*International Journal of Arts and Social Science* www.ijassjournal.com
*ISSN: 2581-7922,*
*Volume 9 Issue 3, March 2026*

# Structural and Functional Analysis of 4-Word Clusters in Sociolinguistic Academic Writing ——A Comparison Between Expert and Student Writers

**Abstract:** This study investigates the structural and functional differences in 4-word lexical clusters between expert and student writers in sociolinguistics. The results reveal distinct patterns. Expert writers predominantly use noun and prepositional clusters to convey dense, discipline-specific information, with functions that are overwhelmingly research-oriented. In contrast, student writers employ a wider range of verb and clausal structures. Their clusters serve more diverse functions, including text-oriented and participant-oriented phrases like it is important to, which help organize arguments and engage the reader. These findings suggest a developmental shift from a focus on sentence construction to one on conceptual knowledge. Academic writing pedagogy should therefore emphasize discipline-specific, high-density phrases to help students bridge this gap.

**Keywords:** lexical clusters, corpus linguistics, academic writing

## I. Introduction

Academic discourse serves as an important medium for disseminating knowledge and facilitating scholarly communication. It not only conveys scientific information and produces credible texts but also expresses rich interpersonal meanings (Hyland, 2002). However, not all students of academic writing can quickly and appropriately use suitable grammatical structures and lexical combinations as native speakers do. Therefore, domestic scholars need to learn from expert writing techniques and adjust their learning goals and targets systematically, according to their own proficiency levels. Lexical clusters are recurrent multi-word sequences that significantly cross the boundaries between vocabulary and syntax, serving as fundamental building blocks of discourse. They effectively fulfill certain grammatical, textual, or pragmatic functions. As minimal units combining form and meaning with textual functions, lexical clusters can reveal and clarify features of academic discourse. They represent important explicit markers of second-language students' academic writing ability (Jiang Lei, Kang Mengchao, & Xiao Yao, 2024). Studying lexical clusters in academic discourse can thus provide valuable references for teaching academic writing.

Many scholars have investigated lexical clusters in academic discourse, focusing mainly on textual analysis and comparative research. (1) Regarding textual analysis using lexical clusters, numerous studies have examined

academic English writing. Li Xiaohong (2021) explored the functional characteristics of phrase frames in doctoral dissertation introductions of Chinese linguistics majors. The study concluded that extensive use of discourse-functional phrase frames is constrained by highly conventionalized introduction structures; Farhang-Ju *et al.* (2024) analyzed the relationship between lexical clusters and rhetorical moves in 1000 research article introductions. It concluded that some clusters are specific to certain moves while others are more general, providing a clear instructional framework for academic writing. (2) Currently, lexical cluster research has expanded beyond native-speaker or expert discourse analysis, increasingly focusing on comparative studies of learner academic texts and differences in lexical cluster use between students and experts. Li Mengxiao and Liu Yongbing (2016) examined differences and similarities in four-word clusters used by Chinese and international scholars in empirical applied linguistics journal articles from structural and functional perspectives. Their analysis revealed Chinese scholars frequently used clausal lexical clusters, while native speakers preferred phrasal clustersl; Li Yan and Jiang Yajun (2024), based on a self-built academic corpus, comparatively analyzed four-word clusters in academic texts by Chinese and international authors. Their findings showed Chinese authors used a greater number and frequency of lexical clusters overall but with lower diversity; Shin and Won (2024) explored lexical clusters in academic L2 English in parallel corpora of written and spoken data produced by the same students. The findings showed that students employed phrasal/referential clusters, typical of academic prose, significantly more in their essays than in their speeches.

In summary, learner corpora effectively represent and reflect probability attributes and distribution features of second-language students' authentic language use. Thus, they serve as valuable tools for describing and explaining the characteristics of students at different proficiency levels. However, due to difficulties in corpus collection, there remains considerable scope for exploring lexical cluster research based on learner corpora. To better understand students' academic discourse practices, this study adopts a comparative perspective, investigating structural and functional differences in lexical cluster usage between students and experts in academic writing within the field of sociolinguistics.

## II. Theoretical Framework

This study categorizes the target lexical clusters according to their structural forms and pragmatic functions. In terms of structural classification, this research adopts the earliest twelve-category system proposed by Biber *et al.* (1999) for academic discourse. As pioneers in lexical cluster research, Biber's structural classification framework has become a primary reference in academic lexical cluster studies (Yang Chuanming, Guo Yunjie, & Leng Litian, 2025). Based on this framework and previous relevant studies, and considering the practical use of lexical clusters, the present study simplifies the structural classification into four categories: noun clusters, prepositional clusters, verb clusters, and clausal clusters, as shown in Table 1.

**Table 1. Structure classification of word clusters**

| Structural Categories | Examples |
|---|---|
| noun clusters | the end of the; the case of the |
| verb clusters | is one of the, is the number of |
| prepositional clusters | in the case of; with respect to the |
| clausal clusters | it is possible that; which is to be |

Regarding functional classification, this study is based on the approach proposed by the British applied linguist Ken Hyland (2008), who categorized lexical clusters into three types: research-oriented, text-oriented, and participant-oriented clusters. Specifically, research-oriented clusters focus on describing activities and experiences occurring in the real world. Text-oriented clusters emphasize structuring discourse and organizing logic and hierarchy. Participant-oriented clusters concentrate on authors and readers of the text, including stance clusters that express authors' attitudes and evaluations, as well as engagement clusters that directly interact with readers (Table 2).

**Table 2. Functional classification of word clusters**

| Functional Categories | Examples |
|---|---|
| research-oriented | can be used to; the nature of the |
| text-oriented | it has been shown; in this paper we |
| participant-oriented | may be due to; it is important to |

In summary, this study combines the structural forms and pragmatic functions of lexical clusters, providing a solid theoretical foundation. This framework helps systematically reveal similarities and differences in lexical cluster usage between students and expert scholars in sociolinguistics-related academic writing.

## III. Research Design

### 3.1 Corpus Collection and Processing

To investigate the differences in lexical cluster usage between student and expert academic papers in sociolinguistics, this study created two corpora: the Expert Sociolinguistics Corpus (ECS) and the Student Sociolinguistics Corpus (SCS). The ECS data was downloaded through AntcorGen. Specifically, the relevant research area was selected by clicking through the sequence: Social Sciences→Linguistics→Sociolinguistics. The SCS data was extracted from the British Academic Written English (BAWE) corpus. Eighteen sociolinguistics-related papers, totaling 286,239 characters, were selected from the linguistics section of BAWE. Given that the data downloaded from AntcorGen was significantly larger than that of BAWE, 20 papers were randomly selected from the AntcorGen corpus to maintain a similar size. The ECS subset used in this research includes 287,541 characters. The papers in both corpora selected the Introduction, Materials & Methods, Results & Discussion, and

Conclusion sections as the objects of study. Altogether, the two corpora contain a total of 38 papers which have 573,780 characters.

### 3.2 Research Methods

This study used AntConc 3.5.9 to extract 4-word clusters and their frequencies from the two self-built corpora. In academic writing, lexical clusters that have higher frequency and wider distribution across texts carry greater research value. This study specifically focused on four-word clusters because these often incorporate shorter clusters (two-word and three-word clusters), and occur much more frequently than longer clusters (five-word and six-word clusters), demonstrating relative continuity and stability (Cortes, 2004). The research followed a procedure of corpus construction→frequency calculation→analysis and interpretation, aiming to identify structural and functional differences in lexical cluster usage between students and experts.

### 3.3 Research Questions

This study examines the structural and functional features of lexical clusters used by students and experts in sociolinguistics-related academic writing, aiming to answer the following two research questions:

1. What are the structural features of lexical clusters used by students and experts in sociolinguistics academic writing, and what similarities and differences exist?

2. What are the main functions of lexical clusters used by students and experts in sociolinguistics academic writing, and what similarities and differences exist?

## IV. Results and Discussion

To ensure the validity and representativeness of the analysis, this study first ranked the 4-word clusters based on their frequency of occurrence. It then manually removed those that lacked clear semantic function or were meaningless combinations caused by word segmentation errors. After that, the top 15 4-word clusters with the highest frequencies in each corpus were selected. These clusters were further analyzed based on their raw frequency and their normalized frequency per million words (PMW).

### 4.1 Characteristics of Cluster Structures Across Different Proficiency Levels

To explore the similarities and differences in the use of lexical cluster structures between student and expert writers, this study adopts the classification framework proposed by Biber et al. (1999) to categorize and analyze the high-frequency lexical clusters in the two corpora. The statistical results are presented in Table 3.

**Table 3. Structural Distribution of Lexical Clusters**

| Structural Categories | ECS | SCS |
|---|---|---|
| noun clusters | 9 | 5 |
| verb clusters | 1 | 5 |

| | | |
|---|---|---|
| prepositional clusters | 5 | 0 |
| clausal clusters | 0 | 5 |
| Summary | 15 | 15 |

An analysis of the high-frequency lexical clusters in SCS reveals a relatively balanced distribution of structural types. Among the top lexical clusters, noun clusters, verb clusters, and clausal clusters each account for five items, jointly forming the most frequently used structural modules in students' academic writing. Notably, clausal clusters such as *it is important to* (PMW = 249.8) and verb clusters such as *can be found in* (PMW = 230.6) stand out. Noun clusters in the student writing demonstrate both generality and topic relevance. They include widely used academic expressions such as *the use of the* (PMW = 211.4) as well as more topic-specific clusters like *the way we speak* (PMW = 172.9), which is closely tied to themes in sociolinguistics. However, it is worth noting that prepositional clusters are entirely absent from the list of high-frequency items in the student corpus. This may suggest a limitation in students' ability to use complex prepositional structures to build nuanced and varied logical relationships. Instead, they tend to rely more on verb-based and clausal constructions with clearer discourse functions to advance their arguments.

In contrast, the high-frequency lexical clusters in ECS display a high degree of structural concentration and disciplinary specificity. Noun clusters and prepositional clusters dominate the list, with nine and six items respectively, together accounting for over 90% of the total. This pattern reflects a key characteristic of expert academic writing——a strong tendency toward nominalization and information compression. Expert writers often condense complex concepts, processes, and relationships into dense nominal expressions, such as t*he viability constraint set* (PMW = 229.5), *the dialect change rate* (PMW = 153.0), and *the mixed effects models* (PMW = 153.0). These clusters function as technical terms or core concepts in sociolinguistics, and their frequent use highlights the writers' deep disciplinary knowledge. Meanwhile, prepositional clusters such as *in the case of* (PMW = 286.9) and *as a function of* (PMW = 229.5) are also widely employed. These serve not only to express logical relations but also to precisely define conceptual boundaries and establish theoretical references.

In contrast to the student corpus, verb-based and clausal clusters are rare in the expert writers' high-frequency list. This further illustrates the expert focus on presenting and reasoning about conceptual entities, rather than on merely constructing sentences.

### 4.2 Characteristics of Cluster Functions Across Different Proficiency Levels

This study adopts Hyland's (2008) functional classification framework to categorize and compare the high-frequency lexical clusters identified in both corpora. The statistical results are presented in Table 4.

**Table 4. Functional Distribution of Lexical Clusters**

| Functional Categories | ECS | SCS |
|---|---|---|
| Research-oriented | 14 | 8 |
| Text-oriented | 1 | 4 |
| Participant-oriented | 0 | 3 |
| Summary | 15 | 15 |

In SCS, the functional distribution of lexical clusters shows a diversified pattern. Research-oriented clusters are the most frequent, with a total of eight, primarily used to describe specific research objects, actions, or phenomena. For example, *the way we speak* and *adjective is referring to*. This suggests that student writers are able to construct discussions around concrete research topics. However, what stands out more prominently is the significant presence of text-oriented and participant-oriented clusters. Text-oriented clusters such as *have been*

*found to* and *can be seen in*, alongside participant-oriented clusters like *it is important to* (PMW = 249.8), collectively reveal an important psychological tendency in student writing. They are highly concerned with constructing coherent discourse and guiding or persuading the reader. These clusters help students organize argumentative structures and clearly express their rhetorical intentions. This is not only a typical strategy employed to ensure clarity and persuasiveness, but also reflects their developing authorial identity, which still relies on external markers to establish a writer's voice.

In contrast, the functional distribution in ECS is characterized by a strong focus on research-oriented functions. Among all high-frequency clusters, as many as 14 serve this function. These clusters go beyond basic description and are primarily used to name and refer to core concepts, theoretical models, research variables, and measurement indices within the discipline. For example, *the viability constraint set* and *the mixed effects models*. This indicates that expert writing is highly content-driven. The language directly points to the objective world of research and theoretical construction, through which disciplinary knowledge is conveyed. Most notably, participant-oriented clusters are entirely absent from the expert high-frequency list. This suggests that authorial identity and academic authority in expert writing are internalized within disciplinary discourse. They are subtly constructed through fluent use of technical terminology, rigorous presentation of evidence, and precise lexical choices, rather than through explicit self-mention or directive statements.

## V. Conclusions

This study conducted a comparative analysis of two self-compiled corpora in the field of sociolinguistics. It systematically examined the structural and functional use of four-word lexical clusters, with the aim of uncovering the features of academic discourse construction across different proficiency levels. The findings clearly demonstrate that expert and student writers exhibit systematic and significant differences in their selection and use of lexical clusters. These differences reflect deeper distinctions in their levels of academic discourse proficiency, cognitive focus, and discourse construction strategies. These findings are expected to offer significant pedagogical implications for academic writing instruction, particularly within sociolinguistics. By highlighting specific structural and functional differences, educators can develop targeted interventions. This could involve creating corpus-informed teaching materials and consciousness-raising activities that focus on high-frequency, discipline-specific lexical clusters. Such an approach can help students bridge the gap between their current writing practices and the conventions of expert academic discourse in sociolinguistics, ultimately enhancing their communicative competence.

This study also has some limitations. Firstly, although the corpus sizes were balanced, the representativeness of the BAWE corpus for student writing in sociolinguistics needs further consideration, as variation in student proficiency and sub-disciplinary focus may affect the results. Secondly, since the analysis was confined to the field of sociolinguistics, the findings may not be directly generalizable to other academic disciplines. Future research could therefore expand the scope to include a wider range of subject areas and conduct cross-disciplinary comparisons.

**Reference:**

[1] Biber, D., Johansson, S., Leech, G., et al. The Longman grammar of spoken and written English[M]. London: Longman, 1999: 990-992.

[2] Cortes, V. Lexical bundles in published and student disciplinary writing: Examples from history and biology[J]. English for Specific Purposes, 2004(23).

[3] Farhang-Ju M, Jalilifar A, Keshavarz M H. Specificity and generality of lexical bundles in the rhetorical moves of Applied Linguistics research article introductions[J]. Journal of English for Academic Purposes, 2024, 69: 101387.

[4] Hyland, K. Authority and invisibility: Authorial identity in academic writing[J]. Journal of Pragmatics, 2002(8): 1091-1112.

[5] Hyland K. Academic clusters: Text patterning in published and postgraduate writing[J]. International journal of applied linguistics, 2008, 18(1): 41-62.

[6] Jiang Lei, Kang Mengchao, Xiao Yao. Structural and Functional Features of Lexical Bundles Used by English Learners at Different Proficiency Levels[J]. Journal of Northeastern University (Social Science), 2024,26(03):127-136.

[7] Li Mengxiao, Liu Yongbing. A comparative corpus-based study of lexical bundles used in Chinese and native expert academic writings[J]. Modern Foreign Languages (Bimonthly), 2016,39(04):507-515+584.

[8] Li Xiaohong. Functional Features of Phrase Frames in EAP Writings——A Corpus-Based Contrastive Study of Chinese and English PhDs' Dissertation Introductions[J]. Technology Enhanced Foreign Language Education, 2021,(01):98-104+16.

[9] Li Yan, Jiang Yajun. On the Structures and Functions of Lexical Bundles in English Academic Discourse by Chinese and L1-English Writers[J]. Technology Enhanced Foreign Language Education, 2024,(06):45-53+111.

[10] Shin Y K, Won D O. To what extent do L2 learners produce genre-appropriate language? A comparative analysis of lexical bundles in argumentative essays and speeches[J]. Journal of English for Academic Purposes, 2024, 69: 101389.

[11] Yang Chuanming, Guo Yunjie, Leng Litian. Comparative Study on Lexical Bundles in AI-Generated and Scholar-Written Abstracts of Chemistry Journal Articles[J]. Chinese Journal of Chemical Education,2025,46(08):93-102.