Four-Word Bundles in English Abstracts of Chinese and English Linguistics Journal Articles: A Corpus-based Comparative Study

ZHOU Yi-tao
School of Foreign Languages and Literatures, Southwest Minzu University, Chengdu, China

As the most frequent word sequence in register, lexical bundles are the basic components of discourse. Their use is of great help to the fluency and accuracy of academic writing. In the past two decades, the study of lexical bundles has focused on the body part of academic papers, but less on the abstract. In order to identify and compare the differences in the use of lexical bundles in English abstracts between international and Chinese authors, this paper selects English abstracts from Chinese and International high-level academic journals in the discipline of linguistics from 2018 to 2020 to build two corpora of English abstracts of academic writings: Corpus of English Abstracts from International Linguistics Journal Articles (ILJA) and Corpus of English Abstracts from Chinese Linguistics Journal Articles (CLJA). This paper analyzes the differences of English abstract writing between Chinese and international high-level linguistics journals, and provides enlightenment for abstract writing and teaching.

Keywords: English abstracts, linguistics journal, corpus, four-word bundles

Introduction

As a substantial academic genre, abstracts “constitute the gateway that leads readers to take up an article, journals to select contributions, or organizers of conferences to accept or reject papers” (Lores, 2004). As an integral part of an academic discourse, it has its own structure and textual characteristics, which can convey the central idea of a research article. The abstract also provides a preliminary overview for those who may need to decide whether to read the entire article or not. In addition, in case that readers have read this article before, abstracts can help them recall the basic content of the article quickly.

Multi-word expression is an important component of fluent linguistic production and a key factor in successful language learning (Hyland, 2008). It has been studied under many rubrics, which Biber, Johansson, Leech, Conrad, and Finegan (1990) call lexical bundles and Scott (1996) refers to as clusters. As a unique linguistic construct and the most frequent sequences of words in a register, lexical bundles usually function as basic building blocks of discourse (Biber et al., 2004). Coxhead and Byrd (2007) pointed out that these lexical bundles are initial to language users for at least three reasons. First, the repetition of lexical bundles offers users (and particularly language learners) ready-made sets of words to work with. Second, they help define fluent use and therefore expertise and legitimate disciplinary membership. Third, they reveal the lexicogrammatical community-authorized ways of making-meanings. Despite this, typically overusing a limited number of familiar

ZHOU Yi-tao, graduate student, School of Foreign Languages and Literatures, Southwest Minzu University.
phrases and lacking a diverse enough multi-word sequences in a native-like manner often become shortcoming to
domestic scholars in abstract writing (Li & Schmitt, 2009).

There have been many linguistic studies of academic writing published over the past 30 years. The research
on lexical bundles follows the pioneering work of Bengt Altenberg (1998), who creates a method to identify the
recurrent word combination defined by frequency, and classifies it by combining grammar and functional
analysis. Obviously, a method that only identifies and classifies formulaic units based on frequency and usage
breadth has a clear methodological advantage, although researchers use different frequency and distribution
criteria. Hyland (2008) argued that bundles are not only central to the creation of academic discourse, but that
they offer an important means of differentiating written texts by discipline. While most of the previous studies
focus on the academic research articles in science and medicine (Wei, 2015; Huxin, 2015, 2017). Few researches
have surveyed to what extent the lexical bundles in Chinese scholars’ English abstracts are comparable to those
of international experts in specific disciplines. Therefore, this paper takes a frequency-driven lexical bundle
approach and compares the four-word bundles used in journal article of English abstract writing by international
writers and Chinese scholars in the discipline of linguistics. The study offers a realistic and accurate description
of the lexical bundle use in abstracts, providing a valuable reference for Chinese scholars and practical
pedagogical implications for teachers and learners.

I. Previous Research on Lexical Bundles

1.1 The Concept and Definition of Lexical Bundles

Biber (1999) refers to these multi-word sequences as ‘lexical bundles’, defined simply as the most frequent
recurring sequences of words. Lexical bundles can be regarded as extended collocation: bundles of words that
show a statistical tendency to co-occur. The research on lexical bundles follows the pioneering work of Bengt
Altenberg (1998), who creates a method to identify the recurrent word sequences defined by frequency, and
classifies it by combining grammar and functional analysis. Obviously, a method that only identifies and
classifies formulaic units based on frequency and usage breadth has a clear methodological advantage, although
researchers use different frequency and distribution criteria. The term ‘lexical bundle’ was first used in the
Longman Grammar of Spoken and Written English (Biber et al., 1999, Chapter 13), which compared the
recurrent sequences of words in conversation and academic prose.

1.2 Classification of Lexical Bundles

In a series of lexical bundle studies conducted by Biber and colleagues (Biber & Barbieri, 2007; Biber &
Conrad, 1999; Biber, Conrad, & Cortes, 2003, 2004; Biber, Johansson, Leech, Conrad, & Finegan, 1999), it was
found that conversation and academic prose present distinctive distribution patterns of lexical bundles. For
example, most bundles in conversation are clausal, whereas most bundles in academic prose are phrasal. Since
some overlap among these structural categories of lexical bundles in the Longman Grammar of Spoken and
Written English, “NP-based”, “PP-based”, and “VP-based” broad structural categories were later distinguished
by Chen and Baker (2010).
Table 1

*The Structural Categories*

| Category          | Bundle structure                                           | Example                  |
|-------------------|-----------------------------------------------------------|--------------------------|
| NP-based bundles  | Noun phrase with post-modifier fragment                    | *the way in which*       |
|                   | Noun phrase with *of* fragment                             | *the nature of*          |
| PP-based bundles  | Preposition + *of* fragment                                | *as a result of*         |
|                   | Other prepositional phrase fragment                        | *on the other hand*      |
| VP-based bundles  | Copula *be* + NP/AdjectiveP                                | *is one of the*          |
|                   | VP with active verb                                        | *has a number of*        |
|                   | Anticipatory *it* + VP/AdjectiveP + (complement-clause)    | *it is possible to*      |
|                   | Passive verb + PP fragment                                 | *is based on*            |
|                   | *(VP +) that-clause fragment*                              | *should be noted that*   |
|                   | *(verb/adjective +) to-clause fragment*                    | *are likely to be*       |
|                   | VP with active verb                                        | *has the potential to*   |
|                   | Pronoun/Noun + verb                                        | *this article describes a*|
| Other             | Other bundles                                              | *as well as the*         |

1.3 Previous Studies on Lexical Bundles

Ellis and Simpson Vlach (2008) build a relatively complete academic lexical bundle list based on larger corpus, which provides reference for lexical bundle teaching. This exploration of establishing academic lexical bundle list set a good example for later research. On the basis of Biber’s classification of lexical bundles, Chen and Baker (2010) summarized three structural categories: lexical bundles based on noun phrases (NP-based bundles), lexical bundles based on prepositional phrases (PP-based bundles) and lexical bundles based on verb phrases (VP-based bundles) which established a foundation for the follow-up study. Ellis (2008) uses the empirical method to analyze the effectiveness of the teaching of lexical bundles. At present, there are relatively few empirical studies on lexical bundles, so his research can be used for reference in academic English teaching. Biber (2009) examined the lexical bundles used in spoken and written language, and the patterns in conversation tend to be fixed sequence (combinations of content words and function words).

Previous studies have explored the differences in the use of bundles from the aspects of academic writing genre, article type, different subject categories, etc. few studies have focused on high-level journal papers at home and abroad. Moreover, in recent years, the study of lexical bundles in academic terminologies mainly focuses on the introduction and full text, while the study of abstracts is relatively less. Therefore, this study explores the use of bundles in abstracts of high-level journals at home and abroad.

II. Research Design and Methodology

2.1 Research Question

This study selects English abstracts of Linguistics articles from Chinese and international linguistic journal articles for comparative analysis, aiming to answer the following questions: (1) What are the four-word bundles frequently used in the two self-built corpora of English abstracts respectively? (2) What are the main structural differences between these frequently used four-word bundles in the two self-built corpora of English abstracts?
2.2 Research Materials

According to the impact factor analysis report of Linguistics journals in Journal of Citation Report, this study selects the following five journals as the self-built corpus: *Applied Linguistics, Journal of Memory and Language, Language Teaching, Language Learning, Journal of Second Language Writing*. All research papers published from 2018 to 2020, excluding review articles and book reviews, and those with less than 100 words in each abstract are not included. A total of 473 abstracts were selected to build Corpus of English Abstracts from International Linguistics Journal Articles (ILJA). Based on the impact factors published by CNKI, the English abstracts of Chinese scholars’ papers are selected from the research papers of Linguistics in six foreign language journals with high impact factors, excluding review articles, book reviews, conference articles, etc. They are *Modern Foreign Languages* (《现代外语》), *Foreign Language World* (《外语界》), *Foreign Language Education* (《外语教学》), *Foreign Language Learning Theory and Practice* (《外语教学理论与实践》), *Foreign Language Teaching and Research* (《外语教学与研究》), and *Foreign Languages in China* (《中国外语》). The work place of the first author is a domestic university or institution and the first author is from Chinese mainland. If the tokens in each abstract is less than 100, it will not be included. A total of 610 abstracts were selected from 2018 to 2020 to build the Corpus of English Abstracts from Chinese Linguistics Journal Articles (CLJA).

Table 2

| Corpus       | Representation                        | Word count | Average length of text | No. of texts |
|--------------|---------------------------------------|------------|------------------------|--------------|
| ILJA         | International Linguistics Journal Articles | 80,236     | 170                    | 473          |
| CLJA         | Chinese Linguistics Journal Articles   | 80,838     | 132                    | 610          |

2.3 Research Method

In this study, the key criteria regarding how to generate a list of lexical bundles using automated corpus tools are set up from the three aspects as follows: The first criterion is to select the four-word bundles. The second criterion is the cut-off frequency, the set at a cut-off frequency of 5 occurrences (about 50 times per million words) and the CLJA, a higher cut-off frequency of 7 occurrences (about 68 times per million words) since the texts are 1/4 more than the ILJA. This setting is to ensure the representativeness of the extraction of lexical bundles. The last issue concerns the requirement that lexical bundles have to occur in different texts. Since the cut-off frequency is 5 and 7 times for respective corpus, the distribution criterion in this study is at least 5 texts for ILJA and at least 7 texts for CLJA. In the data collection phase, this study is using free software Antconc 3.2.1 to extract data.

III. Research and Discussion

3.1 General Presentation of Research Results

Based on the above criteria, lexical bundles retrieved from international journals corpus and Chinese journal corpus by software Antconc 3.2.1. There are 52 target bundles are found in IJLA and 54 target bundles are found in CLJA.
As can be seen in Table 3, the most frequent word bundle in ILJA is *the extent to which*, which occurs 20 times in 19 abstracts. Next, *in the context of*, *this study investigated the* and *the results show that* occur 15, 14 and 11 times respectively. On the other hand, it can be seen from Table 4 that *the results show that* has the highest frequency of 52 times, distributed in 52 abstracts. And *it is found that*, *on the basis of* and *from the perspective of* are used 40, 37, and 36 times respectively.

Table 5

*The Shared Bundles*

| Bundles shared                      | Frequency in ILJA | Frequency in CLJA |
|-------------------------------------|-------------------|-------------------|
| The results show that               | 11/3%             | 52/7.9%           |
| As well as the                      | 6/1.6%            | 20/3.3%           |
| In the field of                     | 7/1.9%            | 12/1.8%           |
| Can be used to                      | 6/1.6%            | 7/1.1%            |
| On the development of               | 5/1.3%            | 8/1.2%            |

As a total of 5 four-word bundles (9%) is shared by the two corpora. Even the shared bundles occur in both materials, it doesn’t mean that it is used equally frequently by Chinese and international scholars.
The structural framework for this research is adopted from Biber et al. (1999)’s structural categories and developed by Chen and Baker (2010). There are four main categories: NP (noun phrase)-based bundles, PP (preposition phrase)-based bundles, VP (verb phrase)-based bundles, and other bundles. Figure 1 is the structural distribution in International Abstract Corpus and Chinese Abstract Corpus.

Figure 1 shows that VP-based bundles are the most frequently used structure in CLJA while NP-based bundles and PP-based bundles have high frequency in ILJA. The number of PP-based bundles used by international scholars is significantly higher than that of domestic writers. Compared with international scholars, Chinese scholars overuse the VP-based bundles. Chinese scholars tend to use the “noun phrase + verb phrase fragment” structure (such as the results indicate that, this paper explores the), which is similar to the “subject + predicate” pattern in Chinese and reflects the traces of Chinese transfer. Chen, Baker (2010) and Pan (2016) also found that expert scholars tend to use phrase bundles, on the contrary, lower-level writers tend to use clause bundles, which mainly including noun + verb phrase fragment and anticipatory it + VP/adjective phrase. Noun phrase with of phrase fragment used frequently both in ILJA and CLJA. Noun phrases embedded with prepositional post-modifiers are pervasive in abstract writing. Previous research (e.g., Biber & Gary, 2010; Biber et al., 2011) also found that academic writers depend on prepositional post-modifier, especially of-phrase, to adds extra information and achieve grammatical complexity, signifying the “compressed” style of academic.

Conclusion

In this study, I built two corpora of English abstract from Chinese and international Linguistics journal articles. Through a comparative analysis of the structure and function of lexical bundles in the two corpora, the results show that Chinese scholars tend to employ some lexical bundles excessively or insufficiently. The loose structure of the abstract is due to the overuse of clause bundles (VP-based bundles). In terms of function, Chinese and international scholars use the highest proportion of text-oriented bundles, followed by research-oriented
bundles and the lowest proportion of participant-oriented bundles. These findings are helpful for second language learners to master the writing norms of abstracts in high-level academic journals and to use bundles properly. Therefore, teachers should teach according to the context and pragmatic function of bundles in native speakers’ academic writing corpus, and strengthen writing practice. Students should increase the awareness of lexical bundles, increase the amount of reading and learn to summarize the commonly used bundles. In view of the problems of small-scale corpus and manual annotation error, it is suggested that large-scale corpus should be built in the future, and more professional help should be used for manual annotation to ensure the reliability of experimental data.

References

Altenberg, B. (1998). On the phraseology of spoken English: The evidence of recurrent word-combinations. In A. Cowie (Ed.), Phraseology: Theory, analysis, and applications. UK: Oxford University Press.

Biber, D. A. (2009). Corpus-driven approach to formulaic language in English: Multi-word patterns in speech and writing. International Journal of Corpus Linguistics.

Biber, D., Conrad, S., & Cortes, V. (2004). If you look at: Lexical Bundles in University Teaching and Textbooks. Applied Linguistics.

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

Chen, Y. H., & Baker, P. (2010). Lexical bundles in L1 and L2 academic writing. Language Learning & Technology.

Ellis, N., Simpson-Vlach, R., & Maynard, C. (2008). Formulaic language in native and second language speakers: Psycholinguistics, corpus linguistics, and TESOL. TESOL Quarterly.

Hyland, K. (2008). As can be seen: Lexical bundles and disciplinary variation. English for Specific Purposes.

Lores, R. (2004). On RA abstracts: From rhetorical structure to thematic organization. English for Specific Purposes.

Wei, Y. Y. (2015). Lexical Bundles in native and non-native scientific writing: Applying a corpus-based study to language teaching. ELT Journal.

胡新. (2015). 中外科技论文英文摘要的语步词块特征对比研究. 现代外语.