential for the present research topic because improved linguistic features in an AI-revised essay do not necessarily indicate that the student has acquired the corresponding language knowledge. A valid comparison should therefore distinguish between original student drafts, AI-assisted revisions, and independent post-instruction writing. The first comparison can reveal what changes AI introduces directly. The second can help establish whether students subsequently demonstrate similar improvements independently. This methodological distinction is necessary for producing reliable conclusions about the educational value of AI-assisted learning.

The findings of the literature-based analysis suggest that corpus-based methods provide a useful framework for investigating linguistic changes in Uzbek EFL students'

writing following AI-assisted learning. One of the central arguments is that writing development should be evaluated through observable linguistic features rather than relying exclusively on general impressions of text quality. Corpus linguistics makes it possible to identify patterns in vocabulary, grammar, syntax, and discourse organization across different writing samples. This approach is consistent with Biber et al. (1998), who emphasize the value of systematic corpus analysis for examining language variation and use. The findings also support the importance of combining quantitative and qualitative analysis. Numerical indicators can reveal changes in lexical diversity or grammatical accuracy, while qualitative examination helps determine whether those changes contribute to meaningful communication. The potential contribution of AI-assisted learning to vocabulary development is particularly relevant. Lu (2025) provides empirical evidence that AI-generated corpus learning can support grammatical structure learning and lexical bundle acquisition among EFL learners. Similarly, Cheung and Crosthwaite (2025) demonstrate the potential of combining corpus linguistics and generative AI through corpus-informed chat bots. These studies suggest that AI-assisted learning may become more effective when students are exposed to authentic examples of language rather than relying exclusively on AI-generated explanations. For Uzbek university students, this approach may be particularly useful because corpus examples can demonstrate how English vocabulary and grammatical structures function in authentic academic contexts. However, an important methodological issue concerns the interpretation of post-instruction improvements. If students use ChatGPT to rewrite their assignments, their final texts may demonstrate improved lexical sophistication and grammatical accuracy. Nevertheless, such improvements may reflect the linguistic abilities of the AI system rather than the independent development of the learners. Mizumoto et al. (2024) highlight this problem through their comparative analysis of EFL student writing and ChatGPT-generated texts. Therefore, research investigating AI-assisted learning should include independent writing assessments after the

instructional intervention. This makes it possible to determine whether learners can transfer acquired knowledge into new writing situations. Another important issue concerns the measurement of lexical diversity. Traditional type-token ratio may produce misleading comparisons when texts differ considerably in length. McCarthy and Jarvis (2010) discuss the importance of reliable lexical diversity measures and demonstrate why researchers should consider the limitations of individual indices. Consequently, future research involving Uzbek EFL students should employ multiple indicators rather than depending on a single measure. Grammatical accuracy should also be examined in relation to text length and the number of opportunities learners have to produce particular structures. A student who writes longer and more complex sentences may produce more errors simply because the text contains more grammatical structures. The use of normalized frequencies and appropriate error ratios can help address these methodological challenges. The discussion also highlights the importance of learner autonomy and critical AI literacy. Students should understand that AI-generated feedback may contain inaccurate or contextually inappropriate recommendations. They need to develop the ability to evaluate suggestions rather than accept every generated correction. Teachers should therefore encourage learners to compare AI feedback with grammar references, academic writing examples, and authentic corpus data. In this way, AI-assisted learning can become an opportunity for developing linguistic awareness rather than a mechanism for replacing independent writing. Overall, the analysis indicates that corpus-based research can provide valuable insights into the relationship between AI-assisted instruction and EFL writing development. Nevertheless, empirical investigation involving Uzbek university students is necessary before making definitive claims about the size and statistical significance of these effects.

The findings have several implications for English language teaching in Uzbek higher education institutions. Teachers can integrate corpus-based activities with AI-

assisted writing instruction to help students examine authentic vocabulary, collocations, grammatical structures, and academic expressions. Students should be encouraged to write initial drafts independently before using AI tools for feedback. This provides a clear record of their original language performance. After receiving AI feedback, learners can examine their errors, identify recurring problems, and revise their texts. Teachers can then assign a new independent writing task to evaluate whether students have applied the acquired knowledge. Corpus examples can also be used to verify AI-generated suggestions. This helps students understand that language choices should be supported by authentic usage rather than accepted solely because an AI system recommends them. Assessment should consider grammatical accuracy, lexical diversity, syntactic complexity, cohesion, and overall communicative effectiveness. Finally, institutions should establish clear guidelines concerning responsible AI use, student authorship, and academic integrity.

The present study has several limitations. First, it develops a corpus-informed analytical framework and synthesizes existing research rather than reporting an original corpus collected from Uzbek university students. Second, the linguistic features examined in corpus research cannot independently explain every aspect of writing quality. Argument development, originality, critical thinking, and rhetorical effectiveness also require qualitative assessment. Third, lexical diversity and syntactic complexity measures have methodological limitations and should be interpreted carefully. Another limitation concerns differences in students' proficiency levels, previous AI experience, and digital literacy. Future empirical research should collect paired writing samples from Uzbek EFL students before and after a clearly defined AI-assisted learning intervention. Such research should include independent writing assessments, comparable prompts, delayed post-tests, corpus-based measurements, and qualitative analysis. A comparison between original drafts, AI-revised texts, and

SCIENCE CONNECTS PEOPLE

independent post-intervention writing would be particularly valuable for distinguishing direct AI contributions from genuine learner development.

CONCLUSION

The present study examined the potential of corpus-based analysis for investigating changes in Uzbek EFL students' writing before and after AI-assisted learning. The analysis indicates that lexical diversity, grammatical accuracy, syntactic complexity, academic vocabulary, lexical bundles, and discourse organization are important dimensions for evaluating writing development. Previous research suggests that AI-supported corpus learning may contribute to grammatical knowledge, lexical development, and students' engagement with writing activities (Lu, 2025). The integration of generative AI with authentic academic corpora may also support students' understanding of discipline-specific language use (Cheung & Crosthwaite, 2025). However, the study emphasizes that improvements in AI-revised texts should not automatically be interpreted as evidence of independent learner development. Distinguishing original student writing from AI-generated and AI-revised texts is essential for reliable research. For Uzbek university students, combining AI-assisted feedback with corpus-based learning may provide opportunities to improve vocabulary knowledge, grammatical awareness, and academic writing skills. In conclusion, corpus-based analysis offers a systematic approach to examining the linguistic effects of AI-assisted learning. Its integration into EFL research can help teachers and researchers understand not only whether students' writing changes but also which linguistic features develop and whether these improvements can be maintained independently. Further empirical research using authentic Uzbek EFL learner corpora is necessary to establish the measurable effects of AI-assisted learning on students' writing performance.

References

1. Biber, D., Conrad, S., & Reppen, R. (1998). Corpus linguistics: Investigating language structure and use. Cambridge University Press.

2. Cheung, L., & Crosthwaite, P. (2025). CorpusChat: Integrating corpus linguistics and generative AI for academic writing development. Computer Assisted Language Learning. https://doi.org/10.1080/09588221.2025.2506480

3. Kohnke, L., Moorhouse, B. L., & Zou, D. (2023). ChatGPT for language teaching and learning. RELC Journal, 54(2), 537–550. https://doi.org/10.1177/00336882231162868

4. Lu, C. (2025). AI-generated corpus learning and EFL learners' learning of grammatical structures, lexical bundles, and willingness to write. PLOS ONE, 20(7), e0321544. https://doi.org/10.1371/journal.pone.0321544

5. McCarthy, P. M., & Jarvis, S. (2010). MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. Behavior Research Methods, 42, 381–392. https://doi.org/10.3758/BRM.42.2.381

6. McEnery, T., & Hardie, A. (2012). Corpus linguistics: Method, theory and practice. Cambridge University Press.

7. Mizumoto, A., Yasuda, S., & Tamura, Y. (2024). Identifying ChatGPT-generated texts in EFL students' writing: Through comparative analysis of linguistic fingerprints. Applied Corpus Linguistics, 4(3), 100106. https://doi.org/10.1016/j.acorp.2024.100106