UDC 81'322:004.738.5

DOI https://doi.org/10.32782/2522-4077-2025-214.1-5

## ONLINE LEARNING DISCOURSE THROUGH LEMMAS AND COLLOCATION: A CORPUS-BASED METHODOLOGICAL APPROACH

## ДИСКУРС ОНЛАЙН-НАВЧАННЯ ЧЕРЕЗ ЛЕМИ Й КОЛОКАЦІЇ: КОРПУСНИЙ МЕТОДОЛОГІЧНИЙ ПІДХІД

Denysova N. B., orcid.org/0000-0001-9744-4553

Senior Lecturer at the English Language and Communication Department, Borys Grinchenko Kyiv Metropolitan University

This article investigates how online learning is represented in scholarly discourse, drawing on a purpose-built corpus of peer-reviewed scientometric publications (approximately 990,000 tokens). The study examines word frequency and collocational patterns to identify the lexical and phraseological features that shape dominant conceptualizations of digital education. Frequent lemmas such as student, engagement, platform, and satisfaction are shown to play a central role in framing discussions around efficiency, learner involvement, and institutional accountability. The analysis focuses on high-frequency lemmas (student, learning, education, platform) and their recurring combinations, using Sketch Engine to trace how they function within fixed expressions like student engagement, online learning platform, and learning satisfaction. These collocations foreground the evaluative and outcome-oriented tendencies of academic writing on the topic.

The data show that writers tend to repeat the same word combinations, often linking terms like online, learning, and student in similar ways. This pattern seems to follow common habits of academic writing and reflects how online education is typically discussed in research texts.. A close examination of the lemma student reveals a consistent portrayal of learners as recipients of instruction and objects of assessment, with limited emphasis on agency or co-construction. At the same time, the occurrence of terms related to emotional, social, and interactional dimensions signals a parallel concern with the affective and experiential facets of online learning. Overall, the findings suggest that institutional priorities and research traditions significantly shape the linguistic representation of digital education.

Key words: semantics, pragmatics, discourse analysis, academic discourse, corpus linguistics, lemmas, collocations, lexical frequency.

У статті досліджено, як онлайн-навчання репрезентується в академічному дискурсі. Проаналізовано частотність лексем і колокаційні моделі у спеціально сформованому корпусі наукових текстів (приблизно 990 000 токенів), що охоплює рецензовані наукометричні публікації. Такий корпус слугує базою для вивчення повторюваних лем і усталених словосполучень. У центрі уваги — лексеми на зразок *student*, *engagement*, *platform*, *satisfaction*, які формують домінантні наративи про ефективність, залученість і функціонування інституцій. Метою дослідження є виявлення того, як мова академічних текстів відображає ключові освітні запити, пов'язані з цифровим навчанням. Аналіз зосереджено на частотних лемах (*student*, *learning*, *education*, *platform*) та їхніх колокаційних зв'язках, які досліджувалися за допомогою платформи Sketch Engine. Особливу увагу приділено стійким словосполученням на зразок *student engagement*, *online learning platform*, *learning satisfaction*, що відображають фокус на результативність, ефективність і орієнтацію на здобувача освіти.

Результати засвідчують, що дискурс онлайн-освіти характеризується лексичною передбачуваністю та оцінною структурованістю. Частотні сполучення модифікаторів з абстрактними іменниками, зокрема online, learning, student, сигналізують не лише про стилістичні норми академічного письма, а й про концептуальні рамки, в яких здійснюється осмислення цифрової педагогіки. Поглиблений аналіз леми student демонструє схильність до зображення здобувача освіти переважно як об'єкта оцінювання та споживача структурованого контенту. Водночас наявність термінів, що позначають емоційні, соціальні та інтеракційні аспекти, свідчить

про усвідомлення ширшого гуманістичного виміру досвіду навчання. Загалом результати вказують на те, що академічні тексти про онлайн-освіту тяжіють до репрезентації навчального процесу крізь призму інституційних показників, педагогічної ефективності та вимірюваних освітніх результатів.

**Ключові слова:** семантика, прагматика, дискурс-аналіз, академічний дискурс, корпусна лінгвістика, леми, колокації, лексична частотність.

**Introduction.** In recent years, the rapid expansion of digital technologies has transformed the educational landscape. It has put online learning as a central mode of knowledge delivery across institutions worldwide. It still continues to influence teaching practices and educational discourse; it becomes increasingly important to investigate how language shapes and reflects these changes.

Corpus linguistics can be used as a methodological basis, an analytical framework to explore such discourses. It enables researchers to identify commonly occurring lexical patterns and semantic associations in large bodies of text. Lemmas, collocates, and collocations are among the fundamental concepts in corpus-based research, together with analytical tools developed to examine and investigate both the structure and function of language in context.

The aim of the article is to demonstrate how a corpus-based approach can be used to analyze online learning discourse, focusing specifically on lemmas and collocations. Drawing on a purpose-built corpus of peer-reviewed academic and media texts, the study examines high-frequency terms and their phraseological behavior using Sketch Engine. By analyzing expressions like student engagement and online learning platform, the article shows how language helps shape the prevailing narratives about effectiveness, learner roles, and institutional priorities in digital education.

Literature Overview. Corpus linguistics has proved to be a powerful methodology for analyzing authentic language use across genres, disciplines, and modes of communication. Such concepts as lemmas, collocates, and collocations, which enable researchers to explore collocations with precision, not only support theoretical investigations into language structure but also offer practical applications in lexicography, language teaching, translation, and discourse analysis [1; 2; 3].

The lemma, the first of these core concepts, refers to the base or canonical form of a word under which all inflected and derived forms are grouped. Lemmatization plays a foundational role in corpus linguistics as it brings together lexical data, in this way allowing researchers to work with unified word forms and conduct frequency-based analysis with greater accuracy [4]. Lemmatization is particularly valuable in fields of applied research, for instance, biomedicine, where tools like BioLemmatizer standardize domain-specific terminology [5]. Lemmatization also supports lexicographic and pedagogical goals by helping identify core lexical entries and their morphological variants. Another applied approach lies in vocabulary instruction: teaching through lemmas rather than isolated word forms enables learners to recognize and generate multiple word variants, promoting deeper lexical competence [4].

Collocates are the next analytical unit in corpus-based study, referring to words that frequently co-occur with a specified "node" word within a defined span of text. The identification of collocates allows researchers to uncover semantic and syntactic associations that structure meaning in discourse [2]. These associations do not occur randomly, but they are patterned. They often reflect cognitive, cultural, or institutional preferences. In linguistic research, collocate identification is typically supported by statistical measures such as mutual information (MI), t-score, or log-likelihood (LL), which assess the significance of word co-occurrence [6; 7]. Such specialized discourse domains as legal or academic writing have proved that the application of collocate analysis is especially effective in these fields, especially if it can reveal frequently used lexical bundles, field-specific terminology, or evaluative expressions.

Closely related to collocates are collocations, which are semi-fixed or habitual word combinations that tend to co-occur more frequently than would be expected by chance. Unlike collocates, which are statistical in nature, collocations are often considered collocations such as take a risk,

conduct research, or online learning, that contribute to the idiomatic and fluent use of language. Compositionality, fixedness, modifiability, and semantic transparency are categories according to which collocations are often grouped [8; 9]. Their study is critical to understanding not only lexical usage but also stylistic variation across genres and registers. In academic discourse, collocations such as empirical evidence or further research signal disciplinary conventions and genre familiarity, while in pedagogical contexts, collocations are emphasized to help language learners achieve fluency and appropriateness [10; 11].

Collocational studies have also been crucial for translation research as such studies assist in identifying target-language equivalents for fixed expressions that may not have direct lexical parallels [12]. The accurate translation of collocations often requires cultural and contextual sensitivity, as literal translation may distort meaning or sound unnatural. Moreover, in discourse analysis, the study of collocations helps uncover rhetorical and evaluative patterns, thematic coherence, and ideological framing within texts [13]. In such specialized contexts as crisis communication or media reporting, collocations may give certain pictures of topics that are framed, they can indicate what values are emphasized, and which discursive strategies are employed to construct credibility or urgency.

Although lemmas and collocations have been widely studied in areas such as lexicography, academic writing, language learning, and translation, their use in discussions about online education remains underexplored. Given the increasing reliance on digital platforms for teaching and learning, and the rapid proliferation of texts discussing online learning environments, a closer investigation of the collocational patterns within this discourse is timely. This study responds to that gap by employing corpus-driven techniques to examine texts concerned with online education. The analysis shows that certain words and word combinations appear regularly in texts about online learning, suggesting common ways of talking about teaching, students, and digital tools.

**Research Design.** This study uses a corpus-based method to explore how online learning is discussed in academic texts, with a focus on recurring lemmas and collocational patterns. The goal is not to test hypotheses, but to apply a structured, transparent procedure for tracing lexical and semantic features that characterize this domain. This analysis is grounded in real textual data, with patterns of language use identified from the material itself rather than shaped by assumptions. The corpus includes about 990,000 tokens and 796,000 word forms, drawn from academic sources focused on online learning.

Part-of-speech tagging was done by Sketch Engine's default English parser, which recognizes nine main grammatical categories. To support the analysis of word frequency and collocational patterns, inflected forms were grouped under their base forms through lemmatization. The corpus was processed using Sketch Engine, and attention was given to word combinations that occurred regularly within consistent grammatical structures. To identify these patterns, the Word Sketch tool was used to extract collocations based on grammatical relations, such as noun–modifier and verb–object pairs.

To identify lemmas that are unusually frequent in online learning discourse, the pilot corpus was compared with a **reference corpus** of general English (enTenTen20) available in Sketch Engine.

The analysis focused on two aspects of the corpus. First, attention was given to lemma frequency and the presence of key terms in order to determine which lexical items were most prominent in the context of online learning. Second, the study considered collocational patterns, examining how words tend to appear together in characteristic combinations. These patterns offered a view of the collocations commonly used in academic writing on the topic.

Results and Discussion. Lemma Frequency and Keyword Distribution. The most frequent lemmas in the corpus consist largely of function words such as the, of, and, to, and be, which are expected in academic English and provide a structural foundation for discourse. More revealing, however, are the prominent content words that highlight the thematic concerns of the texts. Among these are student, learning, online, education, teacher, model, and research, all of which point to a discourse centered on instructional processes, institutional roles, and pedagogical outcomes. The

recurrence of lemmas such as study, result, use, and performance suggests a research-oriented tone, which clear points to empirical inquiry and assessment in the text under study. The frequent appearance of content-related lemmas suggests that academic discussions of online learning are primarily concerned with learning outcomes, student roles, and teaching approaches within digital settings.

The lemma student was chosen for focused collocational analysis not only because of its high frequency, but also because it functions as a central agentive noun in educational discourse. As one of the most prominent content lemmas, it appears regularly in key expressions such as student satisfaction, student engagement, and student performance. These collocations reflect the evaluative and learner-centered focus of online education. Examining student through Sketch Engine's Word Sketch tool sheds light on the grammatical and semantic associations that structure its use. Patterns of co-occurrence with student point to a strong focus on participation, performance, and learner interaction. These associations suggest that student plays a central role in the way academic texts construct and communicate ideas about online education.

N-gram patterns for the lemma *student* (Figure 2) reveal the repetitive and formulaic nature of its usage in academic discourse on online learning. Students commonly appear in multi-word sequences such as student engagement, student satisfaction, and student performance, indicating a strong tendency toward evaluative and outcome-focused framing. These N-grams demonstrate which fixed word combinations shape representations of learner identity, reducing students to metrics of institutional success. The prominence of such collocations also reflects the influence of managerial discourse in shaping educational priorities and linguistic conventions within scholarly communication.

Many of the collocations found in the corpus including terms like student performance, student satisfaction, and student engagement point to a recurring emphasis on evaluation, outcomes, and measurable involvement. These combinations do more than describe educational practices; they reflect the dominant perspective in which students are primarily seen as individuals to be assessed and managed. Rather than portraying learners as co-creators of knowledge or contributors to the learning process, the language foregrounds their role as recipients of instruction and as data points in

Lemma	Frequency ? ↓	Lemma	Frequency <sup>?</sup> ↓	Lemma	Frequency <sup>?</sup> ↓
1 ,	55,896	11 )	13,972	21 on	5,691
2 the	50,943 ***	12 (	13,903	22 have	5,598 ***
3 .	39,537 ***	13 a	13,827 ***	23 as	5,587 ***
4 of	29,522 ***	14 student	11,384 ***	24 this	5,249 ***
5 and	29,255 ***	15 learning	10,962 ***	25 study	5,046 ***
6 [number]	27,601 ***	16 learn	8,297	26 [	4,944
7 be	24,802 ***	17 -	8,227	27 ]	4,877 ***
8 to	19,659	18 that	7,759	28 use	4,813 •••
9 in	18,130	19 for	7,440 ***	29 ;	4,347 ***
10 online	14,858 ***	20 with	5,806 ***	30 :	4,199 ***

	Lemma	Frequency <sup>?</sup> ↓		Lemma	Frequency <sup>?</sup> ↓	
	by	3,782		41 result	2,350	
	their	3,550		42 not	2,321	
		3,401	•••	43 =	2,241	***
34	from	2,956	•••	44 which	2,228	***
	can	2,883	***	45 course	2,190	***
36	it	2,810	•••	45 or	2,118	***
	model	2,652		47 high	2,068	•••
	teacher	2,580	***	48 more	2,035	***
	education	2,486	•••	49 between	2,033	***
40	research	2,372	***	50 we	1,987	

Fig. 1. Lemma frequency list

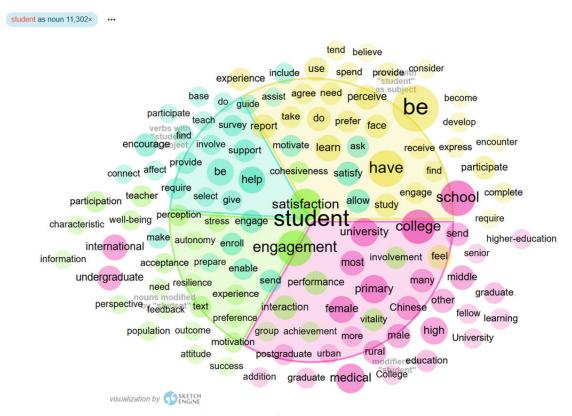


Fig. 2. N-gram of the lemma student

institutional reporting. This tendency reveals how educational discourse often aligns with policy-driven models that prioritize accountability and standardization. In this way, the discourse surrounding online education appears aligned with managerial and results-oriented priorities, privileging standardization, accountability, and quantifiable success over dialogic or exploratory models of pedagogy.

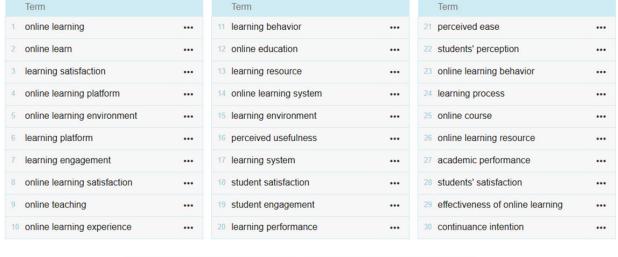


Fig. 3. Collocations with student

The repeated combination of modifiers such as online, learning, and student with central nouns signals a tendency toward fixed lexical pairings that have become conventional within the discourse of digital education. These combinations do not merely serve a descriptive function; they help consolidate institutional language practices, reinforcing dominant ways of framing pedagogical concepts and learner roles within the online environment. Rather than offering variation, the language reflects how certain collocations become routine in academic discussion, shaped by shared institutional and disciplinary habits. Rather than indicating free lexical variation, such pairings suggest a preference for standardized terminology that reinforces dominant ways of conceptualizing education in digital contexts.

Key Multi-Word Terms in Online Learning Discourse. Term extraction using Sketch Engine produced a set of frequently occurring multi-word expressions (MWTs); they characterize the lexical patterns of online learning discourse in scientometric articles. These terms, automatically identified, are based on frequency, grammatical structure, so they offer insight into the dominant conceptual and thematic concerns of the corpus. Figure 4 below presents the top-ranking MWTs identified in the academic texts analyzed.

The list of extracted terms points to a strong focus on how learners interact with and experience online education. Notably frequent expressions include online learning, learning satisfaction, student engagement, learning performance, and online learning platform. These phrases highlight a recurring concern with evaluating both educational outcomes and the effectiveness of the systems that support them. These combinations indicate that much of the discourse is organized around the evaluation of educational outcomes and learner-centered constructs.



Term		Term	
31 self-regulated learning		41 primary school student	
32 behavioral engagement		42 online learning self-efficacy	
33 influencing factor	•••	43 learning method	
34 learning self-efficacy	•••	44 online learning engagement	
35 students' learn	•••	45 behavioural intention	
36 emotional engagement		46 online learner	
perceived ease of use		47 learning activity	
38 discriminant validity		48 face-to-face learning	
39 teacher-student interaction	•••	49 cronbach's alpha	
online learning community		50 social presence	

Fig. 4. Frequent Multi-Word Terms in the Online Learning Corpus (Extracted via Sketch Engine)

A number of collocations reflect the integration of psychological and affective dimensions into the language of online education. For instance, perceived usefulness, perceived ease, students' perception, and self-regulated learning point to the adoption of frameworks from technology acceptance models (e.g., TAM, UTAUT) and educational psychology. Many of the collocations relate to learners' perceptions and engagement, indicating that academic discussions of online learning often foreground how students interpret, navigate, and respond to digital environments.

Other frequently used expressions reflect a strong methodological focus, particularly in relation to measurement tools and research design. Terms like discriminant validity, Cronbach's alpha, and learning method reflect the field's strong focus on quantitative research and measurement. In much of the literature, authors use surveys and statistical modeling to examine student satisfaction, confidence in learning, and academic performance.

From a perspective of collocations, many of the identified terms reflect stable, semi-technical collocations that combine a noun denoting a learning-related outcome or actor (e.g., satisfaction, engagement, system, platform) with a modifier that specifies context (online, learning, student). Many of the frequent terms in the corpus follow familiar phrasing patterns. They often pair words such as online, learning, or student with nouns related to outcomes, systems, or roles. This kind of language use reflects typical academic style, where key ideas are expressed through clear and consistent combinations.

Frequent collocations like teacher-student interaction, face-to-face learning, and online learning community highlight a continued focus on the relationship between digital and traditional modes of education. They also underscore how questions of social presence, learner participation, and evolving definitions of educational legitimacy remain central to discussions about online learning.

Taken together, the multi-word terms suggest that the language of online learning is shaped by institutional priorities, dominant research practices, and common pedagogical models. Frequent references to outcomes such as effectiveness, satisfaction, and performance reflect a strong emphasis on assessment and measurable success. At the same time, terms related to motivation, perception, and interaction highlight a parallel focus on the personal and interpersonal aspects of digital education.

**Conclusions.** This study examined how academic texts represent online learning by analyzing frequent lemmas and collocations in a purpose-built corpus. The results show consistent use of key terms such as *student*, *learning*, *education*, and *platform*, which often appear in stable combinations like *student engagement*, *online learning platform*, and *learning satisfaction*. These collocations are common in scientific articles and academic discussions on online learning. Their frequent use reveals institutional goals, demonstrates assessment practices, and measurable outcomes. The lemma *student* is frequently associated with performance and participation, reinforcing a view of the learner as an object of evaluation. In addition, terms from psychology and educational theory, such as *self-regulated learning*, *perceived usefulness*, and *students' perception*, reflect an interest in how individuals experience and respond to online learning environments.

The analysis suggests that academic writing on this topic often relies on a narrow set of lexical patterns tied to familiar teaching frameworks and research conventions. Corpus tools proved effective for identifying these tendencies, making it possible to work with actual usage rather than assumptions.

Further research could examine whether similar patterns appear in other domains or evolve as educational practices and technologies continue to develop.

## **BIBLIOGRAPHY:**

- 1. Demyanchuk Y. Term-combination in the context of corpus-applied translation studies. *Polonia University Scientific Journal*. 2023. Vol. 59. № 4. P. 16–22. DOI: https://doi.org/10.23856/5902
- 2. Wei Y., Speelman D., Evers-Vermeul J. Applying collocation analysis to Chinese discourse: a case study of causal connectives. *Lingua Sinica*. 2020. Vol. 6. № 1. P. 1–24. DOI: https://doi.org/10.2478/linguasinica-2020-0004
- 3. Qassem M. Style and meaning in translations of the Qur'anic verb-noun collocations into English. *PSU Research Review. 2021*. Vol. 5. № 3. P. 201–214. DOI: https://doi.org/10.1108/prr-12-2020-0041
- 4. Liu H., Christiansen T., Baumgartner W., Verspoor K. Biolemmatizer: a lemmatization tool for morphological processing of biomedical text. *Journal of Biomedical Semantics*. 2012. Vol. 3. № 1. DOI: https://doi.org/10.1186/2041-1480-3-3
- 5. Suparno D., Abshar U., Mulyadi M., Iroth S. Collocation of English, Arabic, and Indonesian COVID-19 terms. *Arabiyat: Jurnal Pendidikan Bahasa Arab dan Kebahasaaraban*. 2021. Vol. 8. № 2. P. 189–208. DOI: https://doi.org/10.15408/a.v8i2.22300
- 6. Březina V., McEnery T., Wattam S. Collocations in context. International *Journal of Corpus Linguistics*. 2015. Vol. 20. № 2. P. 139–173. DOI: https://doi.org/10.1075/ijcl.20.2.01bre
- 7. Prihantoro P. A collocation analysis of 'energy' in Brown Family Corpus. E3S Web of Conferences. 2022. Vol. 359. DOI: https://doi.org/10.1051/e3sconf/202235903012
- 8. Alwazna R. Issues on the translation of certain English collocations into Arabic: from collocations to free constructions in the target language. *English Linguistics Research*. 2018. Vol. 7. № 3. P. 51–58. DOI: https://doi.org/10.5430/elr.v7n3p51
- 9. Stulpinaitė M., Horbačauskienė J., Kasperavičienė R. Lingvistinių kolokacijų vertimo ypatumai. *Studies About Languages*. 2017. № 29. DOI: https://doi.org/10.5755/j01.sal.0.29.15056
- 10. Pellicer-Sánchez A. Learning L2 collocations incidentally from reading. *Language Teaching Research*. 2015. Vol. 21. № 3. P. 381–402. DOI: https://doi.org/10.1177/1362168815618428
- 11. Nizonkiza D., Poel K. Teachability of collocations: the role of word frequency counts. *Southern African Linguistics and Applied Language Studies*. 2014. Vol. 32. № 3. P. 301–316. DOI: https://doi.org/10.2989/160 73614.2014.997061
- 12. Izwaini S. The translation of Arabic lexical collocations. *Translation and Interpreting Studies*. 2016. Vol. 11. № 2. P. 306–328. DOI: https://doi.org/10.1075/tis.11.2.09izw
- 13. Potts A., Bednarek M., Caple H. How can computer-based methods help researchers to investigate news values in large datasets? A corpus linguistic study of the construction of newsworthiness in the reporting on Hurricane Katrina. *Discourse & Communication*. 2015. Vol. 9. № 2. P. 149–172. DOI: https://doi.org/10.1177/1750481314568548

Дата надходження статті: 28.07.2025 Дата прийняття статті: 22.08.2025

Опубліковано: 16.10.2025