# A Corpus-Based Study of Four-Word Lexical Bundles in Chinese and U.S Phd Dissertations

**Min Chen[1], Roslina Abdul Aziz[2], Syamimi Turiman[3]**

**[1]Academy of Language Studies, Universiti Teknologi MARA, 40450, Shah Alam, Selangor, Malaysia; School of Social Management, Jiangxi College of Applied Technology, 2 Wenfeng Road, Ganzhou, Jiangxi Province, 314000, China**

**[2]Akademi Pengajian Bahasa, Universiti Teknologi MARA Cawangan Pahang, Kampus Jengka,26400 Bandar Jengka, Pahang, Malaysia**

**[3]Akademi Pengajian Bahasa, Universiti Teknologi MARA, 40450, Shah Alam, Selangor, Malaysia**

**DOI: https://dx.doi.org/10.47772/IJRISS.2024.8110131**

**Received: 30 October 2024; Accepted: 08 November 2024; Published: 09 December 2024**

## ABSTRACT

Lexical bundles (LBs), which are crucial for smooth language articulation, are deemed a significant unique characteristic in scholarly writing. Many PhD students find that their English academic writing output often falls short of academic expectations compared to native language academic authors in terms of vocabulary sophistication. This study aimed to examine both the functional commonalities and variances in the usage patterns of four-word lexical clusters among Chinese PhD postgraduate EFL students (CH-PhD) and their American counterparts (US-PhD). The study adopted a corpus-based approach, relying on two self-built learner corpora (CH-PhD and US-PhD), and incorporated both quantitative and qualitative data analysis methods. Using AntConc's N-gram tool, 72 bundles in CH-PhD and 37 in US-PhD were extracted. Findings indicate that Chinese PhD students employed a greater number of four-word word clusters in their academic writings. As for the functional types, Chinese PhD learners are more inclined to use research-oriented bundles to provide descriptions to organise writer's actions, while native American PhD students are more inclined to use text-oriented bundles to organise the text. Furthermore, the findings illuminate methods to improve the instruction of LBs in China's academic writing curriculum.

**Keywords:** Lexical bundles; academic writing; Chinese EFL learners; corpus-based.

## INTRODUCTION

Academic writing is a form of scholarly communication in the academic community and one of the necessary ways for researchers to establish their scholarly identity in the academic community and the quality of academic papers is related to whether the papers can be accepted or published. The purpose of a PhD thesis is to prove that the candidate can conduct and communicate academic research and decide whether the writers can get the appropriate degree through thesis writing (Gastel & Day, 2022). Therefore, the quality of academic texts is one of the most important signs of the success or failure of academic writing. With the deepening of international academic exchanges and the circulation of academic information, the ability of PhD EFL learners to write academic papers and communicate in English is increasingly required. Academic writing is one of the necessary English application skills for PhD postgraduate students to carry out scientific research and international academic communication, and the cultivation of graduate students' English academic thesis writing ability has become an indispensable and important link in the development of their international academic communication ability.

The various difficulties faced by EFL learners in producing good quality academic papers has given rise to an in-depth study of the various elements that make up a good academic paper, as well as an exploration of how writers can be taught to learn the use of language in academic contexts. Pan (2024) noted that the majority of Chinese students perceive enhancing academic English writing as challenging and hard to improve. Chen (2021) pointed out that academic writing is particularly challenging for college students due to several issues

including inaccurate choice of academic words, lack of awareness in academic writing conventions and poor logic of the language.

A large number of these studies have harnessed the power of computers to analyse corpora composed of academic texts, with the aim of establishing linguistic and textual patterns in academic discourse and developing systematic descriptions of these patterns (Csomay, 2020; Lu et al., 2021; Römer et al., 2020). One key focus of these studies is the use of chunks or lexical bundles (LBs), which combine linguistic competence and pragmatic competence leading to effective communication (Duraiswamy, 2022). This focus on lexical bundles has opened new avenues for research, advancing our understanding of their role and importance in academic language.

To produce authentic academic texts, students must be proficient in the use of LBs to improve the coherence and logic of the article (Cai, 2021). In view of the importance of LBs in academic writing, EFL writers, especially PhD postgraduates, should be proficient in the use of LBs. Consequently, it is essential to describe and summarize LBs in academic discourse, as well as to analyse and discuss how EFL writers use LBs functionally in English academic writing to build their academic writing skills.

This study focuses on Applied Linguistics theses, as Applied Linguistics is one of the few fields in China where theses are written in English, making students in this field more familiar with lexical bundles. Therefore, the aim of this study is to explore the functional similarities and differences of the most frequent four-word LBs between Chinese PhD EFL learners and native American PhD postgraduates in Applied Linguistics. The research questions guiding this study are as follows:

1. What are the most frequent four-word LBs in the academic writing of Chinese PhD postgraduate EFL and native American PhD postgraduate learners?
2. What are similarities and the differences in the functions of the four-word LBs used by Chinese PhD postgraduate EFL learners and native American PhD postgraduate students in academic writings?

## LITERATURE REVIEW

The term “lexical bundle” first appeared in Longman Grammar of Spoken and Written English (Biber et al., 1999). It refers to repeated expressions, whether or not they are idiomatic or structurally defined, can be seen as extended collocations or a group of words that statistically tend to appear together within a particular register (Biber et al., 1999).According to Chen and Baker (2010), lexical bundles are defined as “continuous word sequences obtained through a corpus-driven approach, based on specific frequency and distribution criteria”(p.30). Lexical bundles are able to increase the speed of language processing and the fluency of language expression due to their holistic storage and usage characteristics (Wray, 2002). Second language acquisition theory focuses on the study of lexical bundles, which are considered to play a key role in the language acquisition process (Larsen-Freeman, 2000). LBs are quite common in academic discourse (Hyland, 2008b) and high-frequency chunks such as "on the other hand, in terms of, the results of the" in academic texts can enhance the authenticity of the language expression and show that the language user is a member of a certain academic community. They indicate the language user's membership in a particular academic community.

Earlier studies on lexical bundles (LBs) have highlighted various factors influencing their usage, particularly across different registers. Several studies have identified notable differences in LB usage between spoken and written language. Biber and Conrad (1999) emphasised the distinctive characteristics of LBs in both speech and writing, while Conrad and Biber (2005) further explored how formal and informal language is applied across diverse bundle structures.

Investigations into LB usage have also concentrated on a range of academic fields. Cortes (2004) conducted an analysis contrasting the application of 4-word LBs in published academic papers on history and biology with the unpublished semester papers from students at three separate educational stages. The study proposed that LBs are indicative of distinct disciplines, with each discipline showcasing unique attributes in the utilization of LBs. Similarly, Hyland (2008b) identified notable differences in the frequency and application of

LBs across the four fields: electrical engineering, business studies, applied linguistics, and microbiology, highlighting how each discipline showcases distinct LB patterns.

Numerous studies have also examined how LBs are used by people with different degrees of second language proficiencies to identify notable differences in their usage. Among them include Qin (2014), who examined the application of five-word LBs across different grades of university graduate students, including first-year, second year, first and second-year PhD programs, beyond second-year PhD programs, and expert levels. Li et al. (2023) applied a corpus-based method to investigate the quantity, function, and quality of four-word lexical bundles created by L2 English writers with limited proficiency and 11 varied L1 backgrounds during a timed English writing evaluation. In a comparative study, Cao (2021) analysed LBs across paradigms and disciplines, finding significant structural and functional differences among research models and fields. Similarly, Bao (2024) conducted a comparative analysis of LBs in dissertation abstracts by Chinese and American university students.

To sum up, a substantial amount of research has examined the use of LBs across various fields, contexts, and levels of proficiency. However, limited attention has been given to comparing the use of LBs by PhD-level EFL students with that of native American PhD students in Applied Linguistics. Understanding the functions of LBs is crucial for enhancing academic writing skills. To expand knowledge on LB usage and functionality, this study aims to explore the functional similarities and differences in the most frequent four-word LBs between Chinese PhD EFL learners and native American PhD students.

## METHODOLOGY

### A. Research design

The study utilised a corpus-based approach, integrating both quantitative and qualitative techniques to examine and contrast the occurrence, roles, and traits of LBs in the dissertations of Chinese PhD postgraduate EFL students (CH-PhD) with those of native American PhD postgraduate students (US-PhD). The quantitative analysis focused on measuring the occurrence and proportion of various functional LBs, their usage trends in the dissertations of Chinese PhD EFL students, and then contrasting these with the results obtained from the US-PhD corpus. The qualitative analysis involved conducting analysis through concordance analysis to ascertain the unique characteristics of LBs utilised by different PhD students.

### B. Corpora

#### 1) Chinese PhD postgraduate EFL learner (CH-PhD) corpus

The CH-PhD corpus comprises 20 dissertations authored by Chinese doctoral candidates specialising in Applied Linguistics. Information was gathered from the China National Knowledge Infrastructure (CNKI), a repository amassing the majority of China's intellectual assets. The research selected 20 doctoral dissertations authored by Chinese postgraduates from five distinct universities in five different Chinese cities, spanning 2022 to 2024, to guarantee the representativeness of the data. For maintaining consistent structural integrity, the theses adhere to the broad PhD structure: Introduction, Literature Review, Methodology, Results and Discussion. This corpus comprises a total of 1288,719 words, averaging 64,436 words in each dissertation. The details of CH-PhD are presented in Table 1.

Table 1 The details of the Corpus of Chinese PhD Postgraduate EFL Learners (CH-PhD)

| Corpus | Total size (word) | Average size (word) | Number of texts |
|---|---|---|---|
| CH-PhD | 1,288,719 | 64,436 | 20 |

#### 2) The native American PhD postgraduate students (US-PhD) corpus

PhD dissertations, sourced from the ProQuest database, are included in the US-PhD corpus for native

American PhD postgraduate students. Following Wood et al. (2001) criteria, the dissertations were gathered from authors connected with U.S. institutions or universities and were also written by persons with English first and last names to confirm English as their primary language. To achieve comparable sizes for both corpora, a selection of 20 dissertations in Applied Linguistic from 8 randomly chosen universities in the US was made, spanning the years 2022 to 2024.The dissertations also adhere to the broad PhD structure: Introduction, Literature Review, Methodology, Result and Discussion. The details of the corpus are as shown in Table 2.

Table 2 The details of corpus of native American postgraduate students (US-PhD)

| Corpus | Total size (word) | Average size (word) | Number of texts |
|---|---|---|---|
| US-PhD | 1,052,312 | 52,615 | 20 |

### 3) Computational Tool

AntConc (Anthony, 2024) was selected for data analysis in this study due to its compatibility with the research requirements, free software license, cross-platform functionality, user-friendly manual, and robust data analysis features. The N/gram function was used to facilitate the generation of 4-word LBs, which were then examined through concordance lines to analyse each specific bundle. The file view feature was also used to identify the specific files where each lexical bundle appeared. Chi-square test analysis was then performed to determine the significance of the variances in frequencies and functions of four-word clusters across the two corpora.

## C. Data analysis

### 1) Lexical bundles identification

In ensuring both representativeness and manageability of the lexical clusters derived, specific identification criteria were established. According to Hyland (2008b), four-word bundles offer a distinct set of structures and functions compared to three-word bundles and are considerably more frequent than five-word sequences. Deng and Liu (2023) further emphasised the importance of four-word bundles in conveying communicative intent in academic writing. Therefore, four-word bundles were selected as the optimal choice for this study, providing an appropriate balance between size and functionality in lexical bundle analysis. As for the normalised frequency, it is usually adopted around from 20-40 occurrences per million words (Biber et al., 2004). Nasrabady et al. (2020) adopted 50, 30, 15 for different lengths of lexical bundles, while Lyu and Gee (2020) set out a higher criterion in his study (60 times per million words). This study utilised a criterion of 30 words per million in each corpus, a norm in contemporary LBs studies (Pan et al., 2016). The converted raw frequency is as shown in Table 3.

Table 3 The normalized frequency of two corpora (30 lexical bundles/one million words)

| Corpus | Size | Normalized frequency threshold | Converted raw frequency | Rounded |
|---|---|---|---|---|
| 1 | 1288,719 | 30/1,000,000 | 38.66 | 39 |
| 2 | 1052,312 | 30/1,000,000 | 31.56 | 32 |

Dispersion is the third essential factor in identifying lexical bundles as it helps “guard against idiosyncratic uses by individual speakers or authors” (Biber et al., 2004, p. 376). This research adopted Hyland’s standard of requiring that LBs appear in at least 10% of the corpus texts, setting a dispersion threshold at three texts across both corpora. With the parameters for bundle length, frequency, and dispersion established, extracting specific LBs becomes more manageable, thereby enhancing understanding of their functions within the dissertations.

### 2) The functional categories of LBs

Biber et al. (2004) identified three primary functional categories for lexical bundles that gained broad

acceptance: stance expressions, discourse organizers, and referential expressions. Building on these guidelines, Hyland (2008b) adapted the categories to better align with research-focused genres, adding subcategories to address specific challenges in research writing. Hyland's framework proposed three main categories and eleven subcategories for the lexical bundles collected, as shown in Table 4.

Table 4 Functional classification of LBs distribution. (Hyland, 2008a)

| Functional types | Sub-types | Examples |
|---|---|---|
| Research-oriented | Location | the end of the |
| | Procedure | in the process of |
| | Quantification | a large number of |
| | Description | the similarities and differences |
| | Topic | language teaching |
| Text-oriented | Transition signals | on the other hand |
| | Resultative signals | the results of the |
| | Structuring signals | in the current study |
| | Framing signals | on the basis of |
| Participant-oriented | Stance features | it is clear that |
| | Engagement features | can be seen in |

**3) Inter-coder reliability**

Bryman (2016) defines reliability as the issue regarding the repeatability or reproducibility of a study's outcomes. In this research, to enhance the dependability of data coding, cross-analysis with multiple coders was employed to resolve any discrepancies. Two coders; an associate professor and a senior English lecturer with a decade of university English teaching experience, were chosen to code the LBs based on the functional classifications applied in this research (see Table 4). This method was implemented to minimise bias and enhance dependability. Cohen's Kappa in SPSS was then used to measure consistency between coders, yielding values from 0.8 to 1.0, indicating near-optimal reliability in functional classification across both datasets, according to Kappa statistics (Landis & Koch, 1977).

## RESULTS AND DISCUSSION

### A. Frequency of LBs in CH-PhD and US-PhD

Antconc software (4.3.1) was utilised to generate all four-word LBs. Following the process of refinement, which involved manually eliminating certain discipline bundles, context-specific bundles, overlapping bundles, and problematic bundles, the LB count was revised from 78 to 72 in CH-PhD and from 42 to 37 in US-PhD. As shown in Table 5, the total number of extracted four-word LBs in the CH-PhD corpus is 72 types and 5,846 tokens, accounting for 0.45% of the CH-PhD corpus. In contrast, the US-PhD corpus contains 37 types and 2,010 tokens, representing 0.19% of the entire corpus, which is significantly lower than in the CH-PhD corpus. The types in CH-PhD are nearly double those in US-PhD (72 vs. 37), while the tokens in CH-PhD are almost three times those in US-PhD (5,846 vs. 2,010).

Vocabulary diversity can be measured by comparing types to tokens. Type/token ratio (TTR) refers to the ratio of type to token in the corpus. The degree of TTR is directly linked to the diversity and profundity of the

vocabulary used by writers. Kyle et al. (2021) highlighted that TTR is widely known to assess the diversity of vocabulary based on corpora. With the rise in TTR's worth, the lexical diversity of the corpus also escalates, resulting in a more diverse and accurate linguistic expression. Table 5 reveals that the TTR for US-PhD surpasses that of CH-PhD (1.84:1.2), indicating a broader range of bundle usage among native American PhD students as opposed to Chinese PhD EFL students.

Table 5 Distribution of 4-word lexical bundles in CH-PhD and US-PhD

| Corpus | Total size | Types | Tokens | Token % | TTR(Type/Toke) | Chi-square | P -value |
|---|---|---|---|---|---|---|---|
| CH-PhD | 1288,719 | 72 | 5846 | 0.45% | 1.23% | 5.335 | 0.019 |
| US-PhD | 1052,312 | 37 | 2010 | 0.19% | 1.84% | | |

There are also twice as many LBs in CH-PhD compared to US-PhD, suggesting a broad spectrum of LBs used by Chinese PhD students. The outcome aligns with the discoveries made by Chen and Baker (2010). In their examination of a series of essays at different CEFR stages (B1, B2, and C1) written by Chinese English students, sourced from the Longman Learner Corpus, they found that less skilled writers used a wider variety of LBs, in contrast to the more academic style of the bundles used by more proficient students. This result aligns with the findings of Al et al. (2020), who analysed two distinct corpora of scientific papers authored by native English and Indonesian professional writers. Their findings indicated that L2 writers used a higher number of lexical bundles than L1 writers.

The datasets used in the current study are similar in scale (CH-PhD:1,288,719 tokens; US-PhD:1,052,312 tokens), yet the quantity of 4-word LBs extracted from CH-PhD is nearly double that in US-PhD 72:37), suggesting a greater dependence of Chinese PhD students on LBs and there is also the tendency of frequently using certain LBs for specific communication tasks. This might be due to the fact that these LBs offer a more accessible method for them to academically express themselves. Furthermore, the occurrence rate of certain LBs in Chinese doctoral dissertations is notably elevated, exemplified by the top 4 bundles recording over 200 instances, in contrast to a single bundle appearing more than 200 times in the US-PhD corpus. This seems to indicate that Chinese learners tend to reuse or overuse some word bundles.

**B. The most frequently used LBs in CH-PhD and US-PhD**

Displayed in Table 6 are the 30 most commonly occurring four-word LBs across both corpora, arranged by frequency. In CH-PhD, the occurrence rate of the top 30 words varied between 296 and 73 instances per million words, in contrast to the significantly lower rate in US-PhD, where it was recorded 210 to 35 times per million words. In CH-PhD, 4 LBs appeared over 200 times per million words, while an additional 10 LBs appeared in excess of 100 instances per million words. The most frequent four-word LBs in CH-PhD was at the same time, which occurred 296 times per million words, followed by the bundle on the other hand, in the process of and as well as the, which occurred 270 times, 206 times and 201 times per million words respectively. The most frequent four-word bundle in US-PhD was on the other hand, with the frequency of 210 times per million words, which was the only bundle that occurred more than 200 times per million words. There are also only 2 bundles that occur more than 100 per million words (in the case of, as well as the).

Table 6 The 30 most frequent four- word bundles in CH-PhD and US-PhD

| CH-PhD | | | US-PhD | | |
|---|---|---|---|---|---|
| Rank | Bundles | Freq./mil | Rank | Bundles | Freq./mil |
| 1 | at the same time | 296 | 1 | on the other hand | 210 |
| 2 | on the other hand | 270 | 2 | in the case of | 121 |

| | | | | | |
|---|---|---|---|---|---|
| 3 | in the process of | 206 | 3 | as well as the | 107 |
| 4 | as well as the | 201 | 4 | it is important to | 96 |
| 5 | from the perspective of | 166 | 5 | in the context of | 95 |
| 6 | in terms of the | 159 | 6 | at the same time | 80 |
| 7 | on the one hand | 158 | 7 | in the present study | 63 |
| 8 | that is to say | 128 | 8 | the meaning of the | 60 |
| 9 | on the basis of | 126 | 9 | the fact that the | 58 |
| 10 | the end of the | 123 | 10 | in terms of the | 56 |
| 11 | in other words the | 122 | 11 | the nature of the | 56 |
| 12 | is one of the | 120 | 12 | in the use of | 54 |
| 13 | in the context of | 108 | 13 | the ways in which | 53 |
| 14 | at the end of | 101 | 14 | it is possible that | 48 |
| 15 | in the present study | 98 | 15 | for the purpose of | 46 |
| 16 | one of the most | 96 | 16 | at the time of | 45 |
| 17 | it is important to | 95 | 17 | can be used to | 45 |
| 18 | in the form of | 91 | 18 | in other words the | 45 |
| 19 | the fact that the | 90 | 19 | the relationship between the | 41 |
| 20 | as can be seen | 89 | 20 | a wide range of | 37 |
| 21 | in the field of | 89 | 21 | in the following section | 37 |
| 22 | as a result of | 87 | 22 | in the form of | 37 |
| 23 | the use of the | 84 | 23 | were more likely to | 37 |
| 24 | as a matter of | 82 | 24 | at the end of | 37 |
| 25 | as shown in table | 80 | 25 | the context of the | 37 |
| 26 | a matter of fact | 79 | 26 | as part of the | 35 |
| 27 | at the beginning of | 76 | 27 | important to note that | 35 |
| 28 | in addition to the | 76 | 28 | in the same way | 35 |
| 29 | can be seen in | 74 | 29 | is one of the | 35 |
| 30 | it is necessary to | 73 | 30 | on the one hand | 35 |

**C. The shared bundles in CH-PhD and US-PhD**

After comparing the two corpora, it is surprising to find that there are 27 bundles shared by PhD students in China and the US, as shown in Table 7 below.

Table 7 List of shared LBs in CH-PhD and US-PhD

| | | Freq. | | P-value |
|---|---|---|---|---|
| Rank | Bundles | CH-PhD | US-PhD | |
| 1 | on the other hand | 270 | 210 | 0.597 |
| 2 | in the case of | 70 | 121[#] | 0.000** |
| 3 | as well as the | 201 | 107 | 0.000** |
| 4 | it is important to | 95 | 96[#] | 0.141 |
| 5 | in the context of | 108 | 95 | 0.597 |
| 6 | at the same time | 296 | 80 | 0.000** |
| 7 | in the present study | 98 | 63 | 0.136 |
| 8 | the meaning of the | 69 | 60 | 0.722 |
| 9 | the fact that the | 90 | 58 | 0.157 |
| 10 | in terms of the | 159 | 56 | 0.000** |
| 11 | the nature of the | 46 | 56[#] | 0.044* |
| 12 | in the use of | 61 | 54 | 0.666 |
| 13 | the ways in which | 49 | 53[#] | 0.156 |
| 14 | it is possible that | 43 | 48[#] | 0.136 |
| 15 | can be used to | 51 | 45 | 0.705 |
| 16 | in other words the | 122 | 45 | 0.000** |
| 17 | the relationship between the | 57 | 41 | 0.534 |
| 18 | a wide range of | 56 | 37 | 0.336 |
| 19 | in the form of | 91 | 37 | 0.000** |
| 20 | at the end of | 101 | 37 | 0.000** |
| 21 | is one of the | 120 | 35 | 0.000** |
| 22 | on the one hand | 158 | 35 | 0.000** |
| 23 | that there is a | 63 | 35 | 0.064 |
| 24 | the extent to which | 71 | 34 | 0.009* |
| 25 | are more likely to | 45 | 33 | 0.637 |
| 26 | in the field of | 89 | 32 | 0.000** |
| 27 | the results of the | 52 | 32 | 0.204 |

(note: *=$p < 0.05$, **=$p=0.000$, [#]= higher frequency)

Table 7 also displays the chi-square test outcomes that determine the frequency differences in each shared four-word LBs in both corpora. The table reveals that the P-values for 12 four-word LBs fall below 0.05 ($p<0.05$), indicating significant variances between these categories in both datasets. Among the 12 four-word LBs, 10 LBs record the P-values equal to 0.000 include in the case of, as well as the, at the same time, in terms of the, the nature of the, in other words the, in the form of, at the end of, is one of the, on the one hand, in the field of, signifying the significant difference between them.

Table 7 also displays the frequency variances between these bundles. Merely 4 LBs exhibit greater frequencies in US-PhD compared to CH-PhD (i.e. in the case of, it is important to, the nature of the, it is possible that), indicating that 23 shared LBs are overused by Chinese PhD students. Some of them are with much higher frequencies; 3 times higher, (i.e. at the same time, in terms of the, in the field of, in the form of, at the end of), and 4 times higher (i.e. in other words, is one of the, on the other hand). The findings indicate that Chinese PhD EFL students, to a certain degree, have acknowledged the significance of LBs and are skilled in using them. They use LBs for structuring sentences in their texts. Nonetheless, the findings also show that Chinese PhD postgraduates tend to frequently employ restricted four-word LBs and exhibit a lack of adaptability in their usage.

**D. The functional characteristics of LBs in CH-PhD and US-PhD**

Following Hyland's categorisation of LBs, the 72 four-word LBs in CH-PhD and 37 four-word LBs in US-PhD are divided into three groups (research-oriented, text-oriented, and participant-oriented) and ten subgroups, each with a distinct percentage spread, as depicted in Table 8.

Table 8 Functional categories distributions in CH-PhD and US-PhD

| Function | CH-PhD (Type) | | US-PhD (Type) | | Chi-squared p-value |
|---|---|---|---|---|---|
| | No. | % | No. | % | |
| Research-oriented | 30 | 41.67% | 15 | 40.54% | 0.0880 |
| Text-oriented | 29 | 40.28% | 17 | 45.95% | 0.3300 |
| Participant-oriented | 13 | 18.06% | 5 | 13.51% | 0.1340 |
| Total | 72 | 100.00% | 37 | 100.00% | 0.0190 |

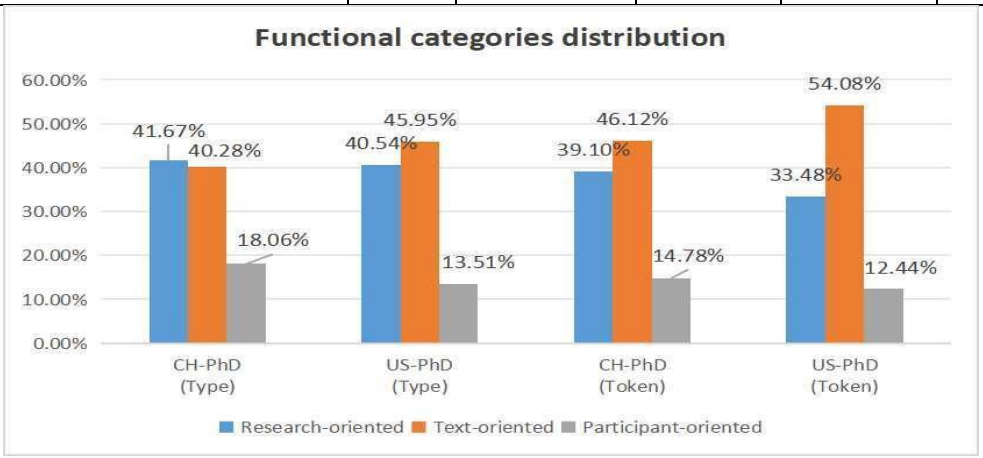


Figure 1 Distributions of functional categories in CH-PhD and US-PhD

Figure 1 shows that in CH-PhD, research-oriented bundles make up 41.67% of all bundles, representing the majority of LB types. Text-oriented bundles, positioned second, account for 40.28% of the total, slightly lower

than research-oriented bundles. The third category consists of participant-oriented bundles, constituting a mere 18.06%. Within the US-PhD framework, text-oriented bundles make up 45.95% of LB varieties and ranked first in US-PhD. Followed by research-oriented bundles, which accounted for 40.54%. Participant-oriented bundles represent the smallest portion, constituting just 13.51%. The findings indicate that Chinese PhD postgraduates predominantly depend on research-oriented bundles, in contrast to US-PhD postgraduates who favor text-oriented bundles.

### 1) Research-oriented bundles

It can be seen from Table 10 that research-oriented bundles ranked first in CH-PhD and second in the US-PhD corpus. This indicates that Chinese PhD students tend to employ more research-oriented bundles in framing their dissertations. Nonetheless, the Chi-square test findings indicate no notable disparities in the application of research-oriented bundles across the two datasets ($p=0.088 > 0.05$). This finding does not align with the findings of Hyland's (2008b) who discovered a notable disparity in the use of research-oriented LBs in the dissertations of master's students compared to doctoral dissertations. He emphasised that master students depend greatly on the tangible aspects of the study to frame their research. However, in the Chinese PhD postgraduate students have had more exposure to academic writing in comparison to master's students. There is only 1.13% difference (41.67%:40.54%) between the use of research-oriented LBs across the corpora, indicating the use of research-oriented LBs by Chinese PhD postgraduate students is closer to the American PhD postgraduate peers.

Table 10 Frequency and percentage of sub- types of research-oriented bundles in CH-PhD and US-PhD

| Function | CH-PhD | | US-PhD | | p-value |
|---|---|---|---|---|---|
| | Types | % | Types | % | |
| Research-oriented | 30 | 41.67% | 15 | 40.54% | 0.0880 |
| Location | 6 | 8.33% | 4 | 10.81% | 0.7520 |
| Procedure | 5 | 6.94% | 3 | 8.11% | 0.4780 |
| Quantification | 8 | 11.11% | 4 | 10.81% | 0.4130 |
| Description | 11 | 15.28% | 4 | 10.81% | 0.1440 |

Of the 30 research-oriented bundles in CH-PhD, 15 bundles are noun-based bundles, accounting for precisely 50% of these bundles. Most of them use the structure "noun + of" phrase (e.g. the end of the, the use of the, the meaning of the, a wide range of etc.). Of the 15 research-oriented bundles in US-PhD, 7 are preposition-based bundles, which accounted for most (46.7%). Most of them use the structure "preposition + of" phrase (e.g. in the use of, for the purpose of, at the time of etc.)

Within the description bundle subgroup, CH-PhD comprises 11 LBs, while US-PhD has 4 LBs. This represents 15.28% of the total LBs in CH-PhD corpus and constitutes 10.81% in US-PhD corpus. The authors from China and the US possess numerous similarities. Each of the four description LBs found in US-PhD is also present in CH-PhD (i.e. the meaning of the, the nature of the, the relationship between the, that there is a). Furthermore, the p-value shows there is no significant difference between them ($p=0.144>0.05$). Through the concordance analysis of the specific use of these bundles, Chinese PhD postgraduate EFL learners tend to use more bundles to specify meanings, important functions and the difference of research objects or contexts. As exemplified by the following extracts:

1. According to Davidson, beyond the literal meaning of a phrase or its words, the notion of metaphorical meaning is empty. (CH-PhD-10)
2. Leadership support (or lack of support) plays an important role in teachers' reactions towards the reform

efforts. (CH-PhD-03)

3. Log-likelihood tests also show there is a significant difference between the Chinese intermediate group and native speakers, but no significant difference between the Chinese advanced group and native speakers. (CH-PhD-01)

The finding suggests that postgraduate EFL students in China heavily depend on detailed descriptions or information during writing which aligns with Hyland (2008a) who reported that Chinese master's dissertations heavily focused on LBs that detailed the subjects or settings of research.

**2) Text-oriented bundles**

It can be seen from Table 11 text-oriented texts account for 40.28% of all the LBs (40.28%) in CH-PhD, which is lower than that in US-PhD (45.95%). This indicates that the American PhD students tend to employ more text-oriented bundles for text organisation. Their understanding of text structuring is more robust, with a heightened focus on the arrangement of scholarly texts and the rational framework of propositions. Nonetheless, the outcomes of the Chi-square test reveal no notable disparities in the application of text-oriented bundles across the two datasets ($p=0.330 > 0.05$).

Table 11 Frequency and percentage of sub- types of text-oriented bundles in CH-PhD and US-PhD

| Function | CH-PhD | | US-PhD | | p-value |
|---|---|---|---|---|---|
| | Types | % | Types | % | |
| Text-oriented | 29 | 40.28% | 17 | 45.95% | 0.3300 |
| Transition signals | 5 | 6.94% | 4 | 10.81% | 0.7750 |
| Resultative signals | 4 | 5.56% | 1 | 2.70% | 0.2420 |
| Structuring signals | 6 | 8.33% | 2 | 5.41% | 0.3730 |
| Framing signals | 14 | 19.44% | 10 | 27.03% | 0.6170 |

As for the structuring signals sub-category, the ratio of these bundles is comparable across both datasets. Nevertheless, PhD students in the U.S. employ a greater variety of bundles to engage readers and enhance their comprehension of the present text, such as in the present study (63), in the following section (38). While Chinese PhD students, other than employing comparable LBs like in the current study, in the next section, focus more on drawing readers' attention to elements like the table or the figure, such as shown in table (80), and as shown in figure (47). This deviates from the findings of earlier research. Hyland (2008b) found that there were no bundles related to tables or figures in the Applied Linguistic corpus.

In both CH-PhD and US-PhD, framing signal bundles represent the largest proportions across all sub-categories. In CH-PhD, they make up 20.83%, representing a fifth of the total bundles (22.22%), in contrast to a higher percentage (27.03%) in US-PhD. There are 6 shared bundles (i.e. in the case of, in the context of, the fact that the, in terms of the, the ways in which, in the form of). Nonetheless, Chinese PhD students predominantly depend on certain specific bundles, with four bundles (i.e. from the perspective of, in terms of the, on the basis of, in the context of) exceeding a frequency of one hundred and just a single bundle surpassing one hundred (i.e. in the case of).

**3) Participant-oriented bundles**

According to Table 12, bundles focused on participants represent the smallest percentage in both CH-PhD and US-PhD corpora, accounting for 18.06% and 13.51% respectively. Compared to research-oriented LBs and text-oriented LBs. The employment of participant-focused LBs in academic writings by Chinese PhD EFL

students bears a closer resemblance to the practices of native American PhD students. They share some common features in using this sub-type. Stance features bundles are a linguistic means by which the writer expresses his attitude towards a certain proposition or point of view, makes a judgment and establishes a proper relationship with the readers (Hyland, 2005).

Table 12 Frequency and percentage of sub- types of participant-oriented bundles in CH-PhD and US-PhD

| Function | CH-PhD | | US-PhD | | Chi-squared p-value |
|---|---|---|---|---|---|
| | Types | % | Types | % | |
| Participant-oriented | 13 | 18.06% | 5 | 13.51% | 0.1340 |
| Stance features | 6 | 8.33% | 4 | 10.81% | 0.7520 |
| Engagement features | 7 | 9.72% | 1 | 2.70% | 0.0470 |

There are three shared bundles in this sub-type, namely it is important to, it is possible that, were more likely to. Chinese PhD students were used to express writers' point of view or stance by using stance bundles to stress the necessaries (it is necessary to) and to emphasise the importance (it should be noted) as exemplified below:

1. Therefore, it is necessary to analyze the development of teachers' PCK over a length of time in its natural way. (CH-PhD- 08)
2. However, it should be noted that since Giora fails to give a strict definition of meaning, whether this result can prove the graded salience hypothesis is still open to question. (CH-PhD- 10)

Table 13 shows that Chinese PhD students utilise seven engagement feature bundles, in contrast to native the American PhD counterparts who utilise just one bundle. In addition, all the 7 bundles are related to "can be seen" (i.e. as can be seen, can be seen in, can be seen from, we can see that, it can be seen, can be seen as, can be seen that). Compared to American students, Chinese students rely more heavily on the "can be seen" bundle, using it with various prepositions (e.g., as, in, from). Most instances of this bundle refer to tables or figures as shown in the extracts below, which are not as commonly present in the US PhD dissertations.

1. As can be seen in Figures 4.15 and 4.16, most teachers believed that IELTS and Innovation conversations were more authentic, with 22 (73.33%) for IELTS and 18 (60.00%) for Innovation. (CH-PhD- 12)
2. As what can be seen from Table 4-17, no significant interaction effect was found with respect to D-value although participants produced the most various vocabularies in simple written task. (CH-PhD-16)

## CONCLUSION

Utilising the self-developed learner corpora of CH-PhD and US-PhD, this research methodically investigates the most common usage patterns of 4-word LBs, the functional classifications in doctoral dissertations by Chinese PhD EFL students and native American PhD students majoring in Applied Linguistics. The two groups share certain commonalities; however, a notable disparity exists between them. Chinese postgraduate EFL students, in contrast to their native American counterparts, tend to rely more heavily on four-word lexical bundles for academic writing. This reliance indicates less diversity in their lexical bundles, suggesting a higher dependence on standardised expressions among these students and lower confidence or proficiency in text construction. Regarding functional categories, PhD postgraduate Chinese EFL learners tend to favor research-oriented bundles for structuring writer actions, in contrast to the native American postgraduates who favor text-oriented bundles for text organisation.

The outcomes of this study bear considerable implications for academic writing pedagogy. Firstly, the data can

be used to enhance EFL students' understanding of commonly utilised LBs within a particular academic field by using the corpus as a reference tool for academic writing as it offers authentic models for the Chinese PhD EFL learners. Secondly, data derived from the US-PhD corpus can help supply the educational resources required in the selection, organisation, and structuring of educational content on field-specific bundles in developing a bundle-based language model (Nasrabady et al., 2020).

## ACKNOWLEDGEMENT

This study was not supported by any grants from funding bodies in the public, private, or not-for-profit sectors.

## REFERENCES

1. Al, F. M. S., Wasito, K. A., & Kharisma, P. C. I. (2020). Lexical bundles of L1 and L2 English professional scholars: A contrastive corpus-driven study on applied linguistics research articles. Journal of Language and Education, 6(4 (24)), 76-89. https://doi.org/10.17323/jle.2020.10719
2. Anthony, L. (2024). AntConc (Version 4.3.1) [Computer Software]. Tokyo, Japan: Waseda University. Available from https://www.laurenceanthony.net/software
3. Bao, K. (2024). Comparative analysis of lexical bundles in dissertation abstracts: insights for teaching academic English to Chinese students. English Linguistics Research, 13(1), 1-8. https://doi.org/10.5430/elr.v13n1p8
4. Biber, D., & Conrad, S. (1999). Lexical bundles in conversation and academic prose. In Out of corpora (pp. 181-190). Brill. https://doi.org/10.1163/9789004653689_017
5. Biber, D., Conrad, S., & Cortes, V. (2004). If you look at …: lexical bundles in university teaching and textbooks. Applied Linguistics, 25(3), 371-405. https://doi.org/10.1093/applin/25.3.371
6. Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). Longman grammar of spoken and written English. Longman London.
7. Bryman, A. (2016). Social research methods. Oxford University Press.
8. Cai, F. (2021). Research on improving English writing ability by lexical chunks approach. Frontiers in Educational Research, 4(8). https://doi.org/10.25236/FER.2021.040808.
9. Cao, F. (2021). A comparative study of lexical bundles across paradigms and disciplines. Corpora, 16(1), 97-128. https://doi.org/10.3366/cor.2021.0210
10. Chen, J. (2021). The challenge of college-level Chinese students academic writing in English. Greensboro College.
11. Chen, Y.-H., & Baker, P. (2010). Lexical bundles in L1 and L2 academic writing. Language Learning & Technology, 14(2), 30-49.
12. Conrad, S., & Biber, D. (2005). The frequency and use of lexical bundles in conversation and academic prose. Lexicographica, 20(2004), 56-71. https://doi.org/10.1515/9783484604674.56
13. 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
14. Csomay, E. (2020). A corpus-based study of academic word use in EFL student writing. Advances in corpus-based research on academic writing, 9-32. https://doi.org/10.1075/scl.95.01cso
15. Deng, L., & Liu, J. (2023). Move–bundle connection in conclusion sections of research articles across disciplines. Applied linguistics, 44(3), 527-554. https://doi.org/10.1093/applin/amac040
16. Duraiswamy, M. (2022). Chunks and Language Development: A Lexical Approach. Notion Press.
17. Gastel, B., & Day, R. A. (2022). How to write and publish a scientific paper. Bloomsbury Publishing USA. https://doi.org/10.5040/9798400666933.ch-035
18. Hyland, K. (2008a). 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
19. Hyland, K. (2008b). As can be seen: Lexical bundles and disciplinary variation. English for Specific Purposes, 27(1), 4-21. https://doi.org/10.1016/j.esp.2007.06.001
20. Kyle, K., Crossley, S. A., & Jarvis, S. (2021). Assessing the validity of lexical diversity indices using direct judgements. Language Assessment Quarterly, 18(2), 154-170. https://doi.org/10.1080/15434303.2020.1844205

21. Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. Biometrics, 33(1), 159-174. https://doi.org/10.2307/2529310
22. Larsen-Freeman, D. (2000). Techniques and principles in language teaching. Oxford University.
23. Li, M., Zhang, X., & Reynolds, B. L. (2023). Exploring lexical bundles in low proficiency level L2 learners' English writing: An ETS corpus study. Applied Linguistics Review, 14(4), 847-873. https://doi.org/10.1515/applirev-2020-0129
24. Lu, X., Casal, J. E., & Liu, Y. (2021). Towards the synergy of genre-and corpus-based approaches to academic writing research and pedagogy. International Journal of Computer-Assisted Language Learning and Teaching (IJCALLT), 11(1), 59-71. https://doi.org/10.4018/ijcallt.2021010104
25. Lyu, M., & Gee, R. W. (2020). Lexical bundles in thesis abstracts by L1 Chinese learners of English and US students. English Language Teaching, 13(1), 141-155. https://doi.org/10.5539/elt.v13n1p141
26. Nasrabady, P., Elahi S. M., & Ehsan G.S. (2020). Exploring lexical bundles in recent published papers in the field of applied linguistics. Journal of World Languages, 6(3), 175-197.
27. Pan. (2024). An analysis of common problems in academic English and a content-based teaching of academic English writing. Journal of Higher Education, 8, 101-104.
28. 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
29. Qin, J. (2014). Use of formulaic bundles by non-native English graduate writers and published authors in applied linguistics. System, 42, 220-231. https://doi.org/10.1016/j.system.2013.12.003
30. Römer, U., Cortes, V., & Friginal, E. (2020). Advances in Corpus-based Research on Academic Writing: Effects of discipline, register, and writer expertise (Vol. 95). John Benjamins Publishing Company. https://doi.org/10.1075/scl.95
31. Wood, A., Flowerdew, J., & Peacock, M. (2001). International scientific English: The language of research scientists around the world. Research Perspectives on English for Academic Purposes, 1, 71-83. https://doi.org/doi.org/10.1017/cbo9781139524766.008
32. Wray, A. (2002). Formulaic language and the lexicon. Cambridge University Press. https://doi.org/10.1017/cbo9780511519772