UDC 37:004:81’25
DOI https://doi.org/10.32782/2307-1222.2026-61-5

CORPUS TOOLS IN LANGUAGE EDUCATION: TRENDS, PRACTICES, AND PEDAGOGICAL INSIGHTS

Olena YEMELIANOVA
Candidate of Philological Sciences, Associate Professor,
Associate Professor at the Department of Germanic Philology
Sumy State University
116 Kharkivska str., Sumy
ORCID: 0000-0002-3277-1227
o.emelyanova@gf.sumdu.edu.ua

Olesia YEHOROVA
Candidate of Philological Sciences, Associate Professor,
Associate Professor at the Department of Germanic Philology
Sumy State University
116 Kharkivska str., Sumy
ORCID: 0000-0002-3225-5580
o.egorova@gf.sumdu.edu.ua

The article is devoted to the study of the corpus approach effectiveness in the process of teaching foreign languages and translation. The aim of the article is to investigate the functionality of the corpus approach in translation-oriented teaching of foreign languages by assessing its impact on students’ vocabulary acquisition, written competence, and students’ engagement; to identify key conditions for the productive inclusion of the corpus approach in modern practices of teaching foreign languages and translation. The authors emphasize the need to analyze the potential of corpus tools to ensure linguistic authenticity, improve learning through the use of corpus data, personalize and contextualize learning, and promote the development of student autonomy. The authors demonstrate the effectiveness of the corpus approach in teaching foreign languages and translation, especially in vocabulary acquisition, which is very important for future translators, and suggest practical examples of working with COCA to study word combinations and word formation. The students’ feedback reveals their enthusiasm and positive attitude towards the use of corpora in training. The application of RhymeZone for the development of creative skills and NetSpeak for the study of lexical-grammatical models is also considered. The authors analyze the advantages of the corpus approach, including scientific validity, objectivity, development of autonomy, increased motivation, differentiated and personalized learning, as well as preparation for real speech interaction. The article highlights certain limitations of corpus-based language pedagogy, which include technological barriers, methodological complexity

and certain challenges for teachers, ethical and legal concerns, as well as issues related to data interpretation. Prospects for further research lie in a thorough study of the potential of the corpus approach in teaching translation and linguistics students the genre specificity of the analyzed texts, their lexical and grammatical features, and methods of accurate translation from the original language to the target language.

Key words: *corpus approach, Corpus of Contemporary American English (COCA)***,** *corpus-oriented language pedagogy, differentiated learning, personalized learning.*

ІНСТРУМЕНТИ КОРПУСНОГО АНАЛІЗУ В МОВНІЙ ОСВІТІ: ТРЕНДИ, ПРАКТИКИ ТА МЕТОДИЧНЕ РОЗУМІННЯ

Олена ЄМЕЛЬЯНОВА
кандидатка філологічних наук, доцентка,
доцентка кафедри германської філології
Сумського державного університету
вул. Харківська, 116, м. Суми
ORCID: 0000-0002-3277-1227
o.emelyanova@gf.sumdu.edu.ua

Олеся ЄГОРОВА
кандидатка філологічних наук, доцентка,
доцентка кафедри германської філології
Сумського державного університету
вул. Харківська, 116, м. Суми
ORCID: 0000-0002-3225-5580
o.egorova@gf.sumdu.edu.ua

Статтю присвячено вивченню ефективності застосування корпусного підходу в процесі навчання іноземних мов та перекладу. Метою статті є дослідити ефективність корпусного підходу в перекладоорієнтованому навчанні іноземних мов шляхом оцінки його впливу на засвоєння словникового запасу студентами, письмову компетентність та залученість учнів; визначити ключові умови для продуктивного включення корпусного підходу до сучасних практик викладання іноземних мов та перекладу. Автори наголошують на необхідності аналізу можливостей корпусних інструментів задля забезпечення лінгвістичної автентичності, покращення навчання через використання корпусних даних, персоналізації та контекстуальності навчання, сприяння розвитку автономії студентів. Автори демонструють ефективність корпусного підходу у навчанні іноземних мов та перекладу, особливо в засвоєнні лексики, що є вельми важливим для майбутніх перекладачів, використовуючи практичні приклади роботи з COCA для вивчення словосполучень та словотвору. Наводяться відгуки студентів, які свідчать про їхній ентузіазм та позитивне ставлення до використання корпусів. Також розглядається застосування RhymeZone для розвитку творчих навичок та NetSpeak

для вивчення лексико-граматичних моделей. Автори аналізують переваги корпусного підходу, включаючи наукову обґрунтованість, об'єктивність, розвиток автономії, підвищення мотивації, диференційоване та персоналізоване навчання, а також підготовку до реальної мовленнєвої взаємодії. Стаття висвітлює певні обмеження корпусно-орієнтованої мовної педагогіки, до яких належать: технологічні бар'єри, методична складність і певні виклики для викладачів, етичні та юридичні питання, а також проблеми інтерпретації даних. Перспектива подальших досліджень полягає у ґрунтовному вивченні потенціалу корпусного підходу до навчання студентів-перекладачів жанровій специфіці аналізованих текстів, їхнім лексичним та граматичним особливостям та способам точного перекладу з мови оригіналу на мову перекладу.

Ключові слова: *корпусний підхід, Корпус сучасної американської англійської мови (COCA), корпусно-орієнтована мовна педагогіка, диференційоване навчання, персоналізоване навчання.*

The problem statement. In the era of rapid development of information technology and digitalization of all spheres of human activity – and that's where we are now – language as an object of study is undergoing profound transformations. Linguistics is apt in attuning to these changes by actively introducing advanced methods of language analysis based on large arrays of real text data. One of the most promising areas in this field is corpus linguistics, which in recent years has merged thoroughly with language pedagogy, becoming an integral part of theoretical research and applied disciplines. Corpus-based language pedagogy (CBLP) is the language teaching approach that uses linguistic corpora, which are the systematized collections of texts or spoken language that have been digitized and annotated for computer processing, as the major source of teaching material. This approach allows us to go beyond the unnatural patterns that are quite often typical of traditional textbooks, especially written by non-native authors, and provide quality learning based on genuine, authentic language data.

The role of corpus instruments in language teaching has undergone significant changes recently. The growing number of scientific research and publications in this domain witnesses these transformations. We observe the shift from the simple use of corpora as a data storage to more integrated and dynamic approaches where corpus tools become a vital part of foreign language teaching and learning processes.

Analysis of recent research and publications. One of the key areas of active exploration is the development and use of specialized educational corpora. In contrast to general corpora, which may be too extensive and complicated for novice students, educational corpora are created with due regard for learners' specific needs. Studies show the efficiency of corpus use that is focused on certain lexical and grammatical phenomena, that help students stay concentrated on relevant material [1; 2].

Integration of corpus tools into the curricula and teaching methods are also actively researched; that is done in parallel with corpus development. We observe a noticeable shift from occasional use to systematic implementation of corpus approaches in teaching foreign languages. M. Callies, for example, provides practical recommendations for

pre- and in-service teachers on how to implement corpus linguistics into the day-to-day teaching practice [3]. Another important direction of research is the application of corpus tools in assessing language competence [4]. Technological innovations play a crucial role in corpus tools development. Corpus technologies became more accessible to a wide range of teachers and students who don't have profound knowledge of programming due to the availability of cloud computing and the development of user-friendly interfaces. In this context, works devoted to the development of intuitive platforms and applications, such as the platform CorpusMate introduced by P. Crosthwaite and V. Baisa [5], are of exceptional importance.

We observe a growing interest in interdisciplinary approaches, where corpus linguistics is combined with other areas of research such as artificial intelligence, machine learning, and psycholinguistics [6].

While corpus linguistics as a scientific field is thought to have emerged in the mid-20th century, the ideas of systematically collecting and analyzing texts have much older roots. The computer era made it possible to process vast volumes of text information automatically. At this period appeared such large-scale projects as the British National Corpus (BNC) [7], Corpus of Contemporary American English (COCA) [8], as well as multilingual and parallel corpora. Further advancement of digital media and Internet technologies have led to an exponential growth of accessible textual data, supporting the development of cutting-edge corpora that are dynamic, regularly updated, and thematically specialized. This trend has brought corpus-based language teaching to the forefront. Data-Driven Learning (DDL) gained exceptional recognition and support among educators who promote the idea that students should «discover» linguistic patterns themselves through work with concordances (i.e. examples of word usage in context), rather than merely memorize rules from a textbook [9; 10].

The research goal. The purpose of the article is to investigate the effectiveness of the corpus-based approach in translation-focused language instruction by estimating its influence on students' vocabulary acquisition, writing competency, and learners' engagement; to identify major conditions for its productive blending within modern language teaching frameworks.

Presentation of the main research material. The Corpus-based Language Pedagogy (CBLP) rests on a set of principles that differentiate it from traditional approaches to language teaching. They are: (1) authenticity of the language material is the major advantage of CBLP. Students work with the texts produced by the native speakers in real-life settings. It is crucial for exploring stylistic variations of the language, idiomatic expressions, jargon, etc.; (2) Data-Driven Learning (DDL) encourages students to analyze corpus data and discover linguistic patterns and infer rules autonomously; (3) focus on usage rather than norm. CBLP demonstrates *what people say* (i.e., how language is actually used), but not *how they should say* (i.e., how it ought to be used according to normative grammar); (4) personalization and contextuality**.** Corpora allow educators to adapt learning to particular students' needs, that is, to provide specialized

professional collection of text for thorough examination and analysis (e.g., academic writing, legal texts, medical communication).

Corpus-based approach in teaching foreign languages and translation proves to be highly efficient. One of the most successful areas of corpora use is vocabulary learning [11, p. 344–347]. The application of the corpus approach vividly demonstrates that word meaning is determined by its context and collocations – in contrast to traditional vocabulary acquisition methods, where students learn isolated words. The observation of practical implementation of the Corpus of Contemporary American English [8] (COCA) during 2023–2025 academic years for teaching translation and linguistic students word families and collocations revealed significant learning gains. Students were thoroughly instructed on how to use each feature offered by COCA. The feature COLLOCATES helps students find natural word pairings and by doing so get a better understanding of natural word usage. By interpreting the result, students learn various types of collocates, such as noun + noun (N+N), noun + adjective (N+Adj.), noun + verb (N+V), and noun + adverb (N+Adv.) with a desired word. The system provides examples of the authentic context where a certain lexeme or collocations are used. Additionally, COCA includes links to reliable external resources that encourage students to look deeper into linguistic phenomena. This fosters critical thinking and promotes creative approaches to language learning.

We, as language instructors, consistently highlight the idea that words should be learned in word families ensuring a wide range of students' vocabulary. The BROWSE feature offered by COCA makes the process of word families formation engaging and fun. Students are taught to use the feature BROWSE to find the words with different prefixes and suffixes that belong to the same word family. Along with enhancing vocabulary skills, this activity improves their word formation skills, enriching their knowledge and preparing for the future job as a translator.

Feedback from students indicates a highly positive reception of corpus-based learning. The vast majority of students (97%) felt positive about using corpora for learning foreign languages and translation. Many reported that it was an absolutely new experience for them and that before that course, they had never used corpora as a language learning tool.

The experimental teaching at the translation department during 2023–2025 as to enlarging students' vocabulary via corpora application and an engaging approach demonstrates sustainable results. Students responded well to creative yet intellectually challenging tasks. Recognizing that rhymed words are generally easier to memorize, we incorporated the use of the RhymeZone corpus [12] in our teaching practice to stimulate students' active involvement into the process of knowledge acquisition and enjoyment from the learning process. One illustrative practice involved a warming-up activity at the beginning of the lesson, where students were asked to compose short poems on a specific topic or using a keyword. Having a limited time slot (usually 5–7 minutes), students worked individually or in groups to produce poems. The session concluded with a secret ballot as to whose poem takes the 1st, 2nd, 3rd place, giving the activity a particular competitive streak.

Corpora are of exceptional assistance in helping students identify lexical-grammatical patterns, improving grammatical competence, and exploring controversial cases of a certain lexeme or structure usage. Unlike theoretical textbook explanations, corpus data give a well-grounded insights based on actual usage.

Among translation and linguistics students NetSpeak [13] corpus has gained particular recognition. It offers clear and reliable examples in both English and German, which is especially relevant for students enrolled in the educational program 035/B12 «Germanic Languages and Literatures (Translation included)»; the tool is user-friendly and easy to navigate.

Corpora can help develop and enhance students' writing skills [14]. While writing essays, scientific articles, or engaging in everyday or business correspondence, students make many mistakes. Corpora allow students to compare their own texts with authentic ones for reference. For instance, Michigan Corpus of Upper-Level Student Papers (MICUSP) [15] provides access to a wide range of academic writing samples, including argumentative essays, creative writing, critiques/evaluations, proposals, reports, research papers, and response papers. It is possible to choose disciplines, student levels, nativeness, textual features. The Academic Word Suggestion Machine (AWSUM) [16] is an indispensable tool for teaching students authentic language structures characteristic of different parts of a research paper, such as: abstract, introduction, method, results, discussion, conclusion. Given that translation students are encouraged to engage in scientific and research work from their first year, familiarizing them with these tools is essential for building advanced writing competence and academic literacy.

The analysis shows that CBLP offers a range of significant benefits for foreign language learners, particularly in higher education contexts. Among its key advantages are: (1) scientific validity and objectivity, as the corpus-based approach is empirically foregrounded. Students are taught to rely on the analysis of authentic language data rather than on subjective preferences of teachers or instructors; (2) development of learner autonomy and metacognition. Working with corpora promotes independent learning, where students formulate questions, test hypotheses, and interpret results based on corpus evidence. This helps prepare autonomous learners ready for lifelong education. Moreover, language through data analysis fosters metacognition, i.e. the awareness of one's cognitive processes, so that students become more conscious of *how* they learn, not just what they learn; (3) increased learner motivation. The interactive nature of corpus tools offers, elements of discovery and play, opportunities to «decipher» the language as a riddle, and guarantees that the learning process becomes more interesting and exciting; (4) differentiated and personalized learning, since corpus tools allow instructors to tailor content to students' specific needs, proficiency levels, and academic disciplines; (5) enhanced learner readiness for real-world language use. Practice shows that corpus-trained students are better at navigating authentic texts, more adept at recognizing idioms, set combinations, collocations, and stylistic devices. They are less prone to use outdated expressions and give preference to natural, contextually appropriate constructions.

Practice shows that corpus-trained students are better at navigating authentic texts, more adept at recognizing idioms, set combinations, collocations, and stylistic

devices. They are less prone to use outdated expressions and give preference to natural, contextually appropriate constructions.

Despite obvious advantages, CBLP faces several serious limitations, among which are the following: technological barriers; methodological complexity for teachers; ethical and legal issues connected with copyright, data privacy, and data bias; challenges of linguistic data interpretation; limitations of spoken corpora.

Given the rapid advancement of information technologies nowadays, the future of CBLP looks promising. The integration of corpora with artificial intelligence is already underway.

Conclusions and prospects for further research. Corpus-based language pedagogy emerged in response to criticism of traditional methods that often relied on the teacher's intuition or artificially constructed examples and did not always accurately reflect authentic language use. The corpus-based approach offers an alternative – teaching based on factual, objective, and verifiable data. Corpus pedagogy transforms students from passive recipients of knowledge into active researchers, engaging in analysis, comparison, and hypothesis formation. This aligns with the contemporary educational paradigm that emphasizes student-centered problem-oriented, and inquiry-based construction of knowledge.

Corpus-based language pedagogy is a powerful approach that, with the proper introduction and instruction, can qualitatively change the process of language learning and professional training. Its strength relies on scientific validity, material authenticity, student autonomy, and critical thinking development. However, it is important to remember that corpora are educational assistants, and do not substitute a teacher in the classroom (both traditional and virtual). Educators should keep abreast with the data at hand, latest technological innovations, and consider linguistic intuition, human factor influence, and the correlation between lexemes frequency use and their cultural value.

Further research should focus on exploring the potential of CBLP in teaching future linguists and translators to work with genre-specific texts, their lexical, grammatical, and pragmatic dimensions. Particular attention should be given to developing methodologies that support accurate and adequate translation from the source language into the target one, and curricula redesign to effectively integrate CBLP into existing linguist and translator training programs.

BIBLIOGRAPHY

1. Hejazi H. A corpus-based investigation of lexical bundles and keyness in B1, B2 and C1 ESL learners' academic writing. M.S. thesis, Dept. Linguist., Univ. Liverpool. Liverpool. UK, 2022. 336 p.

2. Hou H. Teaching specialized vocabulary by integrating a corpus-based approach: Implications for ESP course design at the university level. *English Language Teach*ing, 2014. Vol. 7. № 5. P. 26–35. DOI: https://doi.org/10.5539/elt.v7n5p26

3. Callies M. Integrating corpus literacy into language teacher education: The case of learner corpora. *Learner Corpora and Language Teaching*. Amsterdam. The Netherlands: John Benjamins, 2019. P. 246–264. DOI: https://doi.org/10.1075/slcs.201.12cal

4. Kusumaningrum M.V., Ardi P. A corpus analysis of lexical sophistication in LLT Journal: A journal on language and language teaching. *ELTR Journal,* 2020. Vol. 4. № 1. P. 53–75. DOI: https://doi.org/10.37147/eltr.v4i1.39

5. Crosthwaite P., Baisa V. A user-friendly corpus tool for disciplinary data-driven learning: Introducing CorpusMate. *International Journal of Corpus Linguistics*. 2024. Vol. 29. № 4. P. 595–610. DOI: https://doi.org/10.1075/ijcl.23056.cro

6. Popescu D.A., Bold N., Stefanidakis M. A systematic model of an adaptive teaching, learning and assessment environment designed using genetic algorithms. *Applied Sci*ences. 2025. Vol. 15. Art. № 4039. DOI: https://doi.org/10.3390/app15074039

7. British National Corpus. URL: http://www.natcorp.ox.ac.uk/ (дата звернення: 13.01.2026).

8. Corpus of Contemporary American English (COCA). URL: https://www.english-corpora.org/coca/ (дата звернення: 13.01.2026).

9. Men H. Data-driven learning in enhancing learners' language idiomaticity. *International Journal Emerging Technologies Learning*. 2020. Vol. 15. № 23. P. 27–41. DOI: https://doi.org/10.3991/ijet.v15i23.19023

10. Giampieri P. Data-driven learning in English for academic purposes class. *Language Learning in Higher Educ*ation. 2020. Vol. 10. № 1. P. 217–233. DOI: https://doi.org/10.1515/cercles-2020-2006

11. Yemelianova O.V. English corpora for teaching translation students : Monograph / ed. I.V. Ushchapovska. Sumy, 2025. P. 338–361.

12. RhymeZone. URL: https://www.rhymezone.com/r/rhyme.cgi?Word=translator&typeofrhyme=perfect&org1=syl&org2=l&org3=y (дата звернення: 13.01.2026).

13. NetSpeak. URL: https://netspeak.org/ (дата звернення: 13.01.2026).

14. Durrant P. What can a corpus tell us about school writing? Findings, challenges, and future directions. *Applied Corpus Linguistic*s. 2025. Vol. 5. № 2. P. 100134. DOI: https://doi.org/10.1016/j.acorp.2025.100134

15. Michigan Corpus of Upper-Level Student Papers. URL: https://micusp.elicorpora.info/main (дата звернення: 13.01.2026).

16. Academic Word Suggestion Machine (AWSUM). URL: https://langtest.jp/awsum/ (дата звернення: 13.01.2026).

17. English Corpora. URL: https://www.english-corpora.org/corpora.asp (дата звернення: 13.01.2026).

REFERENCES

1. Hejazi, H. (2022). *A corpus-based investigation of lexical bundles and keyness in B1, B2 and C1 ESL learners' academic writing* (Master's thesis, University of Liverpool). University of Liverpool.

2. Hou, H. (2014). Teaching specialized vocabulary by integrating a corpus-based approach: Implications for ESP course design at the university level. *English Language Teaching, 7*(5), 26–35. DOI: https://doi.org/10.5539/elt.v7n5p26

3. Callies, M. (2019). Integrating corpus literacy into language teacher education: The case of learner corpora. In *Learner corpora and language teaching* (pp. 246–264). John Benjamins. DOI: https://doi.org/10.1075/slcs.201.12cal

4. Kusumaningrum, M.V., & Ardi, P. (2020). A corpus analysis of lexical sophistication in *LLT Journal: A Journal on Language and Language Teaching*. *ELTR Journal, 4*(1), 53–75. DOI: https://doi.org/10.37147/eltr.v4i1.39

5. Crosthwaite, P., & Baisa, V. (2024). A user-friendly corpus tool for disciplinary data-driven learning: Introducing CorpusMate. *International Journal of Corpus Linguistics, 29*(4), 595–610. DOI: https://doi.org/10.1075/ijcl.23056.cro

6. Popescu, D.A., Bold, N., & Stefanidakis, M. (2025). A systematic model of an adaptive teaching, learning and assessment environment designed using genetic algorithms. *Applied Sciences, 15*, Article 4039. DOI: https://doi.org/10.3390/app15074039

7. British National Corpus. (n.d.). https://www.natcorp.ox.ac.uk (Accessed January 13, 2026).

8. Corpus of Contemporary American English. (n.d.). https://www.english-corpora.org/coca/ (Accessed January 13, 2026).

9. Men, H. (2020). Data-driven learning in enhancing learners' language idiomaticity. *International Journal of Emerging Technologies in Learning, 15*(23), 27–41. DOI: https://doi.org/10.3991/ijet.v15i23.19023

10. Giampieri, P. (2020). Data-driven learning in English for academic purposes class. *Language Learning in Higher Education, 10*(1), 217–233. DOI: https://doi.org/10.1515/cercles-2020-2006

11. Yemelianova, O.V. (2025). *English corpora for teaching translation students* (I.V. Ushchapovska, Ed.). Sumy.

12. RhymeZone. Retrieved from https://www.rhymezone.com (Accessed January 13, 2026).

13. NetSpeak. Retrieved from https://netspeak.org/ (Accessed January 13, 2026).

14. Durrant, P. (2025). What can a corpus tell us about school writing? Findings, challenges, and future directions. *Applied Corpus Linguistics, 5*(2), Article 100134. DOI: https://doi.org/10.1016/j.acorp.2025.100134

15. Michigan Corpus of Upper-Level Student Papers. Retrieved from https://micusp.elicorpora.info/main (Accessed January 13, 2026).

16. Academic Word Suggestion Machine. Retrieved from https://langtest.jp/awsum/ (Accessed January 13, 2026).

17. English Corpora. Retrieved from https://www.english-corpora.org/corpora.asp (Accessed January 13, 2026).

Дата першого надходження статті до видання: 13.01.2026
Дата прийняття статті до друку після рецензування: 18.02.2026
Дата публікації (оприлюднення) статті: 10.04.2026