Development of Lexical-Collocational Competence of English Philology Students through Corpus-Informed and AI-Mediated Instruction

Radjabova Khilola Abdurasulovna[2]
Kheelola15@gmail.com
PhD researcher,
Tashkent Institute of Irrigation and Agricultural Mechanization Engineers,
National Research University

Annotation *This study investigates the development of lexical-collocational competence among English philology students through a corpus-based and artificial intelligence-supported instructional approach. The object of the research is students' language performance, while the subject is the formation of their ability to combine lexical units appropriately in academic writing. The relevance of the study is determined by persistent difficulties in mastering collocations in multilingual educational contexts and the influence of interlingual interference. At the same time, rapid advances in digital technologies create new opportunities for improving language instruction. The study is grounded in usage-based and lexical approaches and employs a quasi-experimental design involving 180 students. Traditional instruction was compared with an integrated model incorporating corpus analysis, contrastive techniques, and automated feedback. Statistical analysis, including ANCOVA, confirmed the effectiveness of the proposed methodology. The novelty of the research lies in the development of an integrated model combining corpus resources and artificial intelligence to enhance collocational competence.*

Keywords *Lexical-collocational competence, corpus-based instruction, artificial intelligence, academic writing, interlingual interference, EFL pedagogy*

Ingliz filologiyasi talabalari leksik-kollokatsion kompetensiyasini korpusga asoslangan va sun'iy intellekt vositalari orqali rivojlantirish

Radjabova Xilola Abdurasulovna
Kheelola15@gmail.com
PhD tadqiqotchisi,
"Toshkent irrigatsiya va qishloq xo'jaligini mexanizatsiyalash muhandislari instituti",
Milliy tadqiqot universiteti

Annotatsiya *Mazkur tadqiqot ingliz filologiyasi yo'nalishida tahsil olayotgan talabalar orasida leksik-kollokatsion kompetensiyani rivojlantirish muammosini korpus ma'lumotlariga asoslangan hamda sun'iy intellekt vositalari bilan qo'llab-quvvatlangan ta'lim orqali o'rganishga qaratilgan. Tadqiqotning obyekti filologiya yo'nalishi talabalarining ingliz tilidagi nutq faoliyati, predmeti esa akademik yozuv jarayonida leksik birliklarni mos ravishda birlashtirish ko'nikmalarini shakllantirish jarayonidir. Tadqiqotning dolzarbligi ko'p tilli ta'lim muhitida kollokatsion xatolarning saqlanib qolishi va ona tilining salbiy ta'siri bilan izohlanadi. Shu bilan birga, raqamli texnologiyalar rivoji ta'lim samaradorligini oshirish uchun yangi imkoniyatlar yaratmoqda hamda zamonaviy pedagogik yondashuvlarni joriy etishni*

[2] *Radjabova Kh.A. — ORCID: 0000-0002-1234-237X*

talab etadi. Metodologik asos sifatida leksik yondashuv va foydalanishga asoslangan nazariy qarashlar tanlanib, kvazi-eksperimental usul qo'llanildi. Tadqiqotda 180 nafar talaba ishtirok etdi hamda an'anaviy va integratsiyalashgan ta'lim modellari qiyosiy tahlil qilindi. Statistik tahlillar, jumladan kovariatsion tahlil, metodikaning samaradorligini to'liq tasdiqladi. Tadqiqotning ilmiy yangiligi korpus ma'lumotlari va sun'iy intellekt asosida leksik-kollokatsion kompetensiyani tizimli rivojlantirish modelining ishlab chiqilganligi bilan belgilanadi va ilmiy jihatdan asoslanadi hamda takomillashtiriladi.

Kalit so'zlar *Leksik-kollokatsion kompetensiya, korpus, sun'iy intellekt, akademik yozuv, metodika, pedagogika*

Развитие лексико-коллокационной компетенции студентов английской филологии на основе корпусно-ориентированного и AI-опосредованного обучения

Раджабова Хилола Абдурасуловна
Kheelola15@gmail.com
PhD-исследователь,
"Ташкентский институт инженеров ирригации и механизации сельского хозяйства",
Национальный исследовательский университет

Аннотация *Данное исследование посвящено развитию лексико-коллокационной компетенции студентов английской филологии на основе корпусно-ориентированного обучения с использованием средств искусственного интеллекта. Объектом исследования выступает речевая деятельность студентов, а предметом – формирование навыков правильного сочетания лексических единиц в академическом письме. Актуальность работы обусловлена устойчивыми трудностями овладения коллокациями в условиях многоязычного образовательного пространства, а также влиянием межъязыковой интерференции. Одновременно развитие цифровых технологий создаёт новые возможности для повышения эффективности обучения иностранным языкам. В качестве методологической основы использованы лексический и частотностный подходы, а также применён квазиэкспериментальный метод с участием 180 студентов. Сравнивались традиционная методика обучения и интегрированная модель, включающая корпусный анализ, контрастивный подход и автоматизированную обратную связь. Статистический анализ, включая ковариационный анализ, подтвердил эффективность предложенной методики. Научная новизна заключается в разработке интегрированной модели, объединяющей корпусные ресурсы и технологии искусственного интеллекта для развития коллокационной компетенции на дискурсивном уровне.*

Ключевые слова *Лексические коллокации, коллокационная компетенция, искусственный интеллект, академическое письмо, методика, педагогика*

Introduction

In contemporary applied linguistics and foreign language pedagogy, lexical-collocational competence has emerged as one of the most significant determinants of advanced second language proficiency. The ability to combine lexical units according to the norms of authentic discourse is now widely recognised as essential for achieving fluency, accuracy, and idiomaticity in both spoken and written communication. Empirical research has demonstrated that formulaic language, including collocations, lexical bundles, and prefabricated chunks, constitutes a substantial proportion of natural language use and plays a central role in language processing and production (Conklin & Schmitt, 2021).

Recent developments in usage-based linguistics and cognitive approaches to language acquisition have further reinforced the importance of collocations as primary units of lexical organisation. According to these perspectives, language learning involves the gradual accumulation of statistically salient patterns of co-occurrence rather than the acquisition of isolated words and abstract grammatical rules (Ellis et al., 2020). As learners encounter frequent lexical combinations in authentic input, they begin to internalise them as holistic units, thereby reducing cognitive load and increasing processing efficiency. This phenomenon has been observed across various proficiency levels and linguistic backgrounds, suggesting that collocational competence is a universal component of language expertise (Boers, Dang & Strong, 2020).

Despite the recognised importance of collocations, numerous studies have shown that advanced learners continue to experience persistent difficulties in mastering them. Analyses of learner corpora indicate that non-native speakers often produce atypical lexical combinations due to overgeneralisation, literal translation, or reliance on semantically transparent but pragmatically inappropriate constructions (Paquot, 2019; Laufer & Waldman, 2020). Such errors are particularly noticeable in academic writing, where lexical precision and phraseological competence are crucial for effective argumentation and knowledge construction (Hyland, 2019; Crosthwaite, 2021).

The challenge of developing lexical-collocational competence is especially relevant in multilingual educational contexts characterised by complex patterns of interlingual influence. In Uzbekistan, English philology students typically operate within a linguistic environment shaped by Uzbek as the dominant native language and Russian as an important lingua franca. Typological differences between agglutinative and analytic language systems contribute to negative lexical transfer and the fossilisation of collocational errors in learner production.

At the same time, ongoing educational reforms aimed at internationalisation and digital transformation have created favourable conditions for the integration of innovative pedagogical technologies. Corpus linguistics has become a particularly influential field in language education, providing empirical tools for analysing authentic language patterns and promoting data-driven learning approaches (Brezina, 2021; Egbert & Biber, 2021). Through the use of concordancers and frequency-based analyses, learners can explore real examples of lexical combinations and develop deeper awareness of contextual constraints.

Parallel to the rise of corpus-based pedagogy, artificial intelligence technologies have begun to reshape language learning practices. Automated writing evaluation systems, intelligent tutoring platforms, and large language models offer new opportunities for personalised feedback and adaptive learning environments. Recent research suggests that AI-mediated feedback can significantly enhance lexical accuracy and

learner motivation when combined with reflective pedagogical design (Ranalli, 2022; Kohnke, Moorhouse & Zou, 2023).

Given these developments, there is a growing need to conceptualise integrated methodological frameworks that combine corpus-informed instruction with AI-supported learning strategies. Such frameworks can address the dual challenge of limited authentic input and persistent interlingual interference in multilingual higher education settings. Therefore, the present study seeks to synthesise theoretical insights and pedagogical practices related to lexical-collocational competence development among English philology students and to evaluate the potential of technology-enhanced approaches in the Uzbek context.

Methods

The methodological foundation of the study is based on an interdisciplinary integration of lexical pedagogy, corpus linguistics, cognitive approaches to language learning, and technology-enhanced instruction. Central to this framework is the lexical approach, which emphasises teaching high-frequency lexical combinations as core units of communication rather than focusing exclusively on grammatical structures.

To operationalise these principles in classroom practice, corpus-informed instructional strategies grounded in Data-Driven Learning were employed. Learners engaged in guided exploration of authentic linguistic data through concordance analysis, collocate searches, and register comparisons using digital resources such as COCA, the British National Corpus, and simplified interfaces like SkELL. Such exploratory activities have been shown to facilitate deeper lexical processing and enhance long-term retention of phraseological patterns (Boulton, 2022; Cobb & Boulton, 2023; Frankenberg-Garcia, 2021; Lee et al., 2024).

A contrastive analytical dimension was incorporated to address interlingual interference in multilingual learning contexts. Students compared English collocational patterns with equivalents in Uzbek and Russian, focusing particularly on verb–noun constructions, adjective–noun combinations, and prepositional usage.

In addition to corpus-based exploration, AI-mediated writing feedback systems were integrated into iterative academic writing tasks. These technologies provided automated suggestions concerning lexical choice, phraseological accuracy, and register appropriateness, thereby supporting reflective practice and formative assessment.

Collaborative learning activities were also implemented to facilitate the social construction of lexical knowledge. Through peer feedback sessions, group discussions, and project-based compilation of thematic collocation glossaries, students engaged more actively with authentic language data and developed heightened metalinguistic awareness (Sun & Park, 2023; Zhai et al., 2021).

Overall, the methodological design combined explicit lexical instruction, corpus-based discovery learning, contrastive reflection, and technology-mediated feedback. This integrated approach was intended to support both receptive and productive dimensions of lexical-collocational competence while fostering learner autonomy and academic literacy development.

Results

The study examined the impact of corpus-informed and AI-mediated instruction on the development of lexical-collocational competence among 180 English philology students using a quasi-experimental pre-test/post-test design. Initial analysis confirmed no statistically significant differences between the experimental and control groups, indicating comparable baseline levels of lexical competence.

Following the instructional intervention, the experimental group demonstrated substantially greater improvement in collocational accuracy and lexical diversity. Post-test mean scores increased markedly in

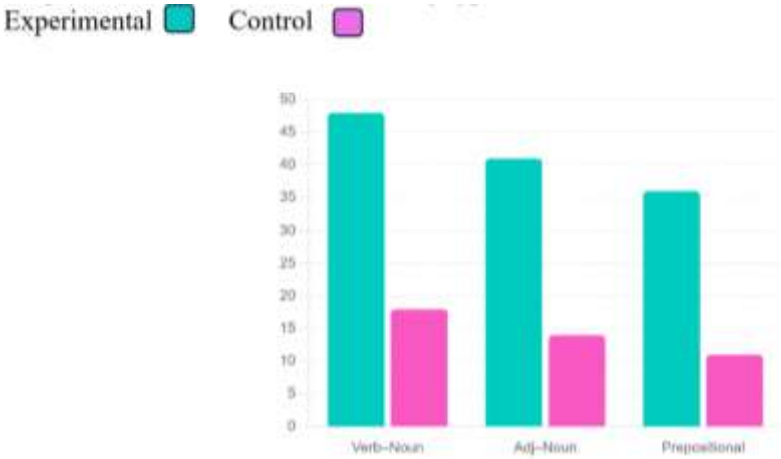
the experimental condition, while gains in the control group remained moderate. ANCOVA results revealed a statistically significant effect of instructional method on post-test performance, with a large effect size ($\eta^2 = .324$), suggesting that approximately one-third of the variance in learning outcomes was attributable to the treatment.

Gain score analysis further confirmed the practical significance of the intervention. The experimental group achieved more than double the improvement of the control group, corresponding to a very large effect size (d = 1.41). Measures of lexical diversity, including Type-Token Ratio and MTLD, also showed stronger growth in the experimental group's written production.

Collocational Error Reduction

Error analysis of student essays demonstrated a substantial reduction in interlingual collocational errors following corpus-informed instruction (Graph 1.). The experimental group reduced verb–noun collocation errors by 48%, adjective–noun errors by 41%, and prepositional collocation errors by 36%. In contrast, the control group achieved reductions of 18%, 14%, and 11%, respectively.

These findings indicate that systematic exposure to authentic corpus data and AI-mediated feedback supports restructuring of learners' mental lexicons and facilitates more accurate lexical selection.



Graph 1. *Collocation error reduction by types*

Correlation Analysis

Pearson correlation analysis revealed a strong positive relationship between corpus consultation frequency and post-test performance ($r = .63$, $p < .001$), suggesting that learners who engaged more actively with corpus tools achieved higher levels of collocational competence. Furthermore, a moderate correlation was observed between AI-feedback usage intensity and lexical diversity growth ($r = .48$, $p < .01$), indicating the complementary role of digital technologies in supporting vocabulary development. The synthesis of empirical research findings and pedagogical practice indicates that systematic corpus-informed instruction contributes to substantial improvements in learners' lexical-collocational competence. One of the most significant outcomes observed across recent studies is the enhancement of learners' ability to recognise frequent lexical patterns in authentic academic texts. When students are regularly exposed to concordance lines and collocate frequency lists, they develop greater sensitivity to the probabilistic nature of lexical co-occurrence. This heightened awareness enables them to distinguish between statistically dominant and marginal lexical combinations, thereby improving lexical accuracy in both controlled and free

production tasks (Boulton, 2022; Sun & Park, 2023).

Discussion

The statistical results provide strong evidence for the effectiveness of corpus-informed and AI-mediated instruction in developing lexical-collocational competence among English philology students. The lack of significant pre-test differences between the experimental and control groups confirms baseline equivalence and supports the internal validity of the research design, allowing post-test improvements to be attributed primarily to the instructional intervention.

Gain score analysis further demonstrated the treatment's effectiveness, as the experimental group achieved more than double the improvement of the control group, indicating a very large educational impact. Such progress reflects noticeable enhancement in learners' ability to produce natural lexical combinations, organise academic discourse more effectively, and communicate with greater confidence in English-medium contexts. These outcomes are consistent with research showing that data-driven learning accelerates vocabulary development and promotes deeper lexical processing.

Improvements in lexical diversity measures, including type-token ratio and measure of textual lexical diversity, suggest that corpus-based instruction supports not only collocational accuracy but also the expansion of learners' productive lexical repertoires. Repeated interaction with authentic corpus input appears to foster flexible lexical retrieval and increased sensitivity to contextual constraints, in line with usage-based theories emphasising the role of frequency and experience in language acquisition. The substantial reduction in interlingual collocational errors, particularly in verb–noun constructions, indicates restructuring of learners' mental lexicons through exposure to statistically dominant target-language patterns. From a psycholinguistic perspective, this process reflects the strengthening of associative links within the L2 network and the gradual weakening of competing L1-based representations.

Correlation findings highlight the importance of learner engagement with digital tools. Frequent corpus consultation was strongly associated with higher post-test performance, suggesting that exploratory language analysis and learner autonomy play a central role in competence development. Similarly, the moderate relationship between AI-feedback use and lexical diversity growth points to the supportive function of automated feedback as a scaffold for refining lexical choices in academic writing.

Overall, the combined use of corpus-informed instruction and AI-mediated feedback appears to generate synergistic cognitive and behavioural learning effects, promoting probabilistic awareness of language patterns and encouraging reflective revision practices. These findings have important implications for curriculum design in multilingual higher education contexts, where digital linguistic resources can compensate for limited natural exposure to English.

However, the quasi-experimental nature of the study limits causal generalisation, and future research should employ randomised and longitudinal designs to examine the durability and transferability of competence gains. Further work is also needed to develop learner corpora and contrastive teaching materials addressing interlingual challenges in multilingual environments.

Conclusion

The present study has demonstrated that the development of lexical-collocational competence represents a fundamental condition for achieving advanced communicative proficiency and academic literacy among English philology students. In contemporary multilingual educational contexts, lexical accuracy and phraseological naturalness are no longer viewed as secondary outcomes of language learning but rather as

core indicators of linguistic expertise and professional readiness. The synthesis of theoretical perspectives and pedagogical evidence confirms that collocations function as essential units of language organisation, facilitating fluent language production, reducing cognitive load, and supporting participation in academic discourse communities.

The findings suggest that corpus-informed instructional practices significantly enhance learners' awareness of authentic lexical patterns and contribute to measurable improvements in both receptive and productive language skills. Through systematic engagement with concordance data and frequency-based analysis, students develop a probabilistic understanding of language use, enabling them to distinguish between acceptable and marginal lexical combinations. This empirical orientation transforms vocabulary learning from a process of memorisation into an exploratory activity grounded in authentic linguistic evidence.

Furthermore, the integration of artificial intelligence technologies into lexical pedagogy has been shown to provide additional pedagogical value. AI-mediated feedback systems support iterative writing development, promote learner autonomy, and enable personalised learning trajectories. When combined with reflective instructional design and collaborative learning practices, such technologies can contribute to more efficient and sustainable competence formation. However, their implementation should be carefully balanced with communicative interaction and critical thinking tasks in order to maintain a holistic approach to language education.

In the specific context of Uzbekistan, the development of lexical-collocational competence is closely linked to broader processes of educational modernisation and internationalisation. Multilingual interference, limited exposure to authentic English input, and the predominance of traditional grammar-centred methodologies remain significant challenges. The adoption of corpus-based and AI-supported pedagogies offers a viable pathway for addressing these issues by fostering empirical language awareness, strengthening academic writing skills, and preparing future teachers to implement innovative instructional strategies in their own professional practice.

The study also highlights the importance of contrastive analysis in multilingual settings. Explicit comparison of collocational patterns across Uzbek, Russian, and English can reduce negative transfer and support the restructuring of learners' mental lexicons. This perspective underscores the need for locally relevant teaching materials, national learner corpora, and targeted teacher training initiatives that reflect the linguistic realities of the region.

Future research should prioritise longitudinal experimental designs investigating the long-term impact of corpus-informed and AI-mediated instruction on lexical development. Additionally, interdisciplinary collaboration between applied linguists, educational technologists, and policy makers will be essential for ensuring the effective integration of digital innovations into higher education curricula.

Overall, the development of lexical-collocational competence through corpus-based and AI-supported methodologies represents a strategically important direction for contemporary philological education. By promoting data-driven language awareness, learner autonomy, and academic precision, such approaches can contribute to the formation of highly qualified English language specialists capable of engaging in global scholarly communication and advancing language education practices in multilingual societies.

References:

1. Conklin, K., & Schmitt, N. (2021). Formulaic language: A review of research and implications for teaching. *Annual Review of Applied Linguistics, 41*, 3–23. https://doi.org/10.1017/S0267190521000014
2. Ellis, N. C., Römer, U., & O'Donnell, M. B. (2020). Usage-based approaches to language acquisition and processing. *Language Learning, 70*(S1), 1–22. https://doi.org/10.1111/lang.12368
3. Boers, F., Dang, T. C. T., & Strong, B. (2020). Comparing the effects of glossing techniques on EFL learners' retention of collocations. *Language Teaching Research, 24*(4), 453–471. https://doi.org/10.1177/1362168818783230
4. Paquot, M. (2019). The phraseological dimension in interlanguage complexity research. *Second Language Research, 35*(1), 121–145. https://doi.org/10.1177/0267658317697796
5. Laufer, B., & Waldman, T. (2020). Verb–noun collocations in second language writing: A corpus analysis. *Language Learning, 70*(1), 1–35. https://doi.org/10.1111/lang.12364
6. Hyland, K. (2019). *Second language writing* (2nd ed.). Cambridge University Press. https://doi.org/10.1017/9781108635541
7. Crosthwaite, P. (2021). Data-driven learning and writing development. *Journal of Second Language Writing, 51*, 100777. https://doi.org/10.1016/j.jslw.2020.100777
8. Brezina, V. (2021). *Statistics in corpus linguistics: A practical guide*. Cambridge University Press. https://doi.org/10.1017/9781108901134
9. Egbert, J., & Biber, D. (2021). Incorporating corpus linguistics in language pedagogy. *TESOL Quarterly, 55*(2), 398–420. https://doi.org/10.1002/tesq.576
10. Ranalli, J. (2022). Automated feedback and L2 writing development. *Language Teaching Research, 26*(3), 1–20. https://doi.org/10.1177/1362168820934886
11. Kohnke, L., Moorhouse, B. L., & Zou, D. (2023). ChatGPT for language teaching and learning. *RELC Journal*. https://doi.org/10.1177/00336882231162868
12. Boulton, A. (2022). Data-driven learning: Taking stock and moving forward. *Language Teaching, 55*(1), 1–20. https://doi.org/10.1017/S0261444821000207
13. Cobb, T., & Boulton, A. (2023). Corpus use in language learning: A meta-analysis. *System, 112*, 102948. https://doi.org/10.1016/j.system.2022.102948
14. Frankenberg-Garcia, A. (2021). Combining corpus and dictionary resources. *International Journal of Lexicography, 34*(1), 1–25. https://doi.org/10.1093/ijl/ecaa021
15. Sun, Y., & Park, M. (2023). Collaborative learning in EFL contexts. *System, 115*, 103026. https://doi.org/10.1016/j.system.2023.103026
16. Zhai, X., et al. (2021). Artificial intelligence in education: A systematic review. *Computers & Education, 172*, 104257. https://doi.org/10.1016/j.compedu.2021.104257