JOURNAL OF MULTIDISCIPLINARY SCIENCES AND INNOVATIONS

VOLUME 04, ISSUE 10
MOPNTHLY JOURNALS

ISSN NUMBER: 2751-4390
IMPACT FACTOR: 9,08

CORPUS-BASED ANALYSIS OF LEXICAL CHUNKS IN SCIENTIFIC WRITING

Radjabova Gulnoza Giyosiddinovna
rad.gulnoza@gmail.com
PhD, Associate Professor
Uzbekistan State World Languages University

Abstract: This paper explores the corpus-based analysis of lexical bundles (also known as chunks) within English scientific texts. The study draws from the academic section of COCA, pinpointing the most common three- and four-word combinations and categorizing their functional roles. Results demonstrate that lexical bundles in scientific writing predominantly fulfill three principal functions: guiding text structure (text-oriented), supporting research presentation (research-oriented), and expressing author stance (stance bundles). The research underscores that understanding these patterns can significantly aid students in advancing their academic writing abilities.

Keywords: lexical bundles, scientific writing, corpus linguistics, COCA, academic skills.

Introduction

Corpus linguistics has revolutionized the study of language by providing systematic access to vast collections of real texts. Unlike intuition-based methods, corpus analyses enable researchers to uncover genuine language patterns and features that might otherwise remain hidden. Lexical bundles, or recurring sequences of words, have emerged as a key area of interest in such research. These multiword units, usually three to five words long, frequently appear in specific contexts, forming the foundation of natural communication. In scientific writing, these bundles are particularly important. Academic texts demand accuracy, clarity, and impartiality, qualities often achieved through the repeated use of set expressions that organize arguments, present findings, and convey evaluative stances (Radjabova G., 2025). Common phrases such as "the results of the," "as can be seen in," or "it is important to note" facilitate the communication of research goals, link evidence, and guide the reader through complex information. These bundles not only foster cohesion but also help define disciplinary conventions, as certain phrases are more common in scientific texts than in other genres like journalism or fiction.

Past studies (Biber et al., 2004; Hyland, 2008; Cortes, 2004) have grouped lexical bundles in academic writing into functional categories—research-oriented, text-oriented, and stance bundles. Each group serves a unique communicative purpose: research-oriented bundles explain methods and findings; text-oriented bundles structure the argument; stance bundles express the writer's evaluation or certainty. Grasping the usage and function of these bundles is key for both linguistic description and academic instruction, since they represent formulaic patterns essential to academic literacy. This article offers a corpus-based examination of lexical bundles in English scientific writing, focusing on the academic sub-corpus of the Corpus of Contemporary American English (COCA). By cataloging and analyzing frequent bundles and their functions, the study aims to clarify how these recurring expressions enhance the communicative power of scientific texts. The research also considers the teaching implications for academic writing, especially for learners using English as a second or foreign language.

Research Methods

The study uses the academic sub-corpus of the COCA as its main dataset. COCA, created by Mark Davies (2009), is a comprehensive English corpus with over a billion words from various genres, including spoken, fiction, magazines, newspapers, and academic sources. This research centers on the academic sub-corpus, which consists of scholarly articles, textbooks, research reports, and other scientific texts from multiple disciplines (Giyosiddinovna, Radjabova G., 2022). Such diversity ensures that the analysis generalizes across fields rather than focusing on a single area, making COCA ideal for examining phraseological patterns in scientific writing. In the initial stage, frequent three- and four-word lexical bundles were identified in the academic sub-corpus. Consistent with Biber, Johansson, Leech, Conrad, and Finegan (1999), a bundle is defined here as a continuous word sequence occurring at least 20 times per million corpus words, ensuring that only common, conventionalized expressions are analyzed. The corpus's n-gram search feature generated frequency lists, which were then filtered to remove irrelevant or random strings (e.g., names, numbers, or fragments lacking semantic or discourse value) (Giyosiddinovna, Radjabova G., 2021).

Further exclusions were made: bundles primarily composed of function words with no clear discourse role were disregarded unless they contributed to cohesion, and strictly technical bundles (e.g., "the polymerase chain reaction") were considered only if they served wider discourse functions. The resulting set of frequent, pedagogically useful bundles was the focus of subsequent analysis (Giyosiddinovna, R. G., 2025).

Identified bundles were sorted into functional categories following Biber, Conrad, and Cortes (2004) and Hyland (2008):

- Research-oriented bundles which help describe research aims, processes, and results (e.g., "the purpose of the," "the results of the," "as shown in figure") by situating the work within a scholarly tradition and guiding readers through empirical sections
- Text-oriented bundles which manage text organization and cohesion ("in the context of," "on the basis of," "as a result of"), linking ideas and facilitating logical flow, crucial for clarity in complex arguments.
- Stance bundles reflect the writer's perspective and certainty ("it is clear that," "it may be argued that," "it should be noted"), balancing objectivity with careful interpretation and aligning with academic conventions (Giyosiddinovna, R. G., 2025).

Each bundle was examined in context and assigned to a category, with ambiguous cases checked against multiple corpus examples. The analysis combined quantitative (frequency counts, normalized per million words) and qualitative (contextual interpretation via concordance lines) methods. The fifty most frequent bundles underwent detailed scrutiny to connect use frequency with discourse function, thus deepening understanding of their role in academic texts (Giyosiddinovna, R. G., 2025). For instance, research-oriented bundles such as the aim of this study, the findings of this research, and data were collected from are frequently used to situate the research within the broader scholarly context. Text-oriented bundles like in terms of the, on the other hand, and in accordance with the help organize the flow and logic of the text. Meanwhile, stance bundles such as it is important to note, the evidence suggests that, and it remains unclear whether allow authors to express evaluation, highlight significance, or signal uncertainty. These and similar bundles are essential tools for structuring scientific arguments and making academic writing more effective and reader-friendly.

To ensure accuracy, several steps were taken: COCA's size and balance guaranteed authentic academic language; established thresholds and taxonomies ensured comparability; and manual verification minimized misclassification. In sum, the methodology was corpus-driven and systematic: the academic sub-corpus of COCA provided data, frequency analysis revealed recurring bundles, functional classification clarified their roles, and qualitative concordance analysis interpreted their context. This comprehensive approach facilitated a nuanced understanding of how bundles function in scientific texts.

Results and Discussion

Analysis of the COCA academic sub-corpus confirms that lexical bundles are integral to scientific writing, helping structure arguments, present information, and guide readers through dense material. Findings echo earlier research (Biber et al., 2004; Cortes, 2004; Hyland, 2008), showing that scientific writing relies on a finite set of common multiword units with predictable roles. The majority of frequent bundles fell into three main categories: research-oriented, text-oriented, and stance bundles. Quantitatively, research-oriented bundles were most prevalent, followed by text-oriented bundles; stance bundles were less common but still significant. This distribution reflects the scientific emphasis on factual reporting and clarity over overt evaluation or subjectivity.

Research-oriented bundles dominated the dataset, describing research methods, aims, and findings (Раджабова, Г., 2025), with phrases like “the results of the,” “the purpose of the,” and “as shown in figure” appearing across disciplines. Their ubiquity shows that scientific writers often rely on established formulations to present findings efficiently. This aligns with the view that academic writing is characterized by routinized expressions that streamline research reporting (Biber, Conrad, & Cortes, 2004; Hyland, 2008). For instance, “the results of the” not only introduces new findings but also fulfills reader expectations for research presentation.

Text-oriented bundles, such as “in the context of,” “on the basis of,” and “with respect to the,” made up the next largest category. These expressions are essential for organizing discourse, connecting ideas, and maintaining logical progression, especially in introductions and literature reviews (Раджабова, Г., 2025).

Cortes (2004) noted that students and novice writers often depend on text-oriented bundles for cohesion; the present study finds that even experienced authors regularly employ these patterns, underlining their importance in academic communication. As Hyland (2008) observes, academic texts are highly intertextual, and text-oriented bundles help authors integrate sources, develop arguments, and guide readers toward conclusions. Stance bundles, while less frequent, are crucial for expressing evaluation and caution. Phrases like “it is clear that,” “it should be noted,” and “it is possible that” are used to introduce claims with varying degrees of certainty. Though not as common as other types, they are vital for balancing objectivity with rhetorical nuance (Hyland, 2005). For example, "it is clear that" asserts a claim’s obviousness, while “it is possible that” introduces cautious speculation. Such expressions help manage the tone and authority of scientific texts. Comparing COCA’s academic corpus with other registers shows the distinctiveness of scientific discourse: while conversation is marked by interactive bundles (e.g., “I don’t know if”), scientific writing emphasizes precision and organization (Biber et al., 1999). This demonstrates the register-specific adaptation of lexical bundles (Radjabova, G. G., 2024).

Pedagogical Implications

The findings highlight the need to teach lexical bundles explicitly in academic writing courses. Since these expressions are fundamental to well-structured, discipline-appropriate texts, instruction in their forms and functions can enhance students’ writing. Formulaic language reduces processing effort and makes communication smoother (Wray, 2002). Teaching research-oriented bundles aids in reporting findings, text-oriented bundles improve flow, and stance bundles help students develop a balanced academic voice. Corpus tools like COCA encourage students to discover authentic usage patterns, fostering greater language awareness and autonomy (Boulton, 2021). Overall, the research demonstrates that lexical bundles underpin the effectiveness of English scientific writing. Research-oriented bundles facilitate reporting, text-oriented bundles ensure coherence, and stance bundles add evaluative nuance (Radjabova, G. G., 2024), all supporting knowledge transfer and disciplinary identity. The results of this study have clear and practical implications for teaching academic writing, especially in contexts where English is a second or foreign language. Explicit instruction in lexical bundles can greatly enhance students' ability to produce coherent, well-structured, and discipline-appropriate academic texts. Integrating lexical bundles into pedagogy offers several concrete benefits:

Firstly, incorporating corpus-based resources such as COCA into classroom activities allows students to observe authentic usage of lexical bundles in context. For instance, students can be tasked with searching for and analyzing frequent bundles such as "the results of the study," "as shown in Figure," or "it is important to note that" in published articles. This encourages data-driven learning, moving students beyond rote memorization and fostering a deeper understanding of how these expressions function within scientific discourse.

Secondly, teachers can design reading and annotation exercises in which students identify and collect lexical bundles while reviewing journal articles. For example, they might highlight expressions like "according to the findings," "in the context of," or "the aim of this research." Class discussions can then focus on categorizing these bundles as research-oriented, text-oriented, or stance bundles, and on understanding their communicative purposes.

Writing activities can further reinforce this knowledge. Students may be given controlled practice exercises, such as completing sentences with appropriate bundles:

"The purpose of this study is to __."
"It should be emphasized that __."
"In accordance with previous research, __."

Such tasks allow learners to internalize the structure and flow of academic arguments. Additionally, genre awareness can be developed by comparing the use of bundles across disciplines. For instance, bundles like "the experimental results show" or "as illustrated in Figure 3" are common in engineering and natural sciences, while "from the perspective of" or "it is generally accepted that" may be prevalent in the social sciences and humanities. This helps students tailor their academic writing to the conventions of their specific field.

Peer review sessions can also be leveraged by asking students to identify lexical bundles in each other's drafts and suggest improvements. For example, substituting informal connectors with formal academic bundles: replacing "and then" with "in addition to this," or "but" with "on the other hand." Maintaining personal lists of useful bundles, such as "the data indicate that," "it remains unclear whether," or "with respect to the," can further support students during the drafting and revision stages. Stance bundles, in particular, help learners express certainty or caution as appropriate, using phrases like "it is likely that," "the evidence suggests that," or "it is apparent that" to convey a nuanced academic voice.

In summary, the integration of lexical bundles into academic writing instruction fosters students' linguistic competence, enhances their genre awareness, and empowers them to construct clear, cohesive, and persuasive scientific texts. By equipping learners with these formulaic sequences, educators support not only grammatical accuracy but also the development of authentic disciplinary identity and greater confidence in academic communication.

Conclusion

This study reaffirms the indispensable role of lexical bundles in English-language scientific discourse. Research-oriented bundles are fundamental for presenting research aims, methodologies, and findings, ensuring that complex information is communicated clearly and systematically. Text-oriented bundles contribute substantially to the organization and coherence of academic texts, allowing writers to structure arguments and link ideas logically across sections. Stance bundles, while less frequent, offer vital resources for expressing evaluation, caution, and authorial presence, enabling writers to maintain both objectivity and rhetorical subtlety. The outcomes of this research indicate that familiarity with these recurrent multiword expressions is not merely a stylistic advantage but a core component of academic literacy. The prevalence of such bundles in scientific writing demonstrates their function as linguistic building blocks, supporting both the transmission of knowledge and the construction of disciplinary identity. For non-native speakers and novice writers, explicit instruction and awareness of lexical bundles can bridge the gap between language proficiency and effective academic communication. Moreover, the integration of corpus-based approaches, such as those enabled by COCA, presents valuable opportunities for both research and pedagogy. These approaches facilitate data-driven learning, encourage critical engagement with authentic texts, and empower students to recognize and apply conventionalized patterns in their own writing. Instructors can leverage these findings to develop targeted teaching materials and activities, fostering greater autonomy and confidence among learners as they navigate the conventions of academic discourse.

In conclusion, the study highlights the significance of lexical bundles not only for linguistic analysis but also for educational practice. Promoting the understanding and use of these formulaic sequences can enhance students’ ability to produce coherent, precise, and credible academic texts across disciplines. Future research might extend this work by examining disciplinary variations in bundle usage, exploring their development in learner writing, or assessing the impact of targeted instruction on academic writing proficiency. Ultimately, the study reaffirms that mastery of lexical bundles is central to successful participation in the global academic community.

References:

1. Biber, D., Conrad, S., & Cortes, V. (2004). If you look at …: Lexical bundles in university teaching and textbooks. Applied Linguistics, 25(3), 371–405.
2. Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). Longman Grammar of Spoken and Written English. Harlow: Pearson Education.
3. Cortes, V. (2004). Lexical bundles in published and student disciplinary writing: Examples from history and biology. English for Specific Purposes, 23(4), 397–423.
4. Davies, M. (2009). The 385+ million word Corpus of Contemporary American English (COCA): Design, architecture, and linguistic insights. International Journal of Corpus Linguistics, 14(2), 159–190.
5. Giyosiddinovna, Radjabova G. "Methodological Characteristics of Corpus Technologies in Teaching Foreign Language." International Journal on Integrated Education, vol. 5, no. 1, 2022, pp. 157-163, doi:10.31149/ijie.v5i1.2645.
6. Giyosiddinovna, Radjabova G. "The Implementation of Spoken Corpora in Creating Teaching Materials." International Journal on Integrated Education, vol. 4, no. 5, 2021, pp. 349-354.
7. Giyosiddinovna, R. G. (2025). The Impact of Gamification on Vocabulary Retention and Student Motivation. Ilm fan taraqqiyotida raqamli iqtisodiyot va zamonaviy ta'limning o'rni hamda rivojlanish omillari, 6(1), 64-69.
8. Giyosiddinovna, R. G. (2025). Linguistic Aspects of Automatic Text Alignment in Parallel Corpora. Western European Journal of Linguistics and Education, 3(05), 146-150.
9. Hyland, K. (2008). As can be seen: Lexical bundles and disciplinary variation. English for Specific Purposes, 27(1), 4–21.
10. Radjabova, G. (2023). Corpus technologies in teaching academic writing. Foreign Languages in Uzbekistan, 1(48), 92-103.
11. Radjabova, G. G. (2024). Adjusting the Perspective of Corpus Linguistics: Bridging Research and the Classroom. American Journal of Modern World Sciences, 1(5), 324-332.