# CORPUS-BASED STUDY ON THE LEXICAL BINDLES IN WRITTEN ENGLISH OF CHINESE ADVANCED ENGLISH LEARNERS

SHI, F.[1*] – ZACHARIAH, T.[1] – SHAARI, N.[2]

[1] *Faculty of Education and Social Sciences, Universiti Selangor, Selangor, Malaysia.*

[2] *Faculty of Engineering and Life Sciences, Universiti Selangor, Selangor, Malaysia.*

**Corresponding author*
*e-mail: alanstone7772003[at]126.com*

(Received 18th July 2024; revised 06th October 2024; accepted 14th October 2024)

**Abstract.** Lexical bundles, as recurrent fixed expressions, play a vital role in achieving fluency and naturalness in language output. Corpus linguistics provides a robust framework for analyzing authentic language use and identifying patterns in large corpus. This paper employs the corpus research approach, combined with theories of second language acquisition, to explore the overall usage patterns of lexical bundles in the writing of Chinese advanced English learners. AntConc software and self-compiled FoxPro programs are utilized to extract word frequency data and perform concordance line searches based on two sampled corpora: the British National Corpus and the Chinese Learner English Corpus. Comparative analysis reveals that although there is no significant difference between Chinese advanced learners of English and native speakers in terms of the overall token frequency of high-frequency lexical bundles, considerable differences in type frequency and distribution patterns are significant. Factors such as first language transfer and avoidance strategies significantly influence the usage patterns of lexical bundles among Chinese students. Therefore, the findings of this paper help provide insights for designing academic English courses for new graduate students, and in the design process of all relevant courses and teaching practices, instructors should take the impact of these distinctive patterns into considerations.
**Keywords**: *corpus linguistics, lexical bundles, Chinese advanced English learners, frequency*

## Introduction

Since the 1980s, with the rapid development of computer technology, corpus linguistics has become a focus of attention in the field of linguistics. The term “corpus” comes from Latin, which means “body”. Nowadays, “corpus” generally refers to a large electronic text database collected based on certain sampling standards (and often needs to be scientifically processed and organized), which can be searched and statistically analyzed using computer analysis tools. Therefore, corpora provide massive real empirical data for language research, and the corpus-based research method has the quantitative advantage compared to other language research methods. In research practice, in order to improve quantitative research reliability of this type, researchers tend to provide manual assistance and apply statistical test methods to further guarantee the reliability of the results. In terms of theoretical perspectives, some researchers ever compared the corpus research method to a critical weight on the scales of the two major linguistic fields, namely structuralism and functionalism. This “weight”, which has been denigrated by Chomsky’s Mentalism, has once again risen from the downturn due to the deep questioning and criticism of Chomsky’s theory by cognitive and other schools. One of the important reasons is that it can provide quite convincing evidence for functionalism, which emphasizes the sociality of language. Therefore, in recent years,

the research methods of corpus linguistics have been widely applied in the fields of teaching, linguistics, translation, and so on.

The concept of lexical bundles just emerged in such linguistic background as mentioned above. Different scholars have given it different terms, such as lexical phrases (Nattinger and DeCarrico, 1992), lexical chunks (Michael, 1993), lexical bundles (Biber et al., 1999), and multi-word items (Schmitt and McCarthy, 1998). Moreover, other common terms include formulaic sequences, prefabs, recurrent word combinations, and "n-grams" (as computational linguists define). Like other hot research topics, the study of lexical bundles also has the problem of inconsistent terminology and divergence in the definition of core concepts. Since this research framework mainly refers to Biber et al. (2021) relevant data, "lexical bundle" uniformly used in this paper is taken as the working definition hereinafter. A large number of related studies have shown that lexical bundles occupy a considerable proportion in both oral and written English of the native speakers. According to the survey of Biber et al. (2021), native speakers tend to produce more than 60,000 three-word lexical bundles and more than 8,500 four-word lexical bundles per million words in written English. Furthermore, lexical bundles, as unified entity of storage and extraction, have the function of reducing the cognitive load in communication and improving the fluency, accuracy, and idiomaticity of language production. Some scholars even claim that the storage and effective retrieval of a large number of lexical bundles is the key to achieving or approaching native speakers' proficiency. In other words, foreign language learners' English proficiency largely lies in the number of high-frequency bundles they have mastered. Therefore, this study aims to: (1) examine significant differences in the utilization of lexical bundles between Chinese advanced English learners and native English speakers in written English; (2) investigate underlying linguistic acquisition factors contributing to these distinctions; (3) offer pedagogical insights to enhance the design of relevant English courses and writing teaching in China.

## *Literature review*

Many scholars and researchers have studied English lexical bundles from different perspectives, and it has been shown that the use of lexical bundles may vary across different groups of writers. Relevant studies in earlier stage include Biber et al. (1999) examination of the distribution of various types of lexical bundles in different English registers based on large corpora; Cowie (1998) empirical research on the oral English produced by native speakers based on the CLL corpus, and so on. In recent years, further and deeper studies have been conducted from different perspectives, with more and more focus on the enhancement of corpus data extraction as well as the effects on foreign language use. Pan et al. (2020) delved into methodological issues in contrastive lexical bundle research, highlighting the importance of corpus design when comparing lexical bundle use across groups. Shin (2019) explored the use of lexical bundles in college students' essays and questioned whether native writers always have an advantage over nonnative writers in this aspect. Güngör and Uysal (2020) focused on crosslinguistic influence in academic texts by compiling a specialized research article corpus containing L1 English, L2 English, and L1 Turkish articles. The study aimed to explore any influence of L1 Turkish on L2 English writing. Ren (2021) as well as Shirazizadeh and Amirfazlian (2021) examined lexical bundles in the context of applied linguistics research articles and textbooks, shedding light on interdisciplinary variations and the variability and functions of these recurrent language patterns. Yakut et al.

(2021) conducted a study on lexical bundles in L1 and L2 English doctoral dissertations, emphasizing the importance of corpus-informed lists and concordances for academic writers producing work in English. Birhan (2021) investigated the effects of teaching lexical bundles on EFL students' academic writing skills improvement, using corpus analysis software to select frequent lexical bundles for classroom instruction. In a study by Kim and Kessler (2022), L2 English university students' use of lexical bundles was examined in relation to writing quality, highlighting the significance of these recurrent language patterns in academic writing. Code (2023) discussed the foundations of familiar language, including formulaic expressions, lexical bundles, and collocations, emphasizing their role in both work and play contexts. These studies collectively contribute to the understanding of lexical bundles and their implications for language use in different circumstances.

In China, there has been a surge of interest in lexical bundle research among college teachers and researchers, with a particular focus on the use of these multi-word sequences in academic writing and EFL contexts. Pan et al. (2020) conducted a cross-disciplinary study of lexical bundles in the academic writing of Chinese university students. The study compared the use of bundles across four disciplines: humanities, social sciences, natural sciences, and engineering. The findings suggest that while there were some common bundles used across all disciplines, there were also significant variations in the types and functions of bundles used in each field. The authors argue that these differences reflect the distinct rhetorical conventions and communicative purposes of each discipline. Ruan (2017) explored the use of lexical bundles in the spoken English of Chinese university students during academic presentations, and the findings indicate that Chinese students tend to rely heavily on certain types of bundles, such as those used for introducing topics or expressing stance, while underusing others, such as those used for summarizing or concluding.

These recent studies demonstrate the growing interest and sophistication of lexical bundle research among Chinese scholars. However, there are still several areas that require further investigation, including the major factors that influence lexical bundle use and the use of more advanced corpus techniques to identify patterns and variations in lexical bundle use compared to native speakers. The difference between this study and the former studies is: first, in terms of data extraction tools, the above studies mostly only use existing corpus retrieval software for data statistics, which inevitably creates certain limitations. Regarding the Foxpro programming software employed in this study, as Fengxiang (2010) said, it is very suitable and effective software for language researchers to process English corpus, which is also an important innovation of this study. In addition, this paper will systematically examine the main patterns of lexical bundles in the writing of Chinese English learners at the advanced stage. In addition, while some studies have highlighted issues in non-native speakers' use of lexical bundles, they have not sufficiently explored the reasons behind these phenomena and their pedagogical implications. This study aims to fill this gap by comparing the usage patterns of lexical bundles between Chinese advanced learners and native speakers based on self-compiled programs, systematically analyzing the frequency, types, and grammatical structures of these bundles to provide empirical evidence and theoretical support for English vocabulary teaching. The specific research questions include: (1) Compared with native speakers, is there significant difference in the use of lexical bundles in the written English produced by Chinese advanced English learners?

(2) If so, what are the specific manifestations? (3) What are the linguistic acquisition causes of the differences and the corresponding teaching implications?

## Materials and Methods

The typical lexical bundles and classifications listed in this study are based on Biber et al. (2021) Grammar of spoken and written English. On this basis, this paper will adopt a combination of quantitative and qualitative research methods, taking advantage of automatic retrieval by self-compiled computer programs while also having manual assistance and analysis to ensure the accuracy of statistical data and the effectiveness of conclusions. Overall, this research mainly focuses on the high-frequency lexical bundles in the output of Chinese English learners (referring to the high-frequency lexical bundles of native speakers), and employs a new method combining general software packages and self-programmed procedures to extract data for descriptive comparative analysis.

### *Corpus*

The Chinese Learner English Corpus (CLEC) and the British National Corpus (BNC) are the data sources for this study. CLEC was constructed by Professors Gui Shichun and Yang Huizhong, and has been widely recognized and adopted by corpus researchers in China. The ST6 subcorpus is composed of written English for exam essays produced by Chinese advanced English learners, with a corpus size of approximately 200,000 words, which can reflect the representative features in written English of Chinese students. For the purpose of comparative analysis, a written English sample of about 1 million words from BNC as the reference corpus, was automatically extracted through self-compiled Foxpro programs. BNC is a large-scale corpus representing British English, co-developed by Oxford University Press and the British Library. Therefore, the 1 million-word BNC written language sample can fully reflect the written language features of native English speakers. Due to the different corpus sizes of the two corpora, statistical processing was performed on the comparative analysis data to make the two subcorpora comparable.

### *Tools*

The concordancing tools employed in this study are self-compiled Foxpro programs and the corpus software AntConc. In addition, in order to analyze and interpret the data more scientifically, SPSS and chi-square tests are used to identify whether the differences between the comparison data reach statistical significance. As mentioned earlier, the Foxpro program can automatically process large-scale corpus text according to the researcher’s specific research purposes. Moreover, AntConc, developed by Professor Laurence Anthony of Waseda University, is a popular corpus software for text processing such as KWIC concordancing and lexical bundle extraction. SPSS is one of the most commonly used professional statistical software in the world, which can conveniently perform statistical analyses such as chi-square tests. For convenience of statement, the two corpora mentioned hereinafter refer to the ST6 subcorpus and the 1-million-word BNC sampling corpus, respectively.

### *Lexical bundle extraction*

The high-frequency lexical bundles of native speakers involved in this study refer to the 3-to-6-word lexical bundles listed by Biber et al. (2021) in the Grammar of spoken and written English. In order to obtain more specific lexical bundle frequency and contextual information for comparative analysis, the AntConc software was employed to perform basic lexical bundle frequency and KWIC retrieval statistics on the two corpora respectively. However, the AntConc software have certain limitations in retrieving standardized complete sentence concordance lines. Foxpro self-compiled programs can help extract detailed contextual information of lexical bundles in the form of complete sentences more accurately and efficiently, which is more convenient for manual screening and in-depth observation and analysis of individual lexical bundles (*Table 1* and *Table 2*)

***Table 1***. *Segment of self-compiled Foxpro program.*

| Category | Description |
|---|---|
| *set defa to e:\corpus* | *strtof(st, 'e:\corpus\meta.txt')* |
| *set safe off* | *use word* |
| *clear* | *appe from e:\corpus\meta.txt sdf* |
| *clos data* | *dele all for word=' '* |
| *crea tabl word(word c(25))* | *pack* |
| *clos data* | *a=''* |
| *st=filetost('e:\corpus\metatxtall.txt')* | *scan* |
| *st=strtr(st, ' ', chr(13))* | *store alltrim(word) to b* |

***Table 2.*** *Examples of full-sentence concordance lines extracted by FoxPro program.*

| No | Description |
|---|---|
| 1 | Everybody study hard in order to find out the way which we can study well and we can get the good result in the exam. |
| 2 | The Student Union of our school will hold a party in order to welcome our American new friends on this Saturday, August, 15, at 7:30 in the evening. |

## Results and Discussion

Based on computer-driven automatic statistics and manual correction, the overall data of two comparative corpora were obtained, and Chi-square tests were performed on this basis. The statistical results are shown in *Table 3*, *Table 4*, and *Table 5*. The statistics show that all the high-frequency lexical bundles (95 in total) listed by Biber et al. (2021) appeared in the BNC sampling corpus. In contrast, only less than half of the lexical bundles appeared in the ST6 corpus representing Chinese advanced learners. *Table 4* shows that the Chi-square test value of 2.853 based on the total token numbers of lexical bundles in *Table 3* is less than the critical value of 3.84, indicating that there is no significant difference in the distribution of lexical bundle token frequencies between the two corpora. However, this does not mean that Chinese students' application of lexical bundles is close to or reaches the level of native speakers, because the test result in *Table 5* (the chi-square value of 16.508 is much greater than the critical value) reveals that there is a significant difference between the two corpora in the frequency of lexical bundle types.

***Table 3***. *Overall distribution of lexical bundles in BNC sampling corpus and ST6.*

| Corpus | BNC | ST6 |
|---|---|---|

| | | |
|---|---|---|
| Size | 1001852 | 226106 |
| Token Total | 3359 | 707 |
| Type Total | 95 | 41 |

***Table 4***. *Chi-square test result of lexical bundle token comparison.*

| Category | Value | df | Asymp. Sig. (2-sided) | Exact Sig. (2-sided) | Exact Sig. (1-sided) |
|---|---|---|---|---|---|
| Pearson Chi-Square | 2.853(b) | 1 | .091 | | |
| Continuity Correction(a) | 2.785 | 1 | .095 | | |
| Likelihood Ratio | 2.896 | 1 | .089 | | |
| Fisher's Exact Test | | | | .093 | .048 |
| Linear-by-Linear Association | 2.853 | 1 | .091 | | |
| N of Valid Cases | 1227958 | | | | |

*Note: a. Computed only for a 2x2 table; b. 0 cells (.0%) have expected count less than 5. The minimum expected count is 748.68.*

***Table 5***. *Chi-square test result of lexical bundle type comparison.*

| Category | Value | df | Asymp. Sig. (2-sided) | Exact Sig. (2-sided) | Exact Sig. (1-sided) |
|---|---|---|---|---|---|
| Pearson Chi-Square | 16.508(b) | 1 | .000 | | |
| Continuity Correction(a) | 15.582 | 1 | .000 | | |
| Likelihood Ratio | 14.224 | 1 | .000 | | |
| Fisher's Exact Test | | | | .000 | .000 |
| Linear-by-Linear Association | 16.504 | 1 | .000 | | |
| N of Valid Cases | 4107 | | | | |

*Note: a. Computed only for a 2x2 table; b. 0 cells (.0%) have expected count less than 5. The minimum expected count is 23.41.*

To comprehensively and systematically compare the distribution and application of different lexical bundles, the following table and figure have been compiled based on data regarding lexical bundle types, frequencies, and Chi-square values (all the high-frequency lexical bundles are classified into 12 types based on their structures). The statistical data shows that there are significant differences between the high-frequency lexical bundles produced by Chinese English learners and native speakers. Out of the 95 high-frequency lexical bundles by native speakers, only 41 non-zero frequency lexical bundles appeared in the ST6 subcorpus, indicating that although Chinese advanced learners are not inferior to native speakers in the total token of lexical bundle output, the richness and variety of their lexical bundle output still needs to be improved. Among the 41 high-frequency lexical bundles used by Chinese English learners, 17 lexical bundles (marked with * in *Table 6)* showed significant differences in frequency compared to native speakers, accounting for about 40% of the total (*Figure 1*). From the perspective of Chi-square values, some lexical bundles are significantly overused, such as "on the other hand" ($\chi2$=222.706), "in order to" ($\chi2$=197.308), "there is no" ($\chi2$=85.334), "at the same time" ($\chi2$=84.519), "the development of the" ($\chi2$=56.934), etc. This further confirms the findings of previous related studies (Hyland, 2008; Wray, 2002; Ellis, 1997), namely the tendency to overuse certain lexical bundles or collocations is a common problem in the development of interlanguage, and the main reason for this tendency is the constraint of interlanguage proficiency, that is, the lack of richness of lexical bundles.

***Table 6***. *High frequency lexical bundles of Chinese learners.*

| No | Lexical bundle | Category | Freq. (ST6) | Freq. (BNC) | Value ($X^2$) |
|---|---|---|---|---|---|
| 1 | *in order to | Other expressions | 121 | 117 | 197.308 |
| 2 | *there is no | Pronoun/noun phrase + be (+…) | 100 | 161 | 85.334 |
| 3 | *on the other hand | Other prepositional phrase fragments | 70 | 23 | 222.706 |

| | | | | | |
|---|---|---|---|---|---|
| 4 | *there is a | Pronoun/noun phrase + be (+…) | 62 | 211 | 5.895 |
| 5 | *at the same time | Other prepositional phrase fragments | 60 | 65 | 84.519 |
| 6 | *is one of the | Linking verb be + noun phrase/adjective phrase | 22 | 38 | 15.889 |
| 7 | the end of the | Noun phrase + of phrase fragments | 22 | 129 | 0.828 |
| 8 | *the fact that | Noun phrase + other postmodifying fragments | 21 | 204 | 10.570 |
| 9 | *the development of the | Noun phrase + of phrase fragments | 19 | 7 | 56.934 |
| 10 | *in the United States | Other prepositional phrase fragments | 17 | 11 | 37.040 |
| 11 | *part of the | Noun phrase + of phrase fragments | 16 | 198 | 15.228 |
| 12 | at the end of | Prepositional phrase + of phrase fragments | 15 | 93 | 0.912 |
| 13 | as well as the | Other expressions | 14 | 37 | 3.694 |
| 14 | at the end of the | Prepositional phrase + of phrase fragments | 12 | 45 | 0.560 |
| 15 | *on the one hand | Other prepositional phrase fragments | 11 | 15 | 11.398 |
| 16 | *on the basis of | Prepositional phrase + of phrase fragments | 10 | 22 | 4.364 |
| 17 | *the use of | Noun phrase + of phrase fragments | 10 | 112 | 7.298 |
| 18 | *the number of | Noun phrase + of phrase fragments | 9 | 16 | 6.123 |
| 19 | *to the development of | Prepositional phrase + of phrase fragments | 9 | 5 | 21.625 |
| 20 | in the case of | Prepositional phrase + of phrase fragments | 9 | 34 | 0.394 |
| 21 | that there is no | (Verb phrase +) that clause fragments | 8 | 20 | 2.489 |
| 22 | the rest of the | Noun phrase + of phrase fragments | 7 | 68 | 3.402 |
| 23 | one of the most | Noun phrase + of phrase fragments | 6 | 38 | 0.422 |
| 24 | *as a result of | Prepositional phrase + of phrase fragments | 5 | 77 | 7.353 |
| 25 | are likely to be | (Verb/adjective +) to phrase fragments | 5 | 12 | 1.741 |
| 26 | it is necessary to | Antecedent it + verb phrase/adjective phrase | 5 | 14 | 1.077 |
| 27 | in the process of | Prepositional phrase + of phrase fragments | 4 | 15 | 0.185 |
| 28 | is based on the | Passive verb + prepositional phrase fragments | 4 | 12 | 0.661 |
| 29 | the fact that the | Noun phrase + other postmodifying fragments | 4 | 31 | 0.856 |
| 30 | can be used to | (Verb/adjective +) to phrase fragments | 3 | 13 | 0.023 |
| 31 | in terms of the | Prepositional phrase + of phrase fragments | 3 | 24 | 0.732 |
| 32 | the beginning of the | Noun phrase + of phrase fragments | 3 | 16 | 0.031 |
| 33 | the case of the | Noun phrase + of phrase fragments | 3 | 18 | 0.136 |
| 34 | the nature of the | Noun phrase + of phrase fragments | 3 | 29 | 1.422 |
| 35 | to be able to | (Verb/adjective +) to phrase fragments | 3 | 37 | 2.720 |
| 36 | *for the first time | Other prepositional phrase fragments | 2 | 49 | 6.471 |
| 37 | at the beginning of | Prepositional phrase + of phrase fragments | 2 | 24 | 1.694 |
| 38 | in the course of | Prepositional phrase + of phrase fragments | 2 | 9 | 0.006 |
| 39 | it can be seen | Antecedent it + verb phrase/adjective phrase | 2 | 11 | 0.034 |
| 40 | it is difficult to | Antecedent it + verb phrase/adjective phrase | 2 | 13 | 0.167 |
| 41 | it is possible that | Antecedent it + verb phrase/adjective phrase | 2 | 5 | 0.619 |

*Note: Lexical bundles marked with * refer to the ones of greater than the critical value of 3.84, indicating a statistically significant difference in frequency distribution.*

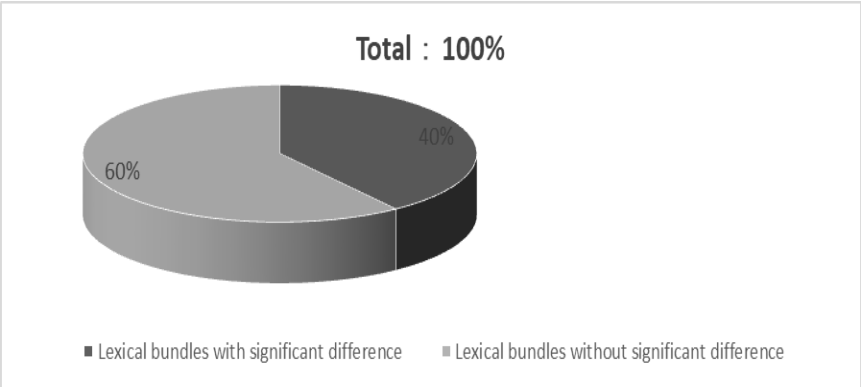


***Figure 1.*** *Difference proportion of lexical bundles produced by Chinese learners.*

The contexts of these lexical bundles with significant differences show that Chinese learners not only have a tendency to overuse high-frequency lexical bundles, but also have problems of misuse (Examples 1 and Examples 2) or inappropriate collocation (eg. the frequency of "on the other hand" is far higher than "on the one hand", because Chinese English learners often output "on the other hand" alone when organizing discourse, without considering its logical collocation with "on the one hand").

Example 1: And many flowers made together in order to made an ornamental Such as [st226106.txt].

Example 2: I hurry home from school. There is no places like our home. It includes [st226106.txt].

Observation of the specific concordance lines and more contextual information found that compared to native speakers, in the specific application of the same type of lexical bundles, the semantic richness of the lexical bundles produced by Chinese English learners is clearly insufficient. Taking the lexical bundle category of (verb/adjective+) to phrase fragment as an example: for native speakers, this category of lexical bundle can express rich semantic information, including judgments of size and quantity, expression of abstract qualities, the possibility of the development and change of things, and previous experience, etc., while Chinese English learners mainly employ them to express the possibility of the development and change of things. The frequency ranking in Table 6 shows that native speakers tend to use more lexical bundles with noun phrase as the main structure, such as "the fact that", "part of the", "the end of the", while Chinese English learners tend to overuse fixed expressions such as "in order to", "on the other hand", "at the same time", which may be the result of rote memorization. In addition, lexical bundles such as "in order to", "on the other hand", "at the same time" are also expressions that can find direct corresponding phrases in Chinese, which may be regarded as "positive transfer" from the mother tongue, so they are more easily accepted and overused by Chinese students.

Among the various categories of high-frequency lexical bundles produced by Chinese English learners (*Table 7*), the ranking order from most to least is: pronoun/noun phrase + be (+...), other prepositional phrase fragments, other expressions, noun phrase + of phrase fragment, prepositional phrase + of phrase fragment, noun phrase + other post-modifying fragments, copula be + noun phrase/adjective phrase, anticipatory it + verb phrase/adjective phrase, (verb/adjective+) to phrase fragment, (verb phrase+) that clause fragment, passive verb + prepositional phrase fragment, adverbial clause fragment, with the adverbial clause fragment being zero frequency. In contrast, the ranking order of high-frequency lexical bundles for native speakers is: noun phrase + of phrase fragment, prepositional phrase + of phrase fragment, pronoun/noun phrase + be (+...), noun phrase + other post-modifying fragments, other prepositional phrase fragments, other expressions, (verb/adjective+) to phrase fragment, anticipatory it + verb phrase/adjective phrase, (verb phrase+) that clause fragment, copula be + noun phrase/adjective phrase, passive verb + prepositional phrase fragment, adverbial clause fragment. Although the output frequency order of the two differs to some extent, there is also a certain consistency in the differences (such as the ranking of anticipatory it + verb phrase/adjective phrase, passive verb + prepositional phrase fragment, and adverbial clause fragment lexical bundles is the eighth, eleventh, and twelfth positions respectively; while the ranking of pronoun/noun

phrase + be (+...), other prepositional phrase fragments, other expressions, noun phrase + of phrase fragment, prepositional phrase + of phrase fragment, and noun phrase + other post-modifying fragment lexical bundles are all in the top six positions of their respective corpora, with only a slight change in order), which seems to demonstrate that Chinese advanced English learners and native speakers tend to share some similarities in the frequency order of lexical bundle types, and will become more and more consistent.

***Table 7**. Summary of lexical bundle categories and frequencies.*

| No | Category (11) | Freq. (ST6) | Category (12) | Freq. (BNC) |
|---|---|---|---|---|
| 1 | Pronoun/noun phrase + be (+…) | 162 | Noun phrase + of phrase fragments | 1321 |
| 2 | Other prepositional phrase fragments | 160 | Prepositional phrase + of phrase fragments | 547 |
| 3 | Other expressions | 135 | Pronoun/noun phrase + be (+…) | 378 |
| 4 | Noun phrase + of phrase fragments | 98 | Noun phrase + other postmodifying fragments | 345 |
| 5 | Prepositional phrase + of phrase fragments | 71 | Other prepositional phrase fragments | 203 |
| 6 | Noun phrase + other postmodifying fragments | 25 | Other expressions | 154 |
| 7 | Linking verb be + noun phrase/adjective phrase | 22 | (Verb/adjective +) to phrase fragments | 136 |
| 8 | Antecedent it + verb phrase/adjective phrase | 11 | Antecedent it + verb phrase/adjective phrase | 128 |
| 9 | (Verb/adjective +) to phrase fragments | 11 | (Verb phrase +) that clause fragments | 74 |
| 10 | (Verb phrase +) that clause fragments | 8 | Linking verb be + noun phrase/adjective phrase | 38 |
| 11 | Passive verb + prepositional phrase fragments | 4 | Passive verb + prepositional phrase fragments | 20 |
| 12 | Adverbial clause fragments | 0 | Adverbial clause fragments | 15 |
| | Total | 707 | | 3359 |

Further comparison (*Table 7*) reveals that the proportion of noun phrase + of phrase fragment lexical bundles produced by native speakers is extremely high, accounting for nearly 40% of the total frequency, while the output rate of Chinese learners is less than 14%. The main difference is that Chinese English learners output far less of the noun phrase + of phrase fragment lexical bundles that native speakers produce more frequently, such as “part of the”, “the use of “. However, the overuse of “the development of the” may be closely related to the rapid development trends of China’s economy, culture, and other aspects. From the perspective of the part-of-speech composition of the lexical bundles, the high-frequency lexical bundles used by native speakers are dominated by the preposition “of”, as the category sum of the noun phrase + of phrase fragment and prepositional phrase + of phrase fragment reaches 1868, accounting for about 57% of the total frequency; while for Chinese English learners, the sum of these two categories is only 169, accounting for approximately 29% of the total frequency. This is closely related to the different structural features of English and Chinese, namely compared to English, which is a language dominated by subordinate structures, Chinese emphasizes parallel structures, so the output frequency of subordinate structures in “of” is significantly lower than that of native speakers. That tendency may reflect the negative transfer of Chinese as mother tongue for English learners.

Zero-frequency lexical bundles produced by Chinese English learners are listed in *Table 8*. It is obvious that although the 54 English lexical bundles listed are zero-frequency in ST6, they are an extremely important component for native speakers, as they cover 10 major categories of lexical bundles including anticipatory it + verb

phrase/adjective phrase, other prepositional phrase fragments, noun phrase + other post-modifying fragments, noun phrase + of phrase fragment, prepositional phrase + of phrase fragment, adverbial clause fragment, pronoun/noun phrase + be (+...), passive verb + prepositional phrase fragment, (verb phrase+) that clause fragment, and (verb/adjective+) to phrase fragment, with a total frequency of 1281, accounting for about 38% of the lexical bundle extraction frequency of native speakers. Among them, the frequency of lexical bundles such as "the secretary of state", "percent of the", "the way in which", "at the time of", "that there is a" is quite high. The reasons mainly lie in the different application trends of function words like articles and prepositions, which lead to differences in output rate (eg. "as a result of" vs. "as a result of the"; "the development of the" vs. "in the development of"; "the number of" vs "in the number of"); Chinese students tend to underuse lexical bundles containing nominal and adjectival clause structures (such as "by the fact that", "it is clear that", "to ensure that", "the extent to which", "the way in which", "the ways in which", which are high-frequency lexical bundles for native speakers but zero-frequency in ST6); Different from native speakers, Chinese students tend to underuse English lexical bundles containing passive voice structures (such as "is said to be", "it has been shown", "it has been suggested", "it should be noted", "should be noted that", which are absent in ST6 and Chinese students only produce relatively simple passive structures). Overall, Chinese English learners tend to "avoid the difficult but focus on the easy" in the output of lexical bundle types. In other words, they overuse what is easy, and underuse what is difficult, which is virtually the effect of the learners' avoidance strategy. Consequently, this strategy is bound to negatively affect the logical rigor and sentence structure level of Chinese learners' English output.

***Table 8**. Frequency list of 54 zero-frequency lexical bundles from ST6 in the BNC sample.*

| Lexical bundle | Freq. | Lexical bundle | Freq. | Lexical bundle | Freq. |
|---|---|---|---|---|---|
| the secretary of state | 404 | by the fact that | 15 | has been suggested that | 8 |
| percent of the | 113 | the role of the | 14 | in the development of | 8 |
| the way in which | 71 | an increase in the | 13 | is shown in figure/fig | 8 |
| at the time of | 44 | in the number of | 13 | it should be noted | 8 |
| that there is a | 41 | the extent to which | 12 | the position of the | 8 |
| the presence of | 35 | in the same way | 11 | the presence of a | 8 |
| as a result of the | 32 | the use of the | 11 | the structure of the | 8 |
| to ensure that the | 30 | the surface of the | 9 | as shown in figure/fig | 7 |
| is likely to be | 26 | as a function of | 9 | the value of the | 7 |
| the time of the | 24 | has been shown to | 9 | there was no significant | 6 |
| in the form of | 23 | in such a way | 9 | in the presence of | 5 |
| the size of the | 23 | is said to be | 9 | in a number of | 5 |
| it is important to | 22 | it has been shown | 9 | in the present study | 5 |
| in the context of | 22 | it has been suggested | 9 | should be noted that | 5 |
| it is clear that | 19 | the basis of the | 9 | the relationship between | 5 |
| it is possible to | 18 | the results of the | 9 | the shape of the | 5 |
| as part of the | 18 | the ways in which | 9 | at the level of | 4 |
| in the absence of | 16 | as we have seen | 8 | the division of labour | 3 |

## Conclusion

Based on concordance statistics from two comparative corpora, this study systematically examined the distribution patterns of high-frequency lexical bundles produced in the written English of Chinese advanced learners and native speakers. The findings reveal that compared to native speakers, Chinese learners tend to employ lexical bundles in their written English that display unique interlanguage features. Although there was no statistically significant difference in the overall frequency of

bundle tokens between the two groups, the number of bundle types showed more apparent divergence, with a considerable number of zero-frequency bundles for Chinese learners that are frequently used by native speakers. Approximately 40% of the high-frequency bundles shared between the CLEC and the BNC exhibited significant differences in usage frequency. By comparing the high-frequency and zero-frequency bundle lists of Chinese advanced learners, it is revealed that the learners' bundle output tendencies are closely related to the part-of-speech composition and syntactic structures of the bundles. The underlying reasons are mainly associated with L1 transfer and avoidance strategies. Evidently, the most preferred English bundles produced by Chinese learners are largely direct equivalents of Chinese idiomatic expressions (e.g., "in order to" corresponding to "为了" in Chinese, "on the other hand" corresponding to "另一方面" in Chinese). In addition, when the expression is too complex or poses much risk in misuse, the learners tend to intentionally avoid using certain bundles (e.g. "the extent to which", "as a function of"). Furthermore, the constraints of interlanguage proficiency and insufficient bundle richness also lead Chinese students to overuse certain bundles.

To sum up, this study primarily conducted descriptive statistical analyses of the relevant frequency data, aiming to explore the distribution patterns and features of lexical bundles produced by Chinese advanced English learners, providing reference for future studies on Chinese learners' lexical bundle output. The findings from the study on the usage patterns of lexical bundles by Chinese advanced English learners offer several valuable insights for English language teaching, including the design of academic English courses for new graduate students and so on. Moreover, in the design process of all relevant courses and teaching practices, instructors should take the negative impact of these distinctive patterns into considerations. The following are some specific pedagogical implications derived from the research: (1) Emphasizing the importance of lexical bundles in curriculum design. Given the significant role that lexical bundles play in fluent and natural language use, it is essential to integrate the teaching of these bundles into the curriculum. Therefore, language instructors should focus on both the recognition and production of high-frequency lexical bundles. (2) Incorporating corpus-based learning. The use of corpora in language teaching has proven to be an effective method for exposing learners to authentic language use. By incorporating corpus-based activities into the classroom, learners can explore real-life examples of lexical bundles. Teachers can guide students in using corpus tools to analyze the frequency, patterns, and contexts of lexical bundle usage. Such activities not only enhance learners' lexical knowledge but also develop their analytical skills, enabling them to become more autonomous learners. (3) Promoting productive practice through writing tasks. To improve learners' productive skills, especially in academic writing, it is crucial to provide ample opportunities for them to practice using lexical bundles. Writing tasks that require the use of specific bundles can help learners become more comfortable and confident in their usage. (4) Developing teaching materials and fostering a lexically rich environment. The creation of teaching materials and resources that focus on lexical bundles can greatly enhance the effectiveness of instruction. Teachers can develop glossaries, flashcards, and digital resources that highlight common lexical bundles and their uses. Additionally, incorporating multimedia resources, such as videos and podcasts, can provide learners with varied and rich input, illustrating how lexical bundles are used in different contexts. To create a classroom environment that is rich in lexical input can facilitate the acquisition of lexical bundles.

In conclusion, the pedagogical implications drawn from the study underscore the importance of lexical bundles in language learning and teaching. By incorporating these strategies into teaching practices, educators can significantly enhance the learners' linguistic competence, particularly in terms of fluency, accuracy, and appropriateness in language use. Meanwhile, there may be some limitations in this study. It is recommended that employing larger-scale sampled corpora or listing and comparing all the English bundles and corresponding frequencies in the two corpora, which would yield more valid and reliable conclusions.

## Acknowledgement

This research is self-funded.

## Conflict of interest

The authors confirm that there is no conflict of interest involve with any parties in this research study.

## REFERENCES

[1] Biber, D., Johansson, S., Leech, G., Conrad, S., Finegan, E. (1999): Longman grammar of spoken and written English. – Longman 1203p.
[2] Biber, D., Johansson, S., Leech, G.N., Conrad, S., Finegan, E., Quirk, R. (2021): Grammar of spoken and written English. – John Benjamins Publishing Company 1220p.
[3] Birhan, A.T. (2021): Effects of Teaching Lexical Bundles on EFL Students' Abstract Genre Academic Writing Skills Improvement: Corpus-Based Research Design. – International Journal of Language Education 5(1): 585-597.
[4] Code, C. (2023): Foundations of Familiar Language. Formulaic Expressions, Lexical Bundles, and Collocations at Work and Play: Diana Sidtis. – John Wiley & Son: Chichester, UK 464p.
[5] Cowie, A.P. (1998): Phraseology: Theory, analysis, and applications. – Oxford University Press 272p.
[6] Ellis, R. (1997): Second language acquisition. – The United States: Oxford 147p.
[7] Fengxiang, F. (2010): Data Processing and Management for Quantitative Linguistics with Foxpro. – RAM-Verlag 233p.
[8] Güngör, F., Uysal, H.H. (2020): Lexical bundle use and crosslinguistic influence in academic texts. – Lingua 242(2): 22p.
[9] Hyland, K. (2008): Academic clusters: Text patterning in published and postgraduate writing. – International Journal of Applied Linguistics 18(1): 41-62.
[10] Kim, S., Kessler, M. (2022): Examining L2 English university students' uses of lexical bundles and their relationship to writing quality. – Assessing Writing 51: 11p.
[11] Michael, L. (1993): The Lexical approach: the state of ELT and a way forward. – Language Teaching Publications 200p.
[12] Nattinger, J.R., DeCarrico, J.S. (1992): Lexical phrases and language teaching. – Oxford University Press 218p.
[13] Pan, F., Reppen, R., Biber, D. (2020): Methodological issues in contrastive lexical bundle research: The influence of corpus design on bundle identification. – International Journal of Corpus Linguistics 25(2): 216-230.

[14] Ren, J. (2021): Variability and functions of lexical bundles in research articles of applied linguistics and pharmaceutical sciences. – Journal of English for Academic Purposes 50: 16p.
[15] Ruan, Z. (2017): Lexical bundles in Chinese undergraduate academic writing at an English medium university. – RELC Journal 48(3): 327-340.
[16] Schmitt, N., McCarthy, M. (Eds.) (1998): Vocabulary: Description, acquisition and pedagogy. – Cambridge University Press 393p.
[17] Shin, Y.K. (2019): Do native writers always have a head start over nonnative writers? The use of lexical bundles in college students' essays. – Journal of English for Academic Purposes 40: 1-14.
[18] Shirazizadeh, M., Amirfazlian, R. (2021): Lexical bundles in theses, articles and textbooks of applied linguistics: Investigating intradisciplinary uniformity and variation. – Journal of English for Academic Purposes 49: 13p.
[19] Wray, A. (2002): Formulaic language and the lexicon. – Cambridge University Press 327p.
[20] Yakut, I., Yuvayapan, F., Bada, E. (2021): Lexical bundles in L1 and L2 English doctoral dissertations. – Journal of Teaching English for Specific and Academic Purposes 18p.