Comparative Analysis of Lexical Bundles in Dissertation Abstracts: Insights for Teaching Academic English to Chinese Students

Kai Bao

Correspondence: Kai Bao, College of Foreign Languages, University of Shanghai for Science and Technology, 516 Jungong Rd., Shanghai, China.

Abstract

Lexical bundle research in academic abstracts has predominantly focused on research articles, with less attention given to dissertation abstracts. This is particularly relevant for Chinese graduate students who are required to provide English abstracts in their dissertations. Addressing this gap, the study compared the structural and functional distribution of lexical bundles in dissertation abstracts by linguistics students from China and the United States to inform academic instruction. Two corpora, the Chinese University Student Collection and the American University Student Collection, each with 700 abstracts, were compiled and analyzed. The findings showed that Chinese students proportionally used more noun phrase (NP) and prepositional phrase (PP)-based lexical bundles, but fewer verb phrase (VP)-based ones, compared to their American counterparts. Additionally, they used a higher proportion of research- and participant-oriented bundles, but fewer text-oriented bundles. These differences highlight distinct structural and functional preferences in lexical bundle usage between the two student groups. This study underscores the importance of adapting instructional strategies to address these differences, enhancing English academic writing skills of Chinese graduate students by acknowledging the diverse linguistic approaches of international student populations.

Keywords: lexical bundles, dissertation abstracts, Chinese graduate students, English academic writing, instructional strategies

1. Introduction

Formulaic language constitutes a significant portion of our everyday language usage (Schmitt and Carter, 2004). These multi-word items, which are stored and retrieved as complete units from our memory, play a critical role in both spoken and written forms of communication. Research by Erman and Warren (2000) quantifies their presence, revealing that formulaic language comprises about 58.6% of spoken English and 52.3% of written English. This notable prevalence emphasizes that thet are "important building blocks of discourse in spoken and written registers" (Biber and Barbieri, 2007, p. 263). Formulaic language competence is increasingly acknowledged as a crucial aspect of language proficiency. This notion stems from the idea that skilled language usage within a particular register often requires proficiency in specific fixed phrases (Cortes, 2004). In academic writing, the effective use of discipline-specific formulaic sequences is considered indicative of a writer's integration into a specific discourse community (Ädel and Erman, 2012). Additionally, learning these formulaic sequences can substantially boost learners' understanding of different genres. Coxhead and Byrd (2007) emphasize that these sequences offer learners pre-constructed sets of phrases for immediate use, simplifying the process of language production. Hunston (2002, p. 174) further observes that formulaic language facilitates the expression of complex ideas with "a single mental effort," enhancing fluency and coherence. Given these benefits, formulaic language is highly valuable in the realm of English for Academic Purposes (EAP) instruction. Its significant role in aiding language acquisition and improving communicative competence in academic settings highlights its essential place in language education. This emphasis on formulaic language in EAP underscores the need for instructional strategies that incorporate these key linguistic components to better equip learners for academic success.

As an important type of formulaic language, lexical bundles are "recurrent expressions, regardless of their idiomaticity, and regardless of their structural status" (Biber et al., 1999, p. 990), and are characterized by their "non-idiomaticity, structural incompleteness, and frequency-driven identification" (Bao and Liu, 2022, p. 2). An example of a lexical bundle commonly used in dissertation abstracts is *I show that*, a sequence frequently employed

by American PhD students in linguistics dissertation abstracts to report research findings. Traditional phraseology often overlooks such sequences due to their grammatical incompleteness. However, lexical bundle research bridges this gap, capturing linguistic elements that conventional approaches might miss. This extension to include less traditionally acknowledged items underlines the comprehensive and nuanced nature of lexical bundle studies in understanding language use. The existing literature underscores the significant presence of lexical bundles in various contexts (Biber et al., 1999; Coxhead and Byrd, 2007; Hyland, 2008). Studies have documented their register-specificity (Biber et al., 1999; Biber and Barbieri, 2017; Huang, 2018), genre-specificity (Hyland, 2008), and discipline-specificity (Cortes, 2004), as well as variations in their usage among different writer groups (Bao and Liu, 2022, 2023; Chen and Baker, 2010; Lu and Deng, 2019). These findings illuminate the diverse applications of lexical bundles in academic discourse, revealing their adaptability and significance across various fields and styles of writing. In addition, empirical studies have emphasized the vital role of formulaic language in language learning and proficiency enhancement (Yu, 2022). A strong link has been identified between high levels of language competence and the effective use of lexical bundles (Kim and Kessler, 2022). This evidence underscores the importance of lexical bundles not only in EAP writing but also in the broader context of language acquisition and proficiency.

The scope of research on lexical bundles in academic writing has largely focused on those found in research article abstracts, leaving a notable gap in the study of dissertation abstracts (Bao, 2022; 2023; Lu and Deng, 2019). This gap is significant as research article abstracts and dissertation abstracts are two distinct genres, each characterized by their own rhetorical structures and linguistic styles (El-Dakhs, 2018). This distinction suggests that the lexical bundles used in these genres may differ in important ways. In the context of Chinese higher education, there has been a remarkable increase in postgraduate admissions over recent years. The year 2023 marked a milestone, with graduate admissions surpassing 1.2 million, doubling from the 0.6 million recorded in 2013. For a majority of these students, particularly in research-intensive disciplines, providing an English abstract in their dissertations is a crucial academic and graduation requirement. These abstracts serve not only as a showcase of their scholarly work but also as a critical element for meeting the academic standards of their programs.

Given the importance of English dissertation abstracts in Chinese academic settings, this study specifically focuses on dissertation abstracts in the discipline of linguistics. The choice of linguistics as the focal discipline is strategic, as it bridges the realms of the humanities and the sciences. It incorporates the theoretical and analytical approaches typical of the humanities, while also embracing empirical methodologies, such as experiments and data analysis, commonly found in the sciences. This unique amalgamation in linguistics allows for a diverse range of linguistic phenomena to be studied from multiple perspectives, making it an ideal discipline for this research. The intersection of humanities and scientific elements in linguistics means that the lexical bundles found in this field's discourse are likely to be applicable and relevant to a wide array of academic disciplines. By identifying structural and functional differences in the use of lexical bundles between Chinese and American linguistics students, the research aims to inform instructional strategies that can minimize the gap between these groups. This approach is particularly relevant given the increasing globalization of academic discourse. Understanding these differences is pivotal for developing teaching methods that encourage foreign language learners to adopt lexical bundles commonly used within the international academic community. Such an approach is vital in promoting greater alignment with international academic standards and practices. By familiarizing foreign language learners with the patterns and practices prevalent among their international peers, educators can help them integrate more seamlessly into the broader scholarly community. Such an approach is invaluable for Chinese students, who often face the challenge of meeting international academic language standards, particularly in English language proficiency. The insights gained from this comparative analysis will provide educators with the tools to tailor their teaching to better support these students, enhancing their ability to produce abstracts that resonate more effectively with an international audience.

2. Research Methodology

2.1 Corpus

Table 1 provides a detailed overview of the data collection methodology for this research. For the creation of the Chinese University Student Collection (CUSC), 700 dissertation abstracts were collected from students in the disciplines of Foreign Linguistics and Applied Linguistics at Chinese universities. These abstracts were sourced from several repositories, including the China National Knowledge Infrastructure, the National Library of China, and select university libraries. The timeframe for these abstracts was carefully chosen to span from 2000 to 2020 for two primary reasons: It ensured the collection of a sufficient volume of sample texts and coincided with the foundational period of many PhD programs in these disciplines in Chinese universities, offering a representative and meaningful cross-section of the academic landscape. In parallel, the American University Student Collection (AUSC) was

developed by collecting 700 dissertation abstracts from students in the field of linguistics at American universities. These were sourced from ProQuest and, to enhance the scope and diversity of the collection, from university libraries as well. The abstracts were selected from the same two-decade span, 2000 to 2020, to ensure methodological consistency and comparability between the CUSC and AUSC. This approach was designed to enable a nuanced comparison of the lexical and structural conventions in dissertation abstracts across these two distinct academic cultures, capturing the evolution and trends in academic writing practices within the field of linguistics during this period.

Table 1. Overview of the compiled corpus

	CUSC	AUSC
Total number of analyzed texts	700	700
Publication year span	2000-2020	
Field of study for abstracts	Foreign Linguistics and Applied Linguistics	Linguistics
Average word count per text	876.73	353.37
Aggregate token count	613,713	247,359

As indicated in Table 1, there are notable differences in the average word count per text and the aggregate token count for the two collections. For the CUSC, the average word count per abstract is significantly higher at 876.73 words, compared to the AUSC, which averages at 353.37 words per abstract. This stark contrast not only reflects differing conventions and expectations in dissertation abstract length between the two academic cultures but may also indicate variations in the depth of detail and complexity of content that students are expected to include in their abstracts. Furthermore, the aggregate token count for the CUSC stands at 613,713, which is considerably larger than the AUSC's count of 247,359 tokens. This substantial difference in the volume of language used could suggest that Chinese students are employing a wider range of lexical items and possibly more complex syntactic structures in their writing. This has implications for the types of lexical bundles that might be present in the abstracts, potentially affecting the density and variety of formulaic language. These quantitative distinctions provide an additional layer of context for analyzing the structural and functional distribution of lexical bundles in the abstracts. Understanding these metrics is important for informing instructional strategies as they offer insights into the norms of academic writing within the linguistic discipline for both Chinese and American educational settings. It also raises questions about the potential impact of cultural and educational factors on the production of written academic discourse, which could be a valuable area for further investigation.

2.2 Data Analysis

Wordsmith Tools 8.0 (Scott, 2020) was employed to identify three-word lexical bundles within the corpus, with a set frequency threshold of 60 occurrences per million words across at least 2% of the abstracts. This focus on three-word bundles aligns with the methodology of preceding research in the field (Azad and Khiabani, 2018; Bao and Liu, 2022, 2023) and acknowledges the higher frequency of occurrence of three-word lexical bundles compared to those with four or five words. Such a focus is intended to capture a broader range of commonly used lexical bundles. The dispersion criterion of 2% is consistent with Bao and Liu's (2022, 2023) findings, which stress the importance of dispersion for researching lexical bundles in the typically succinct genre of abstracts. By setting this rate, the study aims to exclude author-specific bundles and ensure that the remaining lexical bundles are employed by multiple authors, enhancing their validity as commonly recognized and used phrases within the academic community. Additionally, overlapping lexical bundles were merged to reduce the impact of frequency biases, a methodological consideration informed by Chen and Baker (2010). This approach was crucial in yielding a refined list of lexical bundles, resulting in the identification of 274 bundles from the CUSC and 195 bundles from the AUSC.

The structures of lexical bundles were classified using Biber et al.'s (1999) structural taxonomy, originally designed to describe lexical bundles in academic prose (Table 2). To tailor this model more precisely to the genre of dissertation abstracts, several key modifications were made: 1) The category "subject + verb phrases + (that-clause)" was introduced, supplanting the original "noun + verb phrases + that-clause" to better reflect the structures encountered in dissertation abstracts; 2) The categories "other noun phrase fragment" and "other verb phrase fragment" were added. These new categories capture those lexical bundle structures that did not fit neatly into the model's original categories, ensuring a more comprehensive classification; 3) Furthermore, the model was refined by promoting the three sub-categories "predicative adjective + to-clause", "(passive) verb phrase + to-clause", and "to-clause" from their overarching category "(verb/adjective) + to-clause fragment". This change was implemented

to provide more detailed classification for the considerable number of lexical bundles that corresponded to these three specific structures.

Table 2. Biber et al.'s (1999) structural model of lexical bundles

Category	Structure	Example
NP-based	Noun phrase + of	*the use of*
	Noun phrase with other post modifier	*the relationship between*
	Other noun phrase fragments	*Second Language Acquisition*
PP-based	Prepositional phrase + of	*in terms of*
	Other prepositional phrase	*in this dissertation*
VP-based	Be + noun/adjective phrase	*are consistent with*
	Passive verb + prepositional phrase	*based on the*
	Anticipatory *it* + verb/adjective phrase	*it is argued that*
	Verb phrase + *that*-clause	*show that the*
	Subject + verb phrase + (that-clause)	*I argue that*
	Predicative adjective + to-clause	*due to the*
	(Passive) verb phrase + to-clause	*are shown to*
	To-clause	*to account for*
	Adverbial clause fragments	*as opposed to*
	Pronoun/noun phrase + be	*this dissertation is*
	Other verb phrase fragments	*focuses on the*
Other structures	N/A	*as well as*

Note: Bold font indicates sub-categories introduced by this study.

The lexical bundles were then categorized according to Hyland's (2008) functional model, which was specifically tailored for lexical bundles in research articles and dissertations (Table 3). To adapt this model more precisely to the genre of dissertation abstracts, several adjustments were made:1) Two new sub-categories, relationship signals and objective signals, were introduced within the text-oriented category to capture additional functional nuances found in the corpus; 2) Inferential signals and causative signals, previously nested under resultative signals, were elevated to stand as direct sub-categories within the text-oriented category, reflecting their distinct roles; 3) An other bundles category was added to encompass lexical bundles that did not align neatly with the existing functional categories. Recognizing that functional categorization can be subjective, efforts were made to minimize potential bias by enlisting a linguist with a master's degree to independently categorize the frequently used bundles alongside me. To quantify the level of agreement between our categorizations, we applied the Cohen's Kappa coefficient, adhering to the benchmark scale set by Landis and Koch (1977). The resulting Kappa coefficient of 0.889 ($k > 0.80$) indicated an excellent level of inter-coder agreement, attesting to the reliability of the categorization process. For those lexical bundles where functions were initially disputed, the independent coder and the author engaged in thorough discussions of each item until we reached unanimous agreement. This collaborative approach ensured a robust and consensus-driven categorization, further enhancing the validity of the functional framework applied to the lexical bundles in this study.

Table 3. Hyland's (2008) functional model of lexical bundles

Category	Function	Example
Research-oriented	Location	*the course of*
	Procedure	*analysis of the*
	Quantification	*a number of*
	Description	*the use of*
	Topic	*Second Language Acquisition*
Text-oriented	Transition signals	*as well as*
	Inferential signals	*I argue that*
	Causative signals	*the results of*
	Structuring signals	*in this dissertation*
	Framing signals	*in terms of*
	Relationship signals	*the relationship between*
	Objective signals	*in order to*
Participant-oriented	Stance feature	*the importance of*
	Engagement features	*our understanding of*
Other functions	N/A	*is not a*

Note: Bold font indicates sub-categories introduced by this study.

3. Results and Discussion

3.1 Structural Distribution

Table 4 details the structural distribution of lexical bundles in dissertation abstracts by Chinese and American linguistics students. Both groups prominently use NP-based lexical bundles, accounting for over 40.00% of their total usage. However, Chinese students preferentially use NP-based structures, with a lower usage of VP-based bundles compared to their American counterparts. Both groups use PP-based bundles similarly, comprising over 20.00% of their lexical bundles. In sub-category analysis, both Chinese and American students frequently employ the structure "noun phrase + of", which represents over 30.00% of their lexical bundles. The "other prepositional phrase" structure follows, making up nearly 20.00% of the bundles for both cohorts. These particular bundles often function to identify quantities, locative relations, and characteristics of entities, suggesting their importance in crafting linguistics dissertation abstracts (Hyland, 2008). Both cohorts also display comparable usage in structures such as "noun phrase with other post modifier" and "pronoun/noun phrase + be". However, significant differences emerge in categories like "other noun phrase fragment", "subject + verb phrase + (that-clause)", and "passive verb + prepositional phrase". For instance, "other noun phrase fragment" is more prevalent in the CUSC (12.46%) than in the AUSC (2.17%). Conversely, "subject + verb phrase + (that-clause)" is used more by American students (12.00%) than Chinese students (1.53%), and "passive verb + prepositional phrase" appears more in CUSC (5.48%) than AUSC (2.48%). These contrasts could be attributed to the Chinese students' higher use of topic bundles, their avoidance of the first-person singular *I*, and a preference for passive voice, respectively.

Table 4. Distribution of lexical bundles by structure

Category	CUSC		AUSC	
	Token	Percentage	Token	Percentage
NP-based	**9,633**	**47.80%**	**2,532**	**43.67%**
Noun phrase + of	6,237	30.95%	2,087	36.00%
Noun phrase with other post modifier fragments	885	4.39%	319	5.50%
Other noun phrase fragment	2,511	12.46%	126	2.17%
PP-based	**4,804**	**23.84%**	**1,312**	**22.63%**
Prepositional phrase + of	1,063	5.27%	189	3.26%
Other prepositional phrase	3,741	18.56%	1,123	19.37%
VP-based	**4,699**	**23.32%**	**1,686**	**29.08%**
Be + noun/adjective phrase	172	0.85%	69	1.19%
Passive verb + prepositional phrase	1,105	5.48%	144	2.48%
Anticipatory *it* + verb/adjective phrase	215	1.07%	78	1.35%
Verb phrase + *that*-clause	136	0.67%	0	0.00%
Subject + verb phrase + (*that*-clause)	308	1.53%	696	12.00%
Predicative adjective + to-clause	252	1.25%	49	0.85%
(Passive) verb phrase + to-clause	128	0.64%	98	1.69%
To-clause	705	3.50%	128	2.21%
Adverbial clause fragment	279	1.38%	20	0.34%
Pronoun/noun phrase + be	1,001	4.97%	274	4.73%
Other verb phrase fragment	398	1.97%	130	2.24%
Other expressions	1,018	5.05%	268	4.62%

Note: Bold font represents major categories.

For example, the examination of lexical bundles within CUSC and AUSC reveals distinct approaches to presenting research findings, supporting the observed contrasts in structural preferences. Example 1 from CUSC typifies a topic-centered, impersonal presentation, employing passive voice which deflects focus from the researcher to the research outcomes. This aligns with the conventional, formal academic register often utilized by Chinese students, which emphasizes objectivity and minimizes the author's visible role. Conversely, Example 2 from AUSC reflects a personalized, active voice indicative of American academic writing norms. Here, the explicit use of *I* places the researcher at the forefront, suggesting a direct, individual contribution to the scholarly work. These stylistic differences point to deeper cultural and pedagogical norms that influence academic writing and underline the importance of adapting teaching strategies to foster a balance between impersonal and personal academic discourse, particularly for Chinese students aiming to meet international publication standards. The findings illustrate that, while both groups lean towards a phrasal style in academic prose, indicative of the genre's informational focus, notable differences in their approach to structuring dissertation abstracts exist. These variations point to the need for tailored pedagogical approaches to help Chinese students align more closely with the lexical bundle usage patterns prevalent in international academic writing.

1) ***The major findings*** *of the present research are as follows: 1...2...3...* (CUSC)

2) ***I show that*** *these two constructions again mirror the situation...: I propose that the two constructions contribute the same semantic pieces...* (AUSC)

The findings of this study align with the established literature, indicating that lexical bundles in academic prose tend to be more phrasal, utilizing noun and prepositional phrases rather than verb phrases (Biber et al., 1999). This trend is attributed to the genre's emphasis on information density, leading to a shift from clausal to phrasal structures as a means to integrate information more meticulously (Pan et al., 2016; Biber et al., 1999; Lyu and Gee, 2020; Nesi and Basturkmen, 2006). The results suggest that dissertation abstracts carry a heavier information load compared to

research articles, as evidenced by the higher proportion of NP- and PP-based bundles in both the CUSC and AUSC. The greater prevalence of VP-based bundles in the AUSC implies that American students' dissertation abstracts may have a more conversational tone than those of Chinese students, a characteristic linked to the conversational structure of "personal pronoun + lexical verb phrase" outlined by Biber et al. (1999). Notably, American students demonstrate a marked preference for "subject + verb phrase + (that-clause)" structures, often employing the first person singular *I*, despite its discouragement in traditional EAP instruction. In contrast, the predominance of "passive verb + prepositional phrase" structures in the CUSC suggests a more formal and literate style, corroborating Biber et al.'s findings that passive voice is a hallmark of academic writing. These contrasts not only reflect divergent rhetorical styles but also point to the cultural dimensions of academic writing, highlighting the need for pedagogical sensitivity to these differences in EAP curricula.

The observed distributional contrast in the use of lexical bundles between Chinese and American students' dissertation abstracts presents a notable deviation from some previous findings in similar research. Contrary to Pan et al.'s (2016) findings, where Chinese writers in research articles were reported to use a greater proportion of VP-based bundles compared to L1 English writers, this study reveals a different trend. Similarly, this study's results contradict Pan et al.'s (2016) findings, where Chinese writers in research articles were reported to use a greater proportion of VP-based bundles compared to L1 English writers. Additionally, the results diverge from Azad and Khiabani's (2018) research on research article abstracts, which identified a higher prevalence of PP-based and a lower prevalence of VP-based bundles. These contradictions suggest that dissertation abstracts, as a genre, possess distinct features when compared to theses, research articles, and research article abstracts. Dissertation abstracts, characterized by their highly concise nature, typically limit the usage of VP-based structures like *we can see that*. This conciseness aligns with the genre's purpose of serving readers with varied discursive expectations. For instance, while American universities appear more accepting of the use of the authorial *I* in linguistics dissertation abstracts, Chinese universities adhere more strictly to the traditional EAP convention of avoiding the first person. These findings highlight the unique rhetorical and stylistic conventions of dissertation abstracts, reflecting broader academic cultural differences and the specific demands of the genre.

3.2 Functional Distribution

Table 5 reveals that in both corpora, text-oriented lexical bundles constitute the largest category, followed by research- and participant-oriented bundles. This pattern indicates that both Chinese and American student writers predominantly use formulaic sequences for text organization, a moderate amount for describing research activities, and a smaller proportion for engaging with readers. This distribution is likely influenced by the succinct nature of dissertation abstracts, which necessitates clear, explicit, and reader-friendly presentation of key content, thereby requiring meticulous organization of the text. However, a notable difference between the two groups lies in the proportional distribution across the major functional categories. American writers utilize a higher percentage of text-oriented bundles (65.87%) compared to their Chinese counterparts (57.38%), a difference of 8.49%. Conversely, Chinese writers employ a greater proportion of research-oriented bundles (38.44%) compared to American writers (31.05%), a variance of 7.39%. These disparities highlight differing emphases in the functional use of lexical bundles by the two groups. While American students tend to focus more on text organization, Chinese students allocate more lexical resources to detailing their research. This distinction underscores the varied approaches to structuring and presenting academic content in dissertation abstracts by different cultural groups, reflecting their unique academic conventions and stylistic preferences.

Table 5. Distribution of lexical bundles by function

Category	CUSC		AUSC	
	Token	Percentage	Token	Percentage
Research-oriented	**7,748**	**38.44%**	**1,800**	**31.05%**
Location	392	1.95%	87	1.50%
Procedure	1,251	6.21%	403	6.95%
Quantification	904	4.49%	396	6.83%
Description	3,442	17.08%	843	14.54%
Topic	1,759	8.73%	71	1.22%
Text-oriented	**11,564**	**57.38%**	**3,819**	**65.87%**
Transition signals	1,046	5.19%	318	5.48%
Inferential signals	436	2.16%	620	10.69%
Causative signals	1,708	8.47%	760	13.11%
Structuring signals	2,622	13.01%	809	13.95%
Framing signals	3,132	15.54%	754	13.00%
Relationship signals	1,220	6.05%	254	4.38%
Objective signals	1,400	6.95%	304	5.24%
Participant-oriented	**453**	**2.25%**	**123**	**2.12%**
Stance features	320	1.59%	70	1.21%
Engagement features	133	0.66%	53	0.91%
The other functions	389	1.93%	56	0.97%

Note: Bold font represents major categories.

The sub-category distribution of lexical bundles reveals several similarities between Chinese and American writers in their functional use. Both groups allocate the largest share to description bundles, with 17.08% in the CUSC and 14.54% in the AUSC. The smallest proportion is utilized for engagement feature bundles in both collections. Intermediate categories, such as structuring signals, procedure bundles, and transition signals, see similar usage proportions by both groups. This suggests a shared approach to utilizing a significant portion of lexical bundles for describing research, structuring abstracts, introducing research procedures, and facilitating argument transitions. However, differences emerge primarily in the distribution across topic bundles, inferential signals, and causative signals, as well as within framing signals – a notable sub-category where any variation can significantly impact the overall distribution. Chinese writers use 8.73% of their bundles as topic bundles, substantially more than the American writers, who use only 1.22%. Conversely, American writers allocate 10.69% and 13.11% of their bundles to inferential and causative signals, respectively, while Chinese writers use markedly less at 2.16% and 8.47%. This 8.53% and 4.64% difference reflects distinct emphases in the functional deployment of lexical bundles.These contrasts in proportional use within key sub-categories indicate divergent rhetorical strategies. Chinese writers seem to focus more on topic introduction, whereas American writers employ a higher frequency of bundles for inferential reasoning and causation. Such variations illuminate the differing rhetorical preferences and practices of Chinese and American academic writers, offering valuable insights for educators and students alike in understanding and adapting to diverse academic writing conventions.

The findings align with those of Lu and Deng (2019) on dissertation abstracts, Hyland (2008) on dissertations, and Lyu and Gee (2020) on thesis abstracts. Hyland (ibid, p. 56) posits that the prevalent use of text-oriented bundles by PhD students reflects their efforts to create more academic and reader-friendly prose, aiming to engage their readers more effectively. However, our results diverge from those of Zheng and Mao (2018), who studied 230 CSSCI research article abstracts in applied linguistics and found that research-oriented bundles were most prevalent, followed by text- and participant-oriented bundles. This discrepancy highlights the genre-specific nature of lexical bundle usage in dissertation abstracts. An additional point of contrast between this study and Hyland's (ibid) work concerns the proportion of participant-oriented bundles. In Hyland's research on dissertations, these bundles

accounted for 11.20% of the total, whereas in our study of dissertation abstracts, they comprised only 2.00%. This difference likely stems from the concise format of abstracts, which limits the space available for authors to express personal attitudes or extensively engage with readers. This constraint inherent to the abstract genre likely leads to a reduced use of participant-oriented bundles, further underscoring the nuanced differences in lexical bundle usage across various academic writing formats.

It is also revealed that Chinese students tend to use a higher proportion of research-oriented bundles but a lower proportion of text-oriented bundles compared to their American counterparts. This pattern aligns with Lu and Deng's (2019) findings in dissertation abstracts. A closer look at the sub-categories indicates that this disparity is primarily driven by the Chinese students' more frequent use of topic bundles within the research-oriented category and their less frequent employment of inferential and causative signals within the text-oriented category, as compared to American students. This suggests that Chinese students more repetitively reference their research topics, while they less frequently employ lexical bundles that indicate inferential reasoning and causation. However, the observed functional distribution differences in CUSC and AUSC diverge from the findings of Ädel and Erman (2012), Lyu and Gee (2020), and Chen and Baker (2010, p. 44) who noted that "the use of lexical bundles in non-native and native student essays is surprisingly similar" in functional distribution. The contrast in our results could stem from the specific genre of the academic prose being analyzed. It suggests that L1 and L2 speakers might exhibit more pronounced differences in their use of lexical bundles in dissertation abstracts compared to other academic genres. This genre-specific variability highlights the need for nuanced understanding and teaching approaches in academic writing, particularly in guiding non-native speakers to adapt to diverse academic conventions.

4. Conclusion and Implications

In conclusion, this study has uncovered notable structural and functional differences in the use of lexical bundles by Chinese and American PhD students. Structurally, both groups predominantly used NP-based lexical bundles, with 47.80% in the CUSC and 43.67% in the AUSC. This was followed by VP-based bundles (23.32% in CUSC; 29.08% in AUSC) and PP-based bundles (23.84% in CUSC; 22.63% in AUSC). Notably, Chinese students demonstrated a significantly higher usage of "other noun phrase fragment" bundles (12.47% in CUSC compared to 2.17% in AUSC) within the NP category, and a markedly lower usage of "subject + verb phrase + (that-clause)" bundles (1.53% in CUSC compared to 12.00% in AUSC) within the VP category, accounting for the major structural differences observed. Functionally, text-oriented lexical bundles were predominantly used by both groups, but again with differences in proportions (57.38% in CUSC; 65.87% in AUSC). This was followed by research-oriented (38.44% in CUSC; 31.05% in AUSC) and participant-oriented bundles (2.25% in CUSC; 2.12% in AUSC). The Chinese students employed a significantly greater proportion of topic bundles within the research-oriented category (8.73% in CUSC versus 1.22% in AUSC) and a smaller proportion of inferential signals within the text-oriented category (2.16% in CUSC versus 10.69% in AUSC), highlighting the main functional contrasts.

The study's exploration of lexical bundle usage by Chinese and American PhD students reveals distinct academic writing styles influenced by their respective cultural and educational backgrounds. These differences present an opportunity to develop more nuanced and effective pedagogical strategies, particularly within EAP instruction. A key educational objective emerging from this research is to narrow the gap in lexical bundle usage between these groups. By increasing Chinese students' exposure to lexical bundles prevalent in international academic discourse, instructors can help them write in a manner more aligned with global academic norms. Focusing instruction on the structural and functional categories with the most significant disparities, such as encouraging the use of "subject + verb phrase + (that-clause)" structures and incorporating more inferential and causative signals in Chinese students' writing, can further this alignment. Furthermore, a corpus-based approach to instruction, utilizing authentic examples from corpora, can offer tangible insights into effective academic writing practices. This strategy ensures that teaching is grounded in real-world language application, enhancing students' linguistic abilities and their understanding of the stylistic and rhetorical elements of academic writing. Overall, the study underscores the need for more targeted, data-driven instructional strategies in EAP programs. By focusing on reducing the usage gap, concentrating on specific structural and functional categories, and employing a corpus-based approach, EAP instruction can be more effectively tailored to meet the needs of L2 English-speaking students, particularly those from different academic and cultural backgrounds, as they navigate the complex landscape of international academic writing.

Acknowledgments

Authors contributions

This work is a single-author study by Kai Bao, who undertook the study design, data collection, analysis, and manuscript preparation. As the sole contributor, Kai Bao is responsible for all aspects of the work, and there are no additional authorship agreements. The final manuscript was read and approved by Kai Bao for submission.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Competing interests

The author declares no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Informed consent

Obtained.

Ethics approval

The Publication Ethics Committee of the Sciedu Press.

The journal's policies adhere to the Core Practices established by the Committee on Publication Ethics (COPE).

Provenance and peer review

Not commissioned; externally double-blind peer reviewed.

Data availability statement

The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.

Data sharing statement

No additional data are available.

Open access

Copyrights

References

Ädel, 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*, 81-92. https://doi.org/10.1016/j.esp.2011.08.004

Azad, H. R., & Khiabani, S. M. (2018). Lexical bundles in English abstracts of research articles written by Iranian scholars: Examples from humanities. *Iranian Journal of Applied Language Studies*, *10*(2), 149-174. https://doi.org/10.22111/IJALS.2018.4634

Bao, K., & Liu, M. (2022). A corpus study of lexical bundles used differently in dissertations abstracts produced by Chinese and American PhD students of linguistics. *Frontiers in Psychology*, *13*, 1-13. https://doi.org/10.3389/fpsyg.2022.893773

Bao, K., & Liu, M. (2023). Comparative analysis of move-specific lexical bundles in linguistics dissertation abstracts: A study of students from China and the United States. *SAGE Open*, *13*(4), 1-17. https://doi.org/10.1177/21582440231210055

Biber, D., & Barbieri, F. (2007). Lexical bundles in university spoken and written registers. *English for Specific Purposes*, *26*, 263-286. https://doi.org/10.1016/j.esp.2006.08.003

Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *Longman grammar of spoken and written English*. London, England: Longman.

Chen, Y., & Baker, P. (2010). Lexical bundles in L1 and L2 academic writing. *Language Learning and Technology*, *14*(2), 30-49. http://dx.doi.org/10125/44213

Cortes, V. (2004). Lexical bundles in published and student disciplinary writing: Examples from history and biology. *English for Specific Purposes*, *23*(4), 397-423. https://doi.org/10.1016/j.esp.2003.12.001

Coxhead, A., & Byrd, P. (2007). Preparing writing teachers to teach the vocabulary and grammar of academic prose. *Journal of Second Language Writing*, *16*, 129-147. https://doi.org/10.1016/j.jslw.2007.07.002

El-Dakhs, D. A. S. (2018). Why are abstracts in PhD theses and research articles different? A genre-specific perspective. *Journal of English for Academic Purposes*, *36*, 48-60. https://doi.org/10.1016/j.jeap.2018.09.005

Erman, B., & Warren, B. (2000). The idiom principle and the open-choice principle. *Text*, 20, 29-62. https://doi.org/10.1515/text.1.2000.20.1.29

Huang, K. (2018). Register features of lexical bundles used by Chinese EFL majors: A contrastive analysis of spoken and written English. *Foreign Language World*, *2018* (5), 71-79.

Hunston, S. (2002). Pattern grammar, language teaching, and linguistic variation: Applications of a corpus-driven grammar. In R. Reppen, S. M. Fitzmaurice, & D. Biber (Eds.), *Using corpora to explore linguistic variation* (pp. 167-183). Amsterdam: John Benjamins.

Hyland, K. (2008). Academic clusters: Text patterning in published and postgraduate writing. *International Journal of Applied Linguistics, 18*(1), 41-62. https://doi.org/10.1111/j.1473-4192.2008.00178.x

Kim, S., & Kessler, M. (2022). Examining L2 English University students' uses of lexical bundles and their relationship to writing quality. *Assessing Writing*, *51*, 1-11. https://doi.org/10.1016/j.asw.2021.100589

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*, 159-174.

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. https://doi.org/10.1016/j.jeap.2019.03.008

Lyu, M., & Gee, R. W. (2020). Lexical bundles in thesis abstracts by L1 Chinese learners of English and U.S. students. *English Language Teaching*, *13*(1), 141-155. https://doi.org/10.5539/elt.v13n1p141

Nesi, H., & Basturkmen, H. (2006). Lexical bundles and discourse signalling in academic lectures. *International Journal of Corpus Linguistics*, *11*(3), 283-304. https://doi.org/10.1075/ijcl.11.3.04nes

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. https://doi.org/10.1016/j.jeap.2015.11.003

Schmitt, N., & Carter, R. (2004). Formulaic sequences in action: An Introduction. In N. Schmitt (Ed.), *Formulaic sequences: Acquisition, processing and use* (pp. 1-22). Amsterdam: John Benjamins.

Scott, M. (2020). *WordSmith Tools version 9 (64 bit version)*. UK, Stroud: Lexical Analysis Software.

Yu, Y. (2022). The role of psycholinguistics for language learning in teaching based on formulaic sequence use and oral fluency. *Frontiers in Psychology*, *13*, 1-19. https://doi.org/10.3389/fpsyg.2022.1012225

Zheng, L., & Mao, H. (2018). A Comparison of Lexical Bundles in Translated Abstracts of Academic Articles in a CSSCl Journal. *Applied linguistics*, *2018*(2), 89-98. https://doi.org/10.16499/j.cnki.1003-5397.2018.02.010