Structural Analysis of Lexical Bundles in English News on Health from China and UK Newspapers

Mengyu Xu[1,a], Tingting Sun[2,*]

[1]English Language and Literature, Korea Maritime and Ocean University, Busan, 612022, Korea

[2]School of Foreign Languages, Dalian Maritime and Ocean University, Dalian, 116000, China

xumengyu@g.kmou.ac.kr

*corresponding author

Keywords: Lexical bundles, Structural features, News corpora

Abstract: News, as a timely report genre, concludes many parts, such as health, event and so on. Furthermore, the study of lexical bundles has always been a hot issue in English studies. Therefore, this paper uses two self-built corpora: Xinhuanet corpus and The Times corpus, to compare the structural features of lexical bundles to have a clear understanding of lexical bundles. It is found that noun, preposition, and verb lexical bundles are widely used in both Chinese and Western news reports, while dependent clause fragments are rarely employed; however, Western news uses fewer verb structures compared with Chinese news.

1. Introduction

In the past 20 years, the fixed feature of language has attracted scholars in mounting numbers. Howarth(1998) points out that single words can be ensured as one kind of lexical bundle (LB), and what is worth mentioning is that there is no clear relationship between the words and their meanings, use ways or even frequencies. Since the middle of the last century, research and related topics on LBs have been endless and have attracted many scholars' attention. Scholars research not only different language styles, such as spoken and written language, but also LBs in different fields, such as lectures, terminology articles, and textbooks. Nonetheless, there are relatively few LBs researches on news reports. News reports refer to reporting on objective facts that have recently occurred, which means the content reflected in news information must truly convey the facts. The standard news sections are international, education, entertainment, health, etc. Based on the discussion above, this paper aims to analyze the structure of LBs on health news between the news from the Xinhuanet website and The Times website, which are the traditional and official representatives of Chinese and Western news media.

2. Literature Review

2.1 The Definition of Lexical Bundles and Significance

Nowadays, a sea of scholars have begun to research LBs, and thus different learned men pointed out various definitions of LBs. For example, Biber et al.(1999) put forward the "lexical bundle" item. Moreover, countless scholars deem LBs can be defined as more than two words' collocations or combinations. However, based on the view of Jiang (2019), two-word LBs can also be seen as lexical bundles. Lexical bundles, in short, are relatively fixed combinations of words belonging to multi-word units, multi-word expressions, or similar terms (such as cluster, prefab, or prefabricated chunk), formulaic sequence, n-gram, etc. (Liang et al. 2010:13). Cortes (2004) believes that the LBs frequency and range are higher than random chance. According to Biber et al.(1999), there is only nearly 15% of the lexical bundles are integrated structure unites, for example, "I can do it." and " This is my house."

Some evidence shows that numerous LBs are defined as settled bundles, compared with single

words or chunks without fixed features, and LBs have processing advantages. In the research on reading tasks, it is discovered that if sentences contain LBs, the reading space will be faster than the reading space in sentences without LBs because Jiang (2019) pointed out that LBs can shape the meaning of the text in a specific context as to enhance coherence and readability. From the aspect of features, lexical bundles have the following characteristics: familiar concurrency of more than two words; lexical bundles can be adjusted with the change of time and situation; other words can be added to enhance comprehension and grammatical accuracy.

2.2 Theological Framework

LBs research usually involves two aspects, namely recognition and classification. Recognition usually includes three levels, word number contained in the lexical bundles (such as three-word lexical bundles, four-word lexical bundles, or more than five-word lexical bundles), the frequency of LBs, and the text's coverage. Ensuring LBs frequency and coverage is a traditional and necessary step in LBs research because this can make the research results clear and accurate. Nevertheless, there is no uniform rule for frequency setting and text coverage so far. Some scholars adopted ten times per million words, and other authors adopted 20 appearances per million words. There are also researchers choosing the "most conservative" way. Biber & Barbieri (2007) used 40 times per million vocabularies as the criteria for frequency setting. The selection of LBs coverage aims to reduce the particularity of using specific lexical bundles in a single text, but there is no consolidated standard like frequency. Furthermore, they believe that three to five texts are correctly selected as the more appropriate coverage. Hyland (2012) used a ten per cent floating range as the coverage.

Classification is another critical link in lexical research, which usually includes structural and functional classification. Biber et al. (2004), the most significant contribution, divides lexical bundles into two major categories: structural and functional types. These structural taxonomies include verb phrase-based bundles (VP-based bundles), such as "I am not going to", "and this is a", "I mean I do not", and "have a lot of", "is based on the"; dependent clause fragments, like "I want you to", "when we get to", "if we look at", "to be able to"; noun/prepositional phrase-based bundles (NP/PP-based bundles), such as "one of the things", "the way in which", "a little bit more", and "at the end of".

2.3 Previous Studies

Biber et al.(2004) found that the frequency of LBs in classroom discourse was the highest, about three times that of academic papers and four times that of university textbooks. However, in terms of structure types, the proportion of three types of LBs in classroom discourse is balanced, accounting for about one-third. In textbooks and academic papers, noun phrase-based bundles make up a large proportion. Verb phrase-based bundles are less than one-sixth, and the proportion of dependent clause fragments is the lowest. In class discourse, the lexical bundles related to clauses mostly contain declarative and interrogative sentences, and the lexical blocks related to noun phrases and prepositional phrases also appear more frequently.

Hyland (2008) found that although a declining trend in academic papers in the past decades, LBs belonging to noun phrases were still the main form of LBs in academic discourse. However, as the number of VP-based bundles increases yearly, the number of clause-related bundles remains stable. Based on the self-built corpus of journal papers and master or doctoral dissertations on electronic engineering, biology, business and applied linguistics, Hyland found that the proportion of text-oriented lexical bundles in multi-disciplinary academic papers was the largest, followed by research-oriented and interpersonal-oriented lexical bundles, respectively. In his corpus analysis, the proportion of research-oriented lexical bundles in electronic engineering and biology, which belong to science subjects, is higher than that in the other two types of lexical bundles, while the frequency of text-oriented lexical blocks in the two humanities subjects is the highest.

Jiang(2019) found that science subjects reduce the use of research-oriented lexical bundles while constantly increasing interpersonal and text-oriented lexical bundles to meet public readers' communication and interaction needs. With the deepening of the empirical research paradigm, the humanities subjects gradually increase the use of research-oriented lexical bundles.

3. Methodology

3.1 Research Questions

(1). Are there some similarities between the two corpora on structure?
(2). Are there some differences between the two corpora on structure?

3.2 Corpora

The data included two corpora, one is the corpus of Xinhuanet, and the other is the corpus of The Times. The contents of the two corpora are searched from the official websites, and the two URLs are http://www.xinhuanet.com/ and https://www.thetimes.co.uk/, respectively. According to the search for the health news part in the two URLs, the texts can be gotten, and two corpora can be established manually. The time range of these two corpora is from February 1st, 2021, to May 15th, 2021. After statistics, the text number and total words in the corpus of Xinhuanet are 663 and 228221 separately; and those in the corpus of The Times are 89 and 52522. The average text lengths are 344.22 words and 590.13 words separately. These corpora remove news headlines, authors, publication time, pictures, and other related information and only keep the article body.

3.3 Lexical Bundle Identification

Lexical bundle identification is a necessary process of lexical bundle research. Lexical bundles identification mainly includes three aspects: the lexical bundle kind, the frequency of LBs and the text's coverage.

In this study paper, four-word lexical bundles are used. The reason for adopting four-word lexical bundles is that the number of four-word LBs is more than the number of five-word LBs. Moreover, Hyland(2008) mentioned that in contrast to three-word bundles, four-word LBs could offer more explicit features, structures and functions. Cortes(2004) indicated that although four-word bundles are easily found from sorter strings, they are not accessible from longer strings.

Based on what has been mentioned above, the most old-fashioned and conservative way, 40 times per million words, is adopted in this paper. Furthermore, the last one is the minimum range. The minimum range of the corpus of The Times is five, and due to the reason that the text number of the corpus of Xinhuanet is five times the size of the corpus of The Times, the minimum range of corpus of the times is 35.

The tool for generating lexical bundles was WordSmith 8. Furthermore, selecting those LBs that met the criteria is convenient and discarding the rest. The details are as follows.

Table 1 Details of Two Corpora.

	Study corpus	Reference corpus
Number of texts	663	89
Total number of tokens	228221	52522
Average text length	344.22	590.13
Standardized TTR	30.10%	44.65%
Min frequency=40%	9	2
Min range	35	5
Number of lexical bundles	534	30

4. Results and Discussion

4.1 Shared Bundles

The shared bundles in the two corpora are four, namely by the end of, since the start of, the number of people, the start of the, accounting for 0.75% of the total 534 LBs in the Xinhuanet corpus, and 13.33% of the 30 LBs in The Times corpus. Among the four typical LBs, the total proportion of noun phrase-based bundles and prepositional phrase-based bundles is 50%, respectively. No standard LBs coexist in the form of verb phrase-based bundles and dependent clause fragments. The following table 2 shows the concrete situation of four shared bundles.

Table 2 Concrete Situation Of Four Shared Bundles Four

The shared LBs	The sequence number in The times corpus	The sequence number in the Xinhuanet corpus	Structural categories
by the end of	2	193	PP-based bundles
since the start of	11	202	PP-based bundles
the number of people	14	329	NP-based bundles
the start of the	16	356	NP-based bundles

(1). “I think it will be very premature, and I think unrealistic, to think that we're going to finish with this virus by the end of the year,” Dr Michael Ryan, executive director of the WHO's Health Emergencies Program. (study corpus)

(2). By the end of March, we expect an additional 6 million people will receive the first dose,” he said. (study corpus)

(3). British Health Secretary Matt Hancock confirmed the rolloutahead of schedule as the country aims to vaccinate all adults by the end of July. (study corpus)

(4). The vast majority of over-85s will have received their vaccine by the end of the week, the Department of Health said. (reference corpus)

(5). A spokesman said: “We have supplied over 300,000 nasal pharyngeal rapid antigttttttttestsests to the private sector and estimate that by the end of April to have doubled this number as industries look to get back to business.”(reference corpus)

(6). In Sao Paulo, 92 per cent of hospital beds were occupied by the end of last week. (reference corpus)

From the above example, it can be seen that the lexical bundles of “by the end of”, both in the learning corpus and the reference corpus, are often applied with time-specific vocabularies such as year, month, and week. These lexical bundles are more commonly collocated with months. In Chinese health news, “by the end of” is combined with the year, while in British health news, “by the end of” is combined with the week, which shows a small-range time.

According to the above sentences, it can be concluded that, in the reference corpus, the “the start of the” lexical bundle is simple, and all three sentences are combined with “pandemic” to indicate the beginning of the disease. However, in the study corpus, “the start of the” collocates with “diseases” and “months”, which is not as fixed as the reference corpus.

4.2 The Corpus of Xinhuanet

The details about the study corpus are following. There are 534 lexical bundles in the canon of Xinhuanet, in which there are 237 noun phrase-based bundles and 109 preposition phrase-based bundles, making up 64.79% of the whole LBs; 181 verb phrase-based bundles, accounting for 33.90% of the total LBs; seven dependent clause fragments, accounting for 1.31% of the total number of LBs.

Xinhuanet corpus has 85 complete LBs, accounting for 15.92% of the 534 LBs in the Xinhuanet corpus, more than the average proportion mentioned above. There are 56 noun phrase-based bundles, 16 verb phrase-based bundles, and 13 preposition phrase-based bundles, accounting for 65.88%, 18.82%, and 15.29%, respectively, and with no dependent clause fragments.

4.3 The Corpus of the Times

The time's corpus has 30 lexical bundles, in which there are 24 noun phrase or phrase-based piles, four VP-based bundles and two dependent clause fragments, accounting for 80%, 13.33%, and 6.67% singly. There are seven complete LBs in the time's corpus: 1) the South African variant; 2) the world health organisation; 3) the number of people; 4) Hancock, the health secretary; 5) in the coming weeks; 6) more than a year; 7) the oxford AstraZeneca vaccine.

Those account for 26.9% of the total 26 LBs in The Times corpus, in which 6 LBs are NP-based bundles; one lexical bundle is PP-based bundles without VP-based bundles and dependent clause fragments.

4.4 The Similarities between Two Corpora

Although the numbers of LBs in two corpora are different, in comparing the four structures in two other corpora, there is no doubt that, generally, noun phrase or phrase-based bundles account for a vast proportion, and the proportional distribution of dependent clause fragments is small. The reasons for the phenomenon appeared are following. First, generally speaking, there are more clause bundles in spoken materials, but in written English materials, noun phrases or preposition phrase-based piles are mainly used (Biber et al., 1999; Hyland, 2008). Besides, Biber et al.(1999) believed that text in the press reportage has high frequencies of nouns and prepositions. Thirdly, concerning the characteristics of news reports, truthfulness, conciseness, and timeliness are three apparent features. So, the average length of sentences in the news is shorter than in other genres, which determines less chance of using dependent clause fragments in news reports.

4.5 The Differences between Two Corpora

In the Xinhuanet corpus, VP-based bundles accounted for more than one-third of the total number of chunks; the proportion in the Times is significantly smaller than that of the Xinhuanet corpus, which is only about one-eighth. The results show that a little oral tendency of Chinese news editors is employed when constructing a news report. Foreign newspaper writers are better at changing verb phases into nouns or proportional phrases, but this feature in the Chinese news corpus is unclear.

5. Conclusion

This paper studies the typical four-word LBs according to two self-built corpora. Through comparison, it is found that both Chinese and British news reports on health use the same LBs order: NP/PP-based bundles in a more significant proportion, then followed by VP-based bundles, less dependent clause fragments, and the complete bundles in two corpora are higher than the average full bundles proposed by Biber et al. The distinction between the study corpus and reference corpus is that the English corpus uses less proportion of VP-based bundles than the corpus of Xinhuanet. The study contributes to a better understanding of news reports and provides a guideline for future studies. However, there are certain limitations because of the small number of words in the text of the two corpora and the limited time range of the selected text.

References

[1] Biber, D., Johansson, S., Leech, G., Conrad, S., Finegan, E., & Quirk, R. Longman grammar of spoken and written English. Longman, 1999.

[2] Biber, D.& F. Barbieri. Lexical bundles in University spoken and Written Registers. English for Specific Purposes vol.26, no.3, pp.263-286, 2007.

[3] Biber, D., S. Conrad & V. Cortes. If you look at...: Lexical bundles in university teaching and textbooks. Applied Linguistics vol.25, no.3, pp.371-405, 2004.

[4] Cortes, V. Lexical bundles in published and student disciplinary writing: Examples from history and biology. English for Specific Purposes, vol.23, no.4, pp.397-423, 2004.

[5] Howarth, P. Phraseology and second language proficiency. Applied Linguistics vol.19, no.1, pp. 24-44, 1998.

[6] Hyland, K. Bundles in academic discourse. Annual Review of Applied Linguistics.vol.32, pp.150-169, 2012.