

A Corpus-Based Comparative Study of Four-Word Lexical Bundles in Expert and Student Sociological Writing

Wenhui Wang

School of Foreign Languages, Beijing University of Technology, Beijing, China

Abstract: Lexical bundles are the basic building blocks of academic discourse, but their use in the field of sociology has not yet received sufficient attention from the academic community. This study compares the structural and functional characteristics of four-word lexical bundles in two self-compiled corpora of sociological academic writing, one produced by expert writers and the other by student writers. The expert corpus consists of 10 academic texts totaling 52,345 tokens, while the student corpus contains 20 academic texts with 51,947 tokens. Using frequency and dispersion as identification criteria, 201 bundle types are retrieved from the expert corpus, compared with 58 from the student corpus.

Structural analysis shows that noun phrase-based bundles are the dominant type in both corpora, but they occur more extensively in expert writing, indicating a stronger tendency toward compressed informational expressions. By contrast, student writers rely more heavily on prepositional phrase-based bundles. Functional analysis reveals that expert writing contains more research-oriented bundles, especially descriptive ones, whereas student writing features more text-oriented and participant-oriented bundles. The results indicate that student writers have not yet developed stable use of some formulaic patterns typical of published sociology articles. These findings have important pedagogical implications for the instruction of academic writing in sociology.

Keywords: Lexical Bundles; Academic Writing; Sociology; Structural Features; Functional Features

1. Introduction

Academic papers constitute a pivotal academic genre, serving as the primary site for disciplinary knowledge construction [1] and a paradigmatic

medium for scholarly communication [2]. Lexical bundles, defined as recurrent multi-word sequences distributed across a substantial number of texts within a given register [3], serve as foundational building blocks for constructing academic discourse [4]. The use of lexical bundles in academic texts can improve the precision and idiomaticity of language expression, signaling the internal status of the author in a specific language community [4]. Accordingly, research into lexical bundles is conducive to helping novice writers to strengthen their academic writing proficiency.

There has been multiple research exploring the application of lexical bundles across different scholarly registers [5,6], disciplines [7-9], and writer groups [10-12]. However, discipline-specific research remains uneven, and sociology has received far less attention than fields such as applied linguistics or science-related disciplines. Given that the deployment of lexical bundles varies across disciplines [9], findings from other fields cannot be simply generalized to sociology.

Against this research backdrop, the present study constructs two self-compiled corpora of academic English papers to systematically compare the structural and functional features of four-word lexical bundles adopted by expert writers and student writers in the field of sociology. It is anticipated that the research findings can provide targeted pedagogical references for sociological academic writing teaching, facilitate the optimization of novice writers' lexical bundle application strategies, and ultimately improve sociological students' academic English writing proficiency.

2. Literature Review

2.1 The Identification and Classification of Lexical Bundles

Lexical bundles are generally defined as recurrent word sequences characterized by high frequency of occurrence [3]. Frequency is

widely regarded as a fundamental criterion for lexical bundle identification. Given that corpora vary considerably in size, previous studies have adopted different frequency thresholds accordingly. In large written corpora, the standardized frequency threshold is always 20 to 40 times per million words [6,9]; whereas in relatively small corpora, researchers mostly use the original cut-off frequency, about two to ten times [13,14]. In addition to frequency, dispersion, which refers to the number of text where a given word sequence appears, is also a key indicator for extracting lexical bundles. Previous studies have operationalized the dispersion threshold in different ways, such as requiring occurrence in 3 of 82 texts [15], 20 of 176 texts [5], or 10% of the total number of texts [4,9,16].

Once extracted, the target bundles are conventionally taxonomized along both structural and functional dimensions. Along the structural dimension, Biber et al. proposed a three-part lexical bundle classification, taken from university teaching discourse, covering classroom speeches and textbooks. This structure divides the lexical bundles into three categories: those based on verb phrases (VP), those with dependent clauses, and those based on nouns or prepositional phrases (NP/PP) [6]. According to the original structure of Biber, Pan et al. taxonomized formulaic sequences in research articles into NP-, PP-, and VP-anchored bundles, effectively combining Biber's first two structural types into a single category [11,17]. In terms of functional classification, there are mainly two classification methods. Drawing on the university teaching register, including classroom lectures and academic textbooks, Biber et al. developed a tripartite functional taxonomy, delineating bundles into stance expressions, discourse organizers, and referential markers [6]. Hyland devised another classification system to explore the function of lexical bundles within academic texts. In his taxonomy, lexical bundles are classified into three functional types: research-driven, text-driven, and participant-driven bundles. Specifically, research-driven bundles function to delineate investigative practices and empirical realities; text-driven bundles help structure the discourse and develop meaning and argumentation; and participant-driven bundles express writer stance and engage readers. Each of these broad functional groups can be further

divided into more specific subcategories [9]. Considering that the present study focuses specifically on academic writing, the functional annotation of the retrieved bundles was operationalized through the lens of Hyland's taxonomy [9].

2.2 Research on the Experts' and Students' Use of Lexical Bundles

A robust body of literature has juxtaposed the lexical bundle repertoires of student writers with those of published experts within scholarly texts. Cortes examined four-word lexical bundle use among published authors and students at three different proficiency levels across history and biology disciplines [7]. The analysis uncovers a distributional asymmetry that the vast majority of bundles favored by established academics are strikingly underrepresented in student texts. Hyland further compared lexical bundle patterns between doctoral dissertations and master's theses [4]. Findings point to a clear divergence between the two kinds of discourse, with doctoral discourse primarily prioritizing research-oriented bundles, as opposed to the prominent reliance on text- and participant-oriented bundles observed in master's theses. Chen and Baker contrasted four-word lexical bundle deployment in the academic writing of experts and students [18]. Structurally, their findings indicated that writers with higher linguistic proficiency employed more NP- and PP-based bundles but fewer VP-based bundles than those with lower writing ability. Functionally, student writers relied more heavily on discourse-organizing bundles while using far fewer referential bundles in their scholarly texts.

While previous studies have compared the use of lexical bundles between student writers and expert writers, research on lexical bundles specific to certain disciplines remains scarce, and there is almost no research on lexical bundles in the field of sociology. Therefore, this paper seeks to contrast how lexical bundles are utilized by student writers and expert writers within the sociological discipline. This comparative analysis is anchored by the subsequent core queries:

1. What structural similarities and differences exist in the four-word lexical bundles used by expert and student writers in sociology?
2. What functional similarities and differences exist in the four-word lexical bundles used by

expert and student writers in sociology?

3. Methodology

3.1 Data Collection

This study constructed two self-built small-scale corpora of English academic papers: an expert corpus and a student corpus. The expert corpus drew its textual data from the corpus embedded in AntCorGen, consisting of 10 randomly selected English academic papers within the discipline of sociology. The student corpus extracted data from the British Academic Written English Corpus (BAWE, (ox.ac.uk)). With the BAWE corpus downloaded, 20 sociology-focused English academic essays were randomly sampled for analysis.

Table 1. Corpus Statistics

	Expert	Student
The number of papers	10	20
Word tokens	52345	51947
Word types	4828	6373

Only the main body of each paper was retained as valid data; titles, abstracts, keywords, figures, tables, footnotes, references, acknowledgements, appendices were entirely excluded. Following format conversion, textual sorting and manual

proofreading, two final corpora were established: the expert corpus comprising 52,345 tokens and the student corpus comprising 51,947 tokens. The quantitative profile of the two corpora is delineated in table 1.

3.2 Research Procedures

First, this study defined the target research unit and the extraction criteria for lexical bundles. Four-word lexical bundles were chosen as the focal object of analysis since they appear more frequently than five-word bundles and convey richer contextual information than three-word bundles [4]. A frequency threshold of four occurrences or above was applied to recognize four-word lexical bundles. A distribution threshold of 10% of all texts was also implemented to eliminate idiosyncratic word sequences unique to individual authors.

Second, lexical bundles were extracted and preprocessed. With AntConc (version 4.3.1), the extraction of lexical bundles across both corpora was operationalized, strictly adhering to the pre-defined criteria. Redundant overlapping bundles and sequences overly specific to the core research themes of individual papers were subsequently eliminated from the extracted lexical bundles.

Table 2. Structural Categorization Scheme for Lexical Bundles by Xiaofei Lu and Jinlei Deng [19]

Category	Subcategory	Instance
VP-anchored	Linking verb be governing a noun phrase/adjective phrase	is essential for a
	Active-voice VP fragments	serve as an adjuster
	Infinitive-led VP	to examine the relationship
	Passive-voice VP fragments	should be explained by
	Past participle-headed clusters	compared with previous studies
Clause-anchored	PP followed by linking verb be	for this reason is
	NP followed by linking verb be	the purpose is to
	Dummy it followed by linking verb be and adjective phrase	it is important to
	Dummy it followed by passive verb and that	it is assumed that
	Passive verb preceded by NP or complementizer	much is known about
	NP governing an active verb	the study examines the
NP-anchored	NP governing an active verb and that	the chart reveals that
	NP modified by an of-phrase	the change of the
	NP followed by other post-modifiers	the effect on the
PP-anchored	Other NP	a robust body of
	PP integrating an of-phrase	with the habit of
Others	Other PP fragment	in relation to the
	/	As long as the

Third, the study classified the obtained lexical bundles from the aspects of structure and function, and analyzed the classification consequences. In the present study, the identified

lexical bundles are structurally classified based on the taxonomy developed by Xiaofei Lu and Jinlei Deng [19] (see Table 2), which was adapted from the structural classification

frameworks employed by Biber et al. [6] and Pan et al. [17]. Xiaofei Lu and Jinlei Deng divided the structures of lexical bundles into four types, namely NP-based, PP-based, VP-based and clause-based bundles. The functions of the lexical bundles are analyzed based on Hyland's

[9] functional taxonomy (see Table 3). Categorizing the four-word lexical bundles retrieved from both corpora, this study analyzed similarities and differences in bundle use between expert and student writers in sociology.

Table 3. Functional Categorization Scheme for Lexical Bundles by Hyland [9]

Category	Subcategory	Instance
Research-driven clusters	Spatial/ temporal markers	by the end of
	Procedural descriptions	by means of a
	Quantification expressions	a significant volume of
	Property description	the degree to which
	Subject markers	the social system
Text-driven clusters	Markers of transition	in comparison with a
	Markers of result	as a consequence of
	Text-structuring devices	as indicated in the
	Framing devices	in the case of
Participant-driven clusters	Indicators of stance	it is inconvincible that
	Engagement devices	if we look at

4. Results and Discussion

4.1 General Characteristics of Four-word Lexical Bundles

Table 4 presents the overall distribution of four-word lexical bundles retrieved from expert and student papers. As illustrated in the table, expert writers produce higher type counts and token counts of lexical bundles than student writers.

Per-million-word frequencies of lexical bundles were calculated for both corpora via the token normalization formula ($f = \frac{\text{bundles tokens}}{\text{corpus size} \times 1,000,000}$). The normalized results demonstrate that expert writers deployed four-word lexical bundles more frequently in their writing. This finding reveals that expert writers draw more extensively on recurrent formulaic sequences to construct academic discourse.

Table 4. General Characteristics of Four-Word Lexical Bundles

	Expert	Student
Bundle types	201	58
Bundle tokens	1205	319
Per-million-word frequency	23020	6140

4.2 Comparison of Structural Characteristics of Four-Word Lexical Bundles

Table 5 presents the structural distribution of four-word lexical bundles in the two corpora. Both groups exhibit the same overall structural preference that NP-based bundles rank first in

both corpora, followed by PP-based bundles, which suggests phrasal constructions constitute the primary structural foundation of formulaic language in sociological academic writing, regardless of writer expertise.

However, the two groups differ considerably in the proportional distribution across categories. The most prominent difference concerns NP-based bundles, which account for 50.2% in the expert corpus but only 39.7% in the student corpus. This result is similar to the discovery reported by Chen and Baker [18], wherein NP-based bundles are significantly more prevalent among highly proficient cohorts than in less skilled learners. As Halliday observed, noun phrase bundles condense more information within an equal word span than clausal structures do, effectively expressing common actions and events in academic discourse [20]. The expert writer's greater reliance on NP-based bundles thus reflects a more mature command of information compression, enabling them to convey dense content in compact forms characteristic of academic writing. By contrast, the proportion of PP-based bundles used by student writers is higher than that of expert writers. This finding diverges from that of Chen and Baker, who found that higher-proficiency writers exhibit a greater propensity for PP-based structures [18]. This discrepancy may stem from disciplinary differences, as sociological writing may impose distinct rhetorical demands on formulaic language use.

4.3 Comparison of Functional Characteristics of Four-Word Lexical Bundles

Table 6 presents the distribution and proportion of lexical bundles in different functional categories of the two corpora. As shown in the table, research-oriented bundles govern the functional distribution, accounting for more than half of the distinct bundle types in each dataset. Most bundles of this type in both corpora serve to delineate research subjects and contextual backgrounds, which is in line with the argumentative and analytical nature of sociology. Participant-oriented bundles constitute the least frequently used category of both groups, accounting for merely 2% of bundle types in the

expert corpus and 10.4% in the student corpus. This distribution may be subject to the established practices of scholarly discourse, which restricts the author from relying on the bundles of expressing subjective opinions.

Table 5. Distribution of Bundle Structure

Category	Types of Expert Corpus		Types of Student Corpus	
	N	%	N	%
NP-based	101	50.2%	23	39.7%
PP-based	60	29.9%	22	37.9%
VP-based	22	10.9%	5	8.6%
Clause-based	17	8.5%	7	12.1%
Conjunctions	1	0.5%	1	1.7%
Total	201	100%	58	100%

Table 6. Distribution of Bundle Functions

	Subcategories	Expert	Student		
Research-oriented bundles	Location	5% (10)	91.5% (184)	6.9% (4)	72.4% (42)
	Procedure	3.5% (7)		0% (0)	
	Quantification	0% (0)		1.7% (1)	
	Description	70.6% (142)		27.6% (16)	
	Topic	12.4% (25)		36.2% (21)	
Text-oriented bundles	Transition signals	1.5% (3)	6.5% (13)	1.7% (1)	17.2% (10)
	Resultative signals	0.5% (1)		1.7% (1)	
	Structuring signals	2.0% (4)		5.2% (3)	
	Framing signals	2.5% (5)		8.6% (5)	
Participant-oriented bundles	Stance features	2.0% (4)	2.0% (4)	5.2% (3)	10.4% (6)
	Engagement features	0% (0)		5.2% (3)	
Total		100% (201)		100% (58)	

Despite these similarities, notable functional differences exist in bundle employment between the two groups. Expert writers display a notably higher concentration of research-oriented bundles (91.5%) than student writers (72.4%), with description bundles alone accounting for 70.6% of all expert bundle types. This finding indicates that expert writers command a considerably richer repertoire of formulaic expressions for characterizing research subjects and contextual backgrounds. Compared with expert writers, student writers exhibit a much higher density of bundles serving text-oriented and participant-oriented functions. Specifically, framing signals account for 8.6% of the student corpus while only 2.5% of the expert corpus; and the engagement markers, entirely absent in the expert corpus, are present in the student corpus at 5.2%. These results suggest that student writers rely more on explicit discourse organizers and stance-marking expressions as compensatory strategies in academic writing, while expert writers have more thoroughly internalized discipline-specific formulaic

language to construct research content. This finding is generally consistent with Hyland's research, which pointed out that expert writers favored research-oriented bundles, while less experienced writers showed a greater reliance on text-oriented and participant-oriented bundles [4].

5. Conclusion

This study explored how four-word lexical bundles are utilized by learner and professional authors in the sociological domain. The analytical results reveal a detailed picture of both convergence and divergence in four-word lexical bundle usage between the two groups of writers. Structurally, NP-based bundles occupy the largest proportion in both corpora, confirming that phrasal constructions constitute the primary structural foundation of formulaic language in sociological academic writing. However, expert writers utilize a considerably higher proportion of NP-based bundles, demonstrating their more mature command of information compression, whereas student writers tend to gravitate toward

PP-based and clause-based structures. Functionally, research-oriented bundles predominate in both corpora, consistent with the argumentative and analytical nature of sociology. Nevertheless, expert writers exhibit a substantially greater density of research-oriented bundles, particularly description bundles, indicating their richer repertoire of discipline-specific formulaic expressions. By contrast, student writers deploy a greater proportion of text-oriented and participant-oriented bundles, which points to their intensified dependence on overt metadiscoursal devices and stance markers as compensatory strategies.

These empirical observations can provide an actionable inspiration for the instruction of academic writing. Specifically, writing instructors in sociology can implement tailored pedagogical interventions to cultivate learners' metalinguistic awareness of NP-based bundles used for information condensation. At the same time, such instruction should mitigate learners' excessive dependence on text-oriented bundles, steering them toward a more normative, discipline-specific integration of research-oriented bundles.

There are several limitations in this study. First, the relatively small scale of the two corpora may constrain the generalizability of the research findings. Therefore, future studies can employ larger and more diversified datasets to enhance representativeness. Second, the current comparison is confined to the structural and functional levels; future studies can integrate qualitative discourse analysis or interview data to explore the rhetorical and cognitive motivations underlying writers' lexical bundle choices.

References

- [1] Hyland, K. *Academic Discourse: English in a Global Context*. New York: Continuum, 2009.
- [2] Duszak, A. Cross-cultural academic communication: A discourse community view. In A. Duszak(ed.). *Culture and Styles of Academic Discourse*. New York: Mouton de Gruyter, 1997:11-39.
- [3] Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *Longman grammar of spoken and written English*. London: Longman.
- [4] Hyland, K. *Academic clusters: Text patterning in published and postgraduate writing*. *International Journal of Applied Linguistics*, 2008a, 18(1):41-62.
- [5] Biber, D., & Barbieri, F. (2007). Lexical bundles in university spoken and written registers. *English for Specific Purposes*, 26(3), 263-286.
- [6] Biber, D., Conrad, S., & Cortes, V. (2004). If you look at: Lexical bundles in university teaching and textbooks. *Applied Linguistics*, 25(3), 371-405.
- [7] Cortes, V. (2004). Lexical bundles in published and student disciplinary writing: Examples from history and biology. *English for Specific Purposes*, 23(4), 397-423.
- [8] Durrant, P. (2017). Lexical bundles and disciplinary variation in university students' writing: Mapping the territories. *Applied Linguistics*, 38(2), 165-193.
- [9] Hyland, K. As can be seen: Lexical bundles and disciplinary variation. *English for Specific Purposes*, 2008b, 27(1):4-21.
- [10] Li, Y., & Jiang, Y. J. (2024). A contrastive study on chunk structure and pragmatic functions of academic terms in Chinese and English. *Computer-Assisted Foreign Language Education*, (6), 45–53+111.
- [11] Li, Y., & Jiang, F. (2025). Rhetorical step chunk features in the introduction sections of academic papers by Chinese and foreign scholars. *Foreign Languages in China*, 22(4), 61–67+112.
- [12] Li, Y., & Lei, H. (2025). Lexical bundles in L1 and L2 English academic writing: Convergent and divergent usage. *Sage Open*, 2, 1–14.
- [13] Altenberg, B. (1998). On the Phraseology of Spoken English: The Evidence of Recurrent Word-Combinations. In A. P. Cowie (Ed.), *Phraseology: Theory, Analysis, and Applications* (pp. 101-122). Oxford University Press.
- [14] De Cock, S. (1998). A recurrent word combination approach to the study of formulae in the speech of native and non-native speakers of English. *Journal of Pragmatics*, 30(1), 51-81.
- [15] Adel, A., & Erman, B. (2012). Recurrent word combinations in academic writing by native and non-native speakers of English: A lexical bundles approach. *English for Specific Purposes*, 31(2), 81-92.
- [16] Perez-Llantada, C. (2014). Formulaic language in L1 and L2 expert academic

- writing: Convergent and divergent usage. *Journal of English for Academic Purposes*, 14, 84-94.
- [17] Pan, F., Reppen, R., & Biber, D. (2016). Comparing patterns of L1 versus L2 English academic professionals: Lexical bundles in telecommunications research journals. *Journal of English for Academic Purposes*, 21, 60–71.
- [18] Chen, Y. H., & Baker, P. (2010). Lexical bundles in L1 and L2 academic writing. *Language, Learning and Technology*, 14(2), 30-49.
- [19] Lu, X., & Deng, J. (2019). With the rapid development: A contrastive analysis of lexical bundles in dissertation abstracts by Chinese and L1 English doctoral students. *Journal of English for Academic Purposes*, 39, 21–36.
- [20] Halliday, M. A. K. *Spoken and Written Language*. Oxford: Oxford University Press, 1989.