Quantitative Analysis of Lexical Bundles Used in Spoken English and Japanese Elementary and Junior High School English Textbooks

Mikajiri, Noriaki

(Citation)
Journal of Corpus-based Lexicology Studies,5:1-23

(Issue Date)
2023-03-10

(Resource Type)
departmental bulletin paper

(Version)
Version of Record

(JaLCDOI)
https://doi.org/10.24546/0100479380

(URL)
https://hdl.handle.net/20.500.14094/0100479380

英語コーパス学会 語彙研究会(Japan Association for English Corpus Studies, Lexicology SIG)
Journal of Corpus-based Lexicology Studies, Vol. 5 (2022)
ISSN: 2434-169X

Quantitative Analysis of Lexical Bundles Used in Spoken English and Japanese Elementary and Junior High School English Textbooks

MIKAJIRI, Noriaki
(Tokyo University of Foreign Studies Graduate Student)
mikajiri.noriaki.w0@tufs.ac.jp

英語の話し言葉と日本の小学校・中学校の英語検定教科書で使用される レキシカル・バンドルに関する量的分析

三河尻 紀明(東京外国語大学 大学院生)

Abstract

One of the primary goals of English education in Japan is to develop communicative abilities, which require the ability to use formulaic sequences. This study analyzed a corpus of English textbooks used in elementary and junior high schools in Japan to investigate the extent to which they employ target-like lexical bundles, compared to SUBTLEXus, a spoken English corpus. The results showed that the textbooks used significantly more lexical bundles than spoken English, but only a few were common across multiple series of textbooks. The study also discovered that short bundles were commonly used in spoken English, while long bundles deviated from spoken English. These results suggest that textbooks may help students effectively use lexical bundles, but revisions can be made in selecting which bundles to include in the textbooks.

Keywords

corpus, formulaic sequence, lexical bundle, English textbook, spoken English

1. Introduction

In recent vocabulary studies, multiword language units, also known as formulaic sequences, have been one of the most actively researched areas (e.g., Nation, 2022). Researchers have been highlighting the importance of learning and using formulaic sequences for language learning for several decades (e.g., Bolinger, 1979; Nattinger, 1980; Pawley & Syder, 1983). The advent of corpora, computerized text databases of naturally occurring languages (McEnery et al., 2006), has shown that formulaic sequences are central to language. For example, Erman and Warren (2000)

proposed that more than 50% of a language consists of formulaic sequences. Such empirical studies have led to the view that formulaic sequences are a central component of a language. This phenomenon is referred to as "the idiom principle" by Sinclair (1991). Therefore, separating whether an expression is grammatically possible from whether it is likely to be used in actual language use is crucial for language learners. Hence, the importance of learning formulaic sequences has been emphasized in language learning and teaching (e.g., Boers et al., 2006; Lewis, 1993; Wray, 2002).

Lexical bundles, also called *n-grams* (e.g., Szudarski, 2018), are mainly investigated with corpora and are a type of formulaic sequence the present study focuses on. Previous corpus studies show that lexical bundles have the following five characteristics. First, lexical bundles appear in an extensive variety of registers, such as speech (e.g., Biber, 2009), classrooms (e.g., Biber et al., 2004; Csomay, 2012), and academic writing (e.g., Staples et al., 2013). Therefore, lexical bundles are universal in language use. Second, lexical bundles are frequently recurring multiword units that are automatically identified based on a frequency threshold. Biber et al. (1999, p. 990) explained that lexical bundles are "recurrent expressions, regardless of the idiomaticity, and regardless of their structural status." This explanation thus led to the third feature, that is, lexical bundles are often grammatically incomplete units, such as "you want to," and "there's a lot of." Biber et al. (1999) found that the ratio of grammatically complete lexical bundles was only 15% in conversation and less than 5% in academic prose. Fourth, lexical bundles are semantically transparent (e.g., Biber, 2009). In other words, lexical bundles have a transparency of their meanings, unlike other variations of formulaic sequences, such as idioms (e.g., "raining cats and dogs"), where the whole meaning cannot be deduced by adding up the meanings of the component words (i.e., it is impossible to infer the whole meaning "it is raining heavily" from "raining + cats + and + dogs"). However, in the case of lexical bundles, there is a transparency of meaning, where adding up the meanings of component words forms the whole meaning. Lastly, lexical bundles also have a variety of functions, which Biber et al. (2004) organized into three primary roles: stance expressions, discourse organizers, and referential expressions. Therefore, learning and utilizing lexical bundles are essential in language learning because of their use and function in such a wide range of contexts.

Although the term "lexical bundles" has not been mentioned explicitly, the importance of teaching formulaic sequences seems to be recognized in English education in Japan. For example, in the section of the content on English knowledge and skills that are required for students, the Course of Study (The Ministry of Education, Culture, Sports, Science and Technology, 2017a, 2017b, 2018) exemplifies various formulaic

sequences, such as “on the other hand,” which is also listed as a highly-frequent discourse organizing lexical bundles in Biber et al. (2004). The Course of Study, then, explains the importance of carefully selecting those formulaic sequences according to the situation (e.g., when expressing thoughts, answering the phone, or in a self-introduction) and teaching them to students. When teaching in the classroom, the Course of Study says that the goal of English classes is not only to increase knowledge of formulaic sequences but also to teach formulaic sequences to students to be used in the five domains of English use (i.e., listening, reading, spoken interaction, spoken production, and writing). It should be noted, however, that because the Course of Study does not define formulaic sequences precisely, those exemplified in the document include not only lexical bundles (e.g., “on the other hand”) but also institutionalized expressions (such as “How are you?”) that Nattinger and DeCarrico (1992) have classified.

English textbooks embody the contents of the Course of Study, which explains the importance of learning and teaching formulaic sequences. Since textbooks are one of the most critical materials for language learning in an instructed context like classrooms (Meunier, 2012), it is thus necessary to analyze the formulaic sequences used in textbooks for better English education. To date, some recent studies investigated formulaic sequences, including lexical bundles, in English textbooks for schools (e.g., Alzahrani, 2020; Coxhead et al., 2020; Lynn, 2021) or for academic purposes (see Cortes (2023) for a review). However, only a limited number of studies have examined formulaic sequences in English textbooks in Japan, and few studies have focused on lexical bundles. The purpose of this study is to fill this gap by quantitatively analyzing lexical bundles used in English textbooks in Japan and by comparing them with those in spoken English. This is because, given that the Course of Study repeatedly states that students are expected to develop the ability to communicate using English in real-life situations, it is useful to compare textbooks and spoken English. Therefore, this study can potentially refine the understanding of how formulaic sequences, specifically lexical bundles, are used in English textbooks compared to spoken English. This study may also contribute to future textbook analysis on formulaic sequences and to developing better English textbooks in terms of formulaic sequences.

2. Previous Studies on English Textbooks in Japan

2.1 Formulaic sequences

Although there are some previous studies on the vocabulary used in English textbooks in Japan, the number of previous studies on formulaic sequences used in English textbooks is limited, and most of them focused on collocations, another type of

formulaic sequence broadly defined as “pairs of words that are commonly found together” (Szudarski, 2018, p. 76).

Koya (2004a, 2004b) targeted English textbooks in Japan and investigated how verb + noun collocations (e.g., “pay attention”) were used. Koya (2004a) investigated six English textbooks used in Japanese high schools. She compiled a small textbook corpus and analyzed verb + noun collocations appearing in the corpus. Those collocations were compared with those appearing in the Bank of English, a native English corpus. The study concluded that although Japanese English textbooks frequently used collocations appearing in the Bank of English, those collocations appeared only once or twice, which is insufficient for learners to remember the items. Koya (2004b) is also a study on verb + noun collocations used in English textbooks in Japan, and the author compiled another textbook corpus because of the revision of the Course of Study and English textbooks. Koya concluded that collocations did not reflect native English corpora like the Bank of English and the textbooks were not improved by revising the Course of Study in terms of collocations. Moreover, those collocations were not repeated enough for students to learn them.

Takesue (2019) also analyzed collocations in the New Horizon series, the most widely used English textbook in Japanese junior high schools. Takesue created a small corpus of the New Horizon series (about 10,000 words) and analyzed the use of verb collocations (i.e., verb + preposition, verb + noun, verb + adverb, verb + adjective, noun + verb, verb + preposition + noun), such as the number of collocations used in the corpus, the types of collocations, and the frequency of each collocation that appeared in the New Horizon corpus. He reported that about 90% of the verb collocations were used only once or twice. The results are consistent with Koya’s (2004a) analysis of English textbooks. Given that the more vocabulary learners encounter repeatedly, the more it sticks in their memory (e.g., Nation, 2022), these results are surprising and probably indicate that English textbooks tend not to provide sufficient repetition.

2.2 Lexical Bundles

There are limited numbers of studies on formulaic sequences in English textbooks; among those studies, very few have analyzed lexical bundles. Ishikawa (2019) attempted to build a helpful lexical bundle list for English education in Japan. In the study, he set a written English corpus of native English speakers as the target corpus, a corpus of English textbooks for Japanese junior high and high schools and university entrance exams as the input corpus, and the written English corpus of Japanese university students as an output corpus. Ishikawa then compared the use of lexical

bundles in those three corpora. The results showed that the observed lexical bundles used in the output corpus greatly deviated from those of native speakers, although there was some bias due to the corpus design. A comparison of the key lexical bundles of each corpus also revealed that, followed by the use of lexical bundles in English textbooks, the output corpus differed from that of native speakers the most.

Northbrook and Conklin (2018) appear to be the only study that has analyzed English textbooks in Japan thoroughly in terms of lexical bundles. The authors compiled a corpus of English textbooks used in Japanese junior high schools. Lexical bundles were then extracted from the corpus, and their authenticity was analyzed compared to a spoken English corpus. The results showed that short lexical bundles were commonly used among the two corpora, but long lexical bundles deviated entirely from the native English patterns. Moreover, they reported that some lexical bundles were commonly used in both corpora but the contexts in which they were used were quite different. The authors criticized that the developers of English textbooks wrote the scripts primarily for grammar explanations and they probably not referring to the authentic use of formulaic sequences in English.

There seem to be two limitations in the previous studies. The first point is that the Course of Study was revised a few years ago. In 2017, the revised Course of Study for elementary and junior high school was released, and textbooks following the revised Course of Study were published in 2020 for elementary schools and 2021 for junior high schools. Additionally, with the revision of the Course of Study for high schools in 2018, new textbooks corresponding to the new Course of Study will be published sequentially from 2022 to 2024. Thus, it is believed that no studies have examined lexical bundles in those new textbooks yet. The second point is the corpus size and contents. Northbrook and Conklin (2018) used a corpus of English textbooks for Japanese junior high schools, and the size was relatively small, approximately 153,000 words. In contrast, as will be discussed later, the English textbook corpus used in the present study is more than three times as large as this one, and it is possible to increase the reliability of the quantitative analysis. Ishikawa (2019) compiled a nearly one-million-word corpus, but English textbooks for elementary schools were not included. The present study aims to overcome these limitations while continuing the objectives of these previous studies.

3. Methodology

3.1 The Purpose of the Study, Approach, and Research Questions

The present study aims to investigate the differences between the use of lexical bundles in the latest English textbooks in Japan and in the speech of English speakers

in the US. To achieve this, the study will compare the use of lexical bundles in an English textbook corpus as the target corpus and a corpus of English speakers in the US as the reference corpus, as done in previous studies (e.g., Ädel & Erman, 2012; Chen & Baker, 2010). The study will address the following four research questions to provide clarification:

1. Are there any differences between the English textbook corpus and spoken English corpus in terms of (a) the ratio of types to tokens (type-token ratio) of lexical bundles, and (b) the ratio of bundle tokens to the corpus tokens?
2. How many bundle types are commonly used in at least three out of six series of textbooks in the English textbook corpus? What are the bundle tokens of those bundle types?
3. How many bundle types found in RQ2 are also present in the spoken English corpus?
4. Is there a difference in the ratio of bundle tokens between the bundle types extracted in RQ3 and the lexical bundles used in the entire corpus for the English textbook corpus and the spoken English corpus?

The goal of RQ1 is (a) to compare the type-token ratio, which is a measure of lexical diversity and indicates less repetition of the same words when closer to 1 (Szudarski, 2018), and (b) to compare the ratio of bundle tokens to the corpus tokens, which shows how frequently lexical bundles are repeated. For RQ2 to RQ4, the analysis focused on lexical bundles that appeared in at least three series of textbooks. Although analyzing bundles that appear in all textbooks would be ideal for representing their usage without being influenced by the topics (see 3.5.2 for more details), this was not possible for long lexical bundles. Specifically, no 6-word bundles were found in any of the six series of textbooks. Lowering the threshold to four series resulted in only one 6-word bundle being extracted ("what are you going to do"). Therefore, the decision was made to lower the threshold to three in order to increase the number of lexical bundles analyzed. This means that the analysis focused on lexical bundles commonly used in at least three series of textbooks. Note that the term "bundle types" refers to distinct sets of lexical bundles, while "bundle tokens" represents the number of occurrences of these bundles.

3.2 Corpora Employed in the Study

The following two corpora were used in this study: a corpus of English textbooks published in Japan (hereafter, the TX corpus) and a corpus of spoken English by English

speakers in the US (SUBTLEXus; Brysbaert & New, 2009).

The target corpus is the TX corpus. It includes elementary, and junior high school English textbooks published in Japan used under the current Course of Study (The Ministry of Education, Culture, Sports, Science and Technology, 2017a, 2017b). Information on the construction of the TX corpus is summarized in Table 1.

Table 1

English Textbooks Included in the TX Corpus

	Junior high school textbooks		Elementary school textbooks		
Series	Number of volumes	Tokens	Number of volumes	Tokens	Tokens per series
A	3 (1st, 2nd, 3rd grade)	19,601	2 (5th, 6th grade)	44,001	63,602
B	3 (1st, 2nd, 3rd grade)	67,084	2 (5th, 6th grade)	51,077	118,161
C	3 (1st, 2nd, 3rd grade)	65,553	2 (5th, 6th grade)	19,676	85,229
D	3 (1st, 2nd, 3rd grade)	56,170	2 (5th, 6th grade)	33,697	89,867
E	3 (1st, 2nd, 3rd grade)	102,641	2 (5th, 6th grade)	22,105	124,746
F	3 (1st, 2nd, 3rd grade)	33,415			33,415

There are English textbooks available for fifth and sixth-grade students in elementary school, and for first, second, and third-grade students in junior high school. The total number of textbooks in the TX corpus is 28, with 10 textbooks for elementary schools and 18 textbooks for junior high schools. The total tokens (corpus token) of the TX corpus are 515,020. It is important to note that the TX corpus was compiled in the laboratory to which the author belongs, and it is only for in-house use and not publicly available.

The reference corpus is a spoken English corpus featuring English speakers in the US named SUBTLEXus (Brysbaert & New, 2009), compiled by researchers at the Department of Experimental Psychology of Ghent University. They claim that the corpus was created to respond to criticism of English corpora that have been used (i.e., Kučera and Francis (1967) for American English, the CELEX (Baayen et al., 1993) for British English). SUBTLEXus comprises 8,388 transcripts of subtitles of American TV shows and movies. The total size of the SUBTLEXus is approximately 51 million words. The author chose the SUBTLEXus because of the following three reasons: (a) it is a publicly available spoken corpus, (b) it is a corpus of American English, which Japanese English textbooks are based on, and (c) it is considered valuable to compare the textbooks with spoken English used in real communication situations to be able to communicate in English in everyday life, which is described in the Course of Study.

3.3 Creating the Sampled SUBTLEXus

The present study employed two corpora: the TX corpus and the SUBTLEXus. However, there was one disadvantage to comparing those corpora. The SUBTLEXus was about 100 times larger in size than the TX corpus. In corpus studies, researchers often use relative frequency, especially frequencies per million words (e.g., McEnery et al., 2006), when comparing corpora of different sizes. However, it has been pointed out that using relative frequencies that exceed the size of a smaller corpus may over-interpret the data (e.g., Ishikawa, 2021), especially misrepresenting infrequent words (Brezina, 2018). Therefore, the SUBTLEXus was downsized to make it easier to compare.

To downsize the SUBTLEXus, the author wrote a script using Python. The procedure is as follows. First, the "random" module, which can be used to perform random operations, such as generating random numbers, was imported in Python. Second, the SUBTLEXus, downloaded as one text file, was loaded into Python. Note that the SUBTLEXus was written line by line in a single text file, and each line was shuffled randomly for copyright reasons. Third, several attempts were made using the "random" module, varying the number of lines sampled to approximate the size of the TX corpus as closely as possible. Consequently, 516,680 tokens were extracted, sampling 49,520 lines. Finally, the sampled SUBTLEXus was written to a text file as a new corpus (henceforth, the SUBTLEXus means the version of SUBTLEXus sampled through these processes). The Python code for downsizing the original SUBTLEXus is provided in the Appendix.

3.4 Tools for Analyzing Corpus Data and Performing Statistical Analysis

Sketch Engine (Kilgarriff et al., 2004, 2014) was employed in the present study as a corpus tool for extracting lexical bundles. It is a versatile web-based corpus management tool built by Adam Kilgarriff, a linguist and lexicographer, for analyzing linguistic information. When corpus files are uploaded to Sketch Engine, they are automatically part-of-speech (POS) tagged and lemmatized. The two corpora that the present study employed were uploaded to Sketch Engine.

Statistical analysis, on the other hand, was performed on the numerical data obtained from Sketch Engine. All statistical analyses were carried out using R (R Core Team, 2022), a free statistical software environment.

3.5 Analysis Procedure

3.5.1 Lexical Bundle Extraction from the Whole Corpus

Lexical bundles were first extracted from the two corpora, the TX corpus, and

the SUBTLEXus. The "N-grams" function of the Sketch Engine was utilized to extract the lexical bundles used throughout each corpus. The function can extract 2- to 6-word lexical bundles. However, since the number of 2-word bundles is extremely large, this study extracted four types of lexical bundles, ranging from 3- to 6-words. These four types of lexical bundles were extracted using the following set of criteria.

Based on the extraction methods used in previous studies that investigated lexical bundles in textbooks (Northbrook & Conklin, 2018), the minimum frequency of occurrence was set at four times. However, as Biber et al. (1999, p. 990) mentioned, long lexical bundles (i.e., 5- and 6-word bundles) are less common than shorter bundles. Therefore, it was decided to change the minimum frequency of 5- and 6-word bundles to half (i.e., two occurrences), following Biber et al. (1999, p. 1001). Furthermore, in order to ensure that all lexical bundles with exactly the same constituent words could be counted as the same item, the "N-grams" function was set to case-insensitive (e.g., "do you want" was treated the same as "Do you want"). Although a short lexical bundle could be part of other long lexical bundles, this study was set up to extract them as independent lexical bundles. For instance, 3-word bundles, such as "I don't know" are part of the 4-word bundle "I don't know why" or the 5-word bundle "I don't know what to." Also, this study does not treat contractions as a single word. How to treat contraction differs in corpus linguistics, and no unified view has been established (see Gardner (2007) for further discussion). For example, "you've" is a contraction of "you have," and Sketch Engine treats this as two separate words (i.e., "you" and "'ve").

3.5.2 Extraction of Lexical Bundles Used in at Least Three Series of English Textbooks

Like any other type of vocabulary, lexical bundles extracted from a corpus are influenced by the topic of the text in which they are included. Some research methods distinguish between lexical bundles that are topic-related and those that are not topic-related (e.g., Dahunsi & Ewata, 2022; Staples et al., 2013; Yan, 2019). Therefore, for extracting lexical bundles from the TX corpus, the method used in this study was to extract lexical bundles common to at least three series of textbooks in order to minimize the influence of the topic (see also 3.1). To accomplish this aim, sub-corpora were created in each textbook series. Of the 28 textbooks in the TX corpus, all but series F also published elementary school English textbooks under the same name (see Table 1). Therefore, sub-corpora A to F were created for each of these textbook series.

In order to find commonly used lexical bundles among at least three series of textbooks, some scripts were written using Python for each length of lexical bundles. Lexical bundles appearing in at least three sub-corpora (i.e., at least three series) out of six sub-corpora of textbooks were extracted. First, using the "N-grams" function of

Sketch Engine, 3- to 6-word bundles were extracted and downloaded from six sub-corpora (i.e., 24 files were downloaded in total). Second, the data extracted were then aggregated by word length. Third, the number of textbook series in which each lexical bundle appeared was tabulated using a function of "Pandas," a data analysis library of Python, which can tally the frequency of values in the data. Finally, only lexical bundles whose count was three or more (i.e., used in at least three textbook series) were extracted for analysis. Note that the last process was done by visually checking the tally results of the second process and only those with a count of three or more were chosen. Those scripts are provided in the Appendix.

4. Results and Discussions

4.1 RQ1: Are there any differences between the English textbook corpus and spoken English corpus in terms of (a) the ratio of types to tokens (type-token ratio) of lexical bundles, and (b) the ratio of bundle tokens to the corpus tokens?

To address (a), Tables 2–5 were constructed to provide a summary of the number of bundle types and the bundle tokens extracted from the TX corpus and the SUBTLEXus for 3- to 6-word lexical bundles.

Table 2

3-word Bundle Types and Bundle Tokens

	TX corpus	SUBTLEXus
Bundle types	8,504	5,000
Bundle tokens	81,027	48,360

Table 3

4-word Bundle Types and Bundle Tokens

	TX corpus	SUBTLEXus
Bundle types	4,743	1,151
Bundle tokens	34,943	9,406

Table 4

5-word Bundle Types and Bundle Tokens

	TX corpus	SUBTLEXus
Bundle types	4,591	564
Bundle tokens	22,327	2,722

Table 5

6-word Bundle Types and Bundle Tokens

	TX corpus	SUBTLEXus
Bundle types	2,781	134
Bundle tokens	12,054	520

A chi-square test of independence was performed on the lexical bundles of each word length, and the values obtained from both corpora were compared. Consequently, the following results were obtained: 3-word bundles, χ^2 (1) = 0.626, p = .429, φ = .002; 4-word bundles, χ^2 (1) = 8.754, p = .003, φ = .013; 5-word bundles, χ^2 (1) = 0.017, p = .896, φ

= .001; and 6-word bundles, χ^2 (1) = 1.134, p = .287, φ = .009. To answer (b), Tables 6–9 were created to summarize the bundle tokens presented in Tables 2–5 and the corpus tokens (i.e., 515,020 for the TX corpus, and 516,680 for the SUBTLEXus).

Table 6

3-word Bundle Tokens and Corpus Tokens

	TX corpus	SUBTLEXus
Bundle tokens	81,027	48,360
Corpus tokens	515,020	516,680

Table 7

4-word Bundle Tokens and Corpus Tokens

	TX corpus	SUBTLEXus
Bundle tokens	34,943	9,406
Corpus tokens	515,020	516,680

Table 8

5-word Bundle Tokens and Corpus Tokens

	TX corpus	SUBTLEXus
Bundle tokens	22,327	2,722
Corpus tokens	515,020	516,680

Table 9

6-word Bundle Tokens and Corpus Tokens

	TX corpus	SUBTLEXus
Bundle tokens	12,054	520
Corpus tokens	515,020	516,680

A chi-square test of independence was performed and the following results were obtained: 3-word bundles, χ^2 (1) = 7427, p < .001, φ = .08; 4-word bundles, χ^2 (1) = 14183, p < .001, φ = .115; 5-word bundles, χ^2 (1) = 15044, p < .001, φ = .119; and 6-word bundles, χ^2 (1) = 10488, p < .001, φ = .1.

The results indicate that, except for the 4-word bundles, there was no significant difference in the type-token ratio. According to Mizumoto and Takeuchi (2008), the effect size for the significant difference in the 4-word bundles is very small. Thus, it can be considered that there is no significant difference in the diversity of the lexical bundles used between the TX corpus and the SUBTLEXus. However, when comparing the ratio of bundle tokens of each bundle length to corpus tokens, it was found that English textbooks used significantly more lexical bundles than spoken English in all length categories, although the effect sizes were small.

The findings suggest that English textbooks, which are crucial resources for language learners (Meunier, 2012), are successful in providing a large number of lexical bundles. It was found in a previous study that English learners, even young ones, respond more quickly and accurately to the lexical bundles presented in English textbooks (Northbrook & Conklin, 2019). This is consistent with previous research showing that formulaic sequences can enhance both accuracy and fluency in language

use (e.g., Dechert, 1983). Therefore, textbooks can be considered a useful tool for assisting language learning for students. The fact that textbooks contain more lexical bundles than spoken English suggests that they can facilitate the effective use of English and foster communication skills among elementary and junior high school students.

4.2 RQ2: How many bundle types are commonly used in at least three out of six series of textbooks in the English textbook corpus? What are the bundle tokens of those bundle types?

The purpose of RQ2 is to determine the number of bundle types and bundle frequency that were used in at least three series of textbooks among lexical bundles extracted in RQ1 (as shown in Tables 2–5) in order to exclude bundles related to specific topics. The same frequency of occurrence as in RQ1 was used to extract lexical bundles, but for RQ2, these bundles must appear in at least three out of the six textbook series sub-corpora. The results of this analysis are presented in Table 10.

Table 10

Type and Token of Lexical Bundles Used in at Least Three Series of Textbooks

	TX corpus			
	Bundle types	Ratio of bundle types to total types of the TX corpus (%)	Bundle tokens	Ratio of bundle tokens to total tokens of the TX corpus (%)
3-word bundle	181	2.1	16,651	20.5
4-word bundle	36	0.7	3,772	10.8
5-word bundle	22	0.5	1,155	5.2
6-word bundle	6	0.2	89	0.7

The results indicate that 181 types (2.1%) for 3-word bundles, 36 types (0.7%) for 4-word bundles, 22 types (0.5%) for 5-word bundles, and 6 types (0.2%) for 6-word bundles were commonly used. In terms of the number of bundle tokens, the 3-word bundles accounted for approximately 20.5%, the 4-word bundles for 10.8%, the 5-word bundles for 5.2%, and the 6-word bundles for 0.7%.

The results indicate that there is a limited number of bundle types used in at least three textbook series, particularly for longer bundles, comprising only a few percent. This aligns with the pattern identified by Biber et al. (1999) of decreasing commonality as bundle length increases. However, in terms of bundle tokens here (16,651 + 3,772 + 1,155 + 89 = 21,677 tokens), a small proportion of bundles accounts for roughly 14.4% of total bundle token extracted from the entire corpus (81,027 + 34,943 + 22,327 + 12,054

= 150,351 tokens). This is consistent with the findings of the research on vocabulary, which shows that a significant portion of language comprises a smaller number of high-frequency words, and learners benefit from acquiring these words efficiently through the use of a frequency list. Therefore, it is crucial for learners to learn those words in English (e.g., Nation, 2022).

On the contrary, the results suggest that there is a potential lack of consensus regarding the selection of lexical bundles in English textbooks. Previous studies have also pointed out the absence of a unified approach in terms of the formulaic sequences presented in English textbooks. For example, Koya (2004a) investigated collocations used in Japanese high school textbooks and concluded that there was no general consensus among English textbooks regarding which collocations should be taught. As a result, the absence of a clear concept for the collocations which English textbooks used may extend to lexical bundles.

4.3 RQ3: How many bundle types found in RQ2 are also present in the spoken English corpus?

RQ3 focuses on the bundle types that appeared in at least three series of textbooks found in RQ2. Table 11 shows the results.

Table 11

Bundle Types Extracted in RQ2 That are Also Used in the SUBTLEXus

	Bundle types found in RQ2	Bundle types common to the SUBTLEXus	Ratio (%)
3-word bundle	181	140	77.3
4-word bundle	36	30	83.3
5-word bundle	22	9	40.9
6-word bundle	6	2	33.3

Among lexical bundles extracted in RQ2, 140 out of 181 types (77.3%) for the 3-word bundle, 30 out of 36 types (83.3%) for the 4-word bundle, 9 out of 22 types (40.9%) for the 5-word bundle, and 2 out of 6 types (33.3%) for the 6-word bundle were also found in the SUBTLEXus.

The results indicate that there is a tendency that relatively short lexical bundles

(3- and 4-word bundles) are used in the SUBTLEXus, and it is suggested that the lexical bundles used in at least three series of English textbooks is also frequently used in spoken English. However, for longer bundles (5- and 6-word bundles), there is less agreement, and the proportion of commonly used bundles between two corpora decreases more rapidly than for shorter bundles. This trend is consistent with the findings of Biber et al. (1999), who reported that longer bundles tend to have lower varieties in their high-frequency items as compared to shorter ones.

4.4 RQ4: Is there a difference in the ratio of bundle tokens between the bundle types extracted in RQ3 and the lexical bundles used in the entire corpus for the English textbook corpus and the spoken English corpus?

RQ4 focuses on bundle tokens and aims to investigate whether there are any differences between the two corpora by comparing the bundle tokens of the bundle types extracted in RQ3 (see Table 11) with the bundle tokens of the entire corpus (see Tables 2–5). For instance, in the case of 3-word bundles (see Table 12), the total number of tokens that appear in the TX corpus is 81,027. However, if only considering the 140 bundle types obtained in RQ3, the number is 14,741 tokens. Similarly, there are a total of 48,360 tokens for 3-word bundles that appear in SUBTLEXus but considering bundle types identified in RQ3 (140 types) results in 7,488 tokens.

Table 12

3-word Bundle Tokens of 140 Types Found in RQ3 and Bundle Tokens of Entire Corpus

	TX corpus	SUBTLEXus
3-word bundle tokens of 140 types	14,741	7,488
3-word bundle tokens of entire corpus	81,027	48,360

Table 13

4-word Bundle Tokens of 30 Types Found in RQ3 and Bundle Tokens of Entire Corpus

	TX corpus	SUBTLEXus
4-word bundle tokens of 30 types	3,057	1,094
4-word bundle tokens of entire corpus	34,943	9,406

Table 14

5-word Bundle Tokens of 9 Types Found in RQ3 and Bundle Tokens of Entire Corpus

	TX corpus	SUBTLEXus
5-word bundle tokens of 9 types	631	83
5-word bundle tokens of entire corpus	22,327	2,722

Table 15

6-word Bundle Tokens of 2 Types Found in RQ3 and Bundle Tokens of Entire Corpus

	TX corpus	SUBTLEXus
6-word bundle tokens of 2 types	48	10
6-word bundle tokens of entire corpus	12,054	520

The following statistical results were obtained using the chi-square test of independence for 3- to 5-word bundles and Fisher's exact test for 6-word bundles: 3-word bundles, χ^2 (1) = 110.89, $p < .001$, $\varphi = .027$; 4-word bundles, χ^2 (1) = 58.956, $p < .001$, $\varphi = .035$; 5-word bundles, χ^2 (1) = .337, $p = .562$, $\varphi = .004$; and 6-word bundles, $OR = .207$, $p < .001$. These results show significant differences in the number of tokens of the lexical bundles commonly used in the two corpora, except for the 5-word bundle, although the effect size is very small. Thus, the results indicate that the TX corpus uses significantly more short bundles but that the longer lexical bundles may not be significantly different. For short bundles, the results are consistent with Northbrook and Conklin's (2018) discussion that English textbooks were closer to the use of native English speakers. For long bundles, on the other hand, the results may indicate that English speakers have a minimum repertoire of frequently used long bundles. Northbrook and Conklin's (2018) report is consistent with this result since English speakers use fewer long bundles.

Those results would be consistent with the discussions of previous studies that, as discussed in RQ1, low-proficiency English learners, those whom the English textbooks are intended for, are more likely to use formulaic sequences. However, again, since formulaic sequences have been shown to increase fluency and accuracy (e.g., Boers et al., 2006; Dechert, 1983), repeated use of such formulaic sequences in textbooks should not be criticized as a deviation from the use of English speakers. Instead, in order to develop the ability to communicate in English that the Course of Study aims at, it should be

recognized positively.

5. Pedagogical Implications, Limitations, and Future Directions

5.1 Pedagogical Implications

In the present study, two pedagogical implications can be drawn. Firstly, developers of English textbooks should consult native English data by utilizing corpora. The study found that as the number of words in lexical bundles increased, they became inconsistent with their use in English speech, which is consistent with the findings of Northbrook and Conklin (2018). Although it was an investigation of collocations, Koya (2004b) also concluded that English textbooks did not include the collocations frequently used in a native speakers' corpus. As Nelson (2023) notes, corpora are typically used to evaluate textbooks, including the present study, rather than to develop new ones. This highlights the challenge of creating effective language materials. Therefore, utilizing corpora to develop English textbooks would likely improve Japanese students' ability to communicate effectively in English.

Secondly, the study suggests the need for a lexical bundle list to be taught through English textbooks. As shown in RQ2, only a small percentage of the lexical bundles extracted from the entire TX corpus were used in at least three out of six sub-corpora. Hence, there appears to be no consensus on the lexical bundles used in English textbooks. As noted earlier, Koya (2004a) found that there is no consensus on the collocations used in English textbooks, and this lack of agreement may also extend to lexical bundles. Therefore, a valuable list of lexical bundles for learners needs to be constructed. Some scholars have developed useful phrase lists using corpora (e.g., the Academic Formulas List; Simpson-Vlach & Ellis, 2010). Ishikawa (2019) has worked on a list specially created for Japanese learners of English, and such a list should be further developed. This type of list is useful for textbook development and may contribute more effectively to achieving the goals outlined in the Course of Study.

5.2 Limitations and Future Directions

This study aimed to quantitatively compare lexical bundles used in English textbooks in Japan and with those used in spoken English. However, this study has several limitations. The first limitation is that it was impossible to include high school textbooks in the TX corpus for analysis. As mentioned earlier, with the revision of the Course of Study, new textbooks corresponding to the new Course of Study will be published sequentially from 2022 to 2024. At this point, only the first-year high school textbooks had been published, and it was impossible to compile the corpus completely.

High school English textbooks will use a higher level of vocabulary and may produce results different from those of the present study. Therefore, future research will also require compiling and analyzing a high school English textbook corpus.

The second limitation concerns how the corpus should be designed. Due to this point, the corpora that were used in the present study did not allow for a sufficient qualitative analysis, which is one of the essential aspects of corpus research (e.g., McEnery & Hardie, 2012). For instance, since the TX corpus was created without tagging registers such as dialogue, monologue, and written text, which has been the critical subject of studies on lexical bundles (e.g., Biber, 2009; Biber et al., 2004), it was impossible to specify the registers that lexical bundles were observed in. Moreover, the SUBTLEXus did not provide information on the English level of the texts, and it was not possible to align the English levels in the TX corpus and the SUBTLEXus. This is a critical issue because the English level of Japanese students targeted by English textbooks differs greatly from that of English speakers in the US. For example, Negishi et al. (2012) reported that as many as 80% of Japanese students' English proficiency level remains at A-level, the lowest proficiency level of the Common European Framework of Reference (CEFR; Council of Europe, 2001, 2020), which is currently the most influential language education framework. Therefore, when comparing English usage between a target corpus about English in Japan with a reference corpus of native speakers, future studies need to adjust the levels according to a unified criterion, such as the CEFR, and compare within the same English level.

How to design the corpora and under what conditions comparisons should be made will cause problems (e.g., Yan, 2019). In particular, recent studies have raised various issues about the design of corpora used when studying lexical bundles. Pan et al. (2020), for instance, claimed that the following three things should be considered when using corpora: (a) the total number of words, (b) the number of texts, and (c) the length of those texts. Pan et al. (2019) compared the use of lexical bundles between L1 English writers and L2 English writers by reworking the corpus under different conditions, such as the number of words and texts. The results showed that even when the same corpus was used, the number and structure of the extracted lexical bundles differed greatly depending on the conditions. Thus, corpus design is crucial when studying lexical bundles with multiple corpora. This suggests that if future research focuses on more qualitative aspects, caution is needed to choose appropriately designed corpora according to the research questions.

6. Conclusion

This study aimed to compare the lexical bundles used in a corpus of English textbooks used in Japanese elementary and junior high schools (the TX corpus) and a corpus of spoken English by English L1 speakers (the SUBTLEXus) and to analyze the differences between them quantitatively. The study addressed the following four research questions: (1) Are there any differences between the English textbook corpus and spoken English corpus in terms of (a) the ratio of types to tokens (type-token ratio) of lexical bundles, and (b) the ratio of bundle tokens to the corpus tokens? (2) How many bundle types are commonly used in at least three out of six series of textbooks in the English textbook corpus? What are the bundle tokens of those bundle types? (3) How many bundle types found in RQ2 are also present in the spoken English corpus? (4) Is there a difference in the ratio of bundle tokens between the bundle types extracted in RQ3 and the lexical bundles used in the entire corpus for the English textbook corpus and the spoken English corpus?

This study is believed to be the first lexical bundle analysis of English textbooks in line with the latest Course of Study. The results show that English textbooks use significantly more lexical bundles than spoken English, but very few are commonly used in at least three series of textbooks. It was also found that shorter lexical bundles are used in spoken English, but as their length increases, there is a deviation from spoken English. These results suggest that while English textbooks may support students' learning of lexical bundles, there is room for improvement given the lack of concept or consensus in choosing the lexical bundles used and the deviation from spoken English. Despite some limitations, this study will enrich the understanding of the formulaic sequences used in English textbooks and, through future research, may contribute to better textbook development and better English education in terms of vocabulary.

References

Ädel, A., & Erman, B. (2012). Recurrent word combinations in academic writing by native and non-native speakers of English: A lexical bundles approach. *English for Specific Purpose, 31*(2), 81–92. https://doi.org/10.1016/j.esp.2011.08.004

Alzahrani, A. (2020). The structure and function of lexical bundles in communicative Saudi high school EFL textbooks. *International Journal of Applied Linguistics and English Literature, 9*(5), 1–10. https://doi.org/10.7575/aiac.ijalel.v.9n.5p.1

Baayen, R. H., Piepenbrock, R., & van Rijn, H. (1993). *The CELEX Lexical Database* [CD-ROM]. Linguistic Data Consortium.

Biber, D. (2009). A corpus-driven approach to formulaic language in English: Multi-

word patterns in speech and writing. *International Journal of Corpus Linguistics, 14*(3), 275–311. https://doi.org/10.1075/ijcl.14.3.08bib

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

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

Boers, F., Eyckmans, J., Kappel, J., Stengers, H., & Demecheleer, M. (2006). Formulaic sequences and perceived oral proficiency: Putting a Lexical Approach to the test. *Language Teaching Research, 10*(3), 245–261. https://doi.org/10.1191/1362168806lr195oa

Bolinger, D. (1979). Meaning and memory. In G. Haydu (Ed.), *Experience forms: Their cultural and individual place and function* (pp. 95–112). De Gruyter Mouton.

Brezina, V. (2018). *Statistics in corpus linguistics: A practical guide.* Cambridge University Press. https://doi.org/10.1017/9781316410899

Brysbaert, M., & New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods, 41*(4), 977–990. https://doi.org/10.3758/BRM.41.4.977

Chen, Y., & Baker, P. (2010). Lexical bundles in L1 and L2 academic writing. *Language Learning & Technology, 14*(2), 30–49.

Cortes, V. (2023). Lexical bundles in EAP. In R. Ratkaine Jablonkai & E. Csomay (Eds.), *The Routledge handbook of corpora and English language teaching and learning* (pp. 220–233). Routledge.

Council of Europe (2001). *Common European framework of reference for languages: Learning, teaching, assessment*. Cambridge University Press. https://rm.coe.int/1680459f97

Council of Europe (2020). *Common European framework of reference for languages: Learning, teaching, assessment: Companion volume*. https://rm.coe.int/common-european-framework-of-reference-for-languages-learning-teaching/16809ea0d4

Coxhead, A., Rahmat, Y., & Yang, L. (2020). Academic single and multiword vocabulary in EFL textbooks: Case studies from Indonesia and China. *The TESOLANZ Journal, 28*, 75–88.

Csomay, E. (2012). Lexical bundles in discourse structure: A corpus-based study of classroom discourse. *Applied Linguistics, 34*(3), 369–388. https://doi.org/10.1093/applin/ams045

Dahunsi, T. N., & Ewata, T. O. (2022). An exploration of the structural and colligational characteristics of lexical bundles in L1–L2 corpora for English language teaching. *Language Teaching Research*, Advance online publication. https://doi.org/10.1177/13621688211066572

Dechert, H. (1983). How a story is done in a second language. In C. Faerch & G. Kasper (Eds.), *Strategies in interlanguage communication* (pp. 175–95). Longman.

Erman, B., & Warren, B. (2000). The idiom principle and the open choice principle. *Text & Talk, 20*(1), 29–62. https://doi.org/10.1515/text.1.2000.20.1.29

Gardner, D. (2007). Validating the construct of word in applied corpus-based vocabulary research: A critical survey. *Applied Linguistics, 28*(2), 241–265. https://doi.org/10.1093/applin/amm010

Ishikawa, S. (2019). Eigokyouiku ni okeru rengo: Taagetto / innputto / autoputto no sangen koopasu bunseki wo fumaeta English N-gram List for Japanese Leaners of English (ENL-J) no kaihatsu to riyou [Development and utilization of "English n-gram list for Japanese learners of English (ENL-J)": Importance of multiword units for English language teaching]. In N. Yasunori, Y. Yoshimura, & Y. Yoshikawa (Eds.), *Gengo bunseki no furonthia* [A frontier in language analysis] (pp. 32–47). Kinseido Publishing.

Ishikawa, S. (2021). *Beeshikku koopasu gengogaku* [A basic guide to corpus linguistics] (2nd ed.). Hitsuji Shobo.

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P., & Suchomel, V. (2014). The Sketch Engine: ten years on. *Lexicography, 1*(1), 7–36. https://doi.org/10.1007/s40607-014-0009-9

Kilgarriff, A., Rychlý, P., Smrž, P., & Tugwell, D. (2004). The Sketch Engine. *Proceedings of the 11th EURALEX International Congress*, 105–115.

Koya, T. (2004a). Collocation research based on corpora collected from English textbooks for Japanese upper secondary schools. In Y. Watanabe, I. Nagano, & A. Morita (Eds.), *Collection of papers in honor of Professor Yoshiaki Shinoda* (pp. 99–113). Nanundo.

Koya, T. (2004b). A comparison of verb-noun collocations collected from revised high school English textbooks in Japan. *The Bulletin of the Graduate School of Education of Waseda University, 11*(2), 55–70.

Kučera, H., & Francis, W. N. (1967). *Computational analysis of present-day American English*. Brown University Press.

Lewis, M. (1993). *The Lexical Approach: The state of ELT and a way forward.* Language Learning Publications.

Lynn, E. M. (2021). Comparing lexical bundle use in EAP reading textbooks to lower-division university textbooks. *Pedagogical Linguistics, 2*(1), 30–63. https://doi.org/10.1075/pl.21001.lyn

McEnery, T., & Hardie, A. (2012). *Corpus linguistics: Method, theory and practice.* Cambridge University Press.

McEnery, T., Xiao, R., & Tono, Y. (2006). *Corpus-based language studies: An advanced resource book*. Routledge.

Meunier, F. (2012). Formulaic language and language teaching. *Annual Review of Applied Linguistics, 32*, 111–129. https://doi.org/10.1017/S0267190512000128

Ministry of Education, Culture, Sports, Science and Technology. (2017a). *Syougakkou gakushu shidou yoryo gaikokugo katsudou/gaikokugo hen* [Course of Study of English for elementary school]. https://www.mext.go.jp/component/a_menu/education/micro_detail/__icsFiles/afieldfile/2019/03/18/1387017_011.pdf

Ministry of Education, Culture, Sports, Science and Technology. (2017b). *Chugakkou gakushu shidou yoryo gaikokugo hen* [Course of Study of English for junior high school]. https://www.mext.go.jp/component/a_menu/education/micro_detail/__icsFiles/afieldfile/2019/03/18/1387018_010.pdf

Ministry of Education, Culture, Sports, Science and Technology. (2018). *Koutougakkou gakushu shidou yoryo gaikokugo hen* [Course of Study of English for high school]. https://www.mext.go.jp/content/1407073_09_1_2.pdf

Mizumoto, A., & Takeuchi, O. (2008). Basics and considerations for reporting effect sizes in research papers. *Studies in English Language Teaching, 31*, 57–66. http://www.mizumot.com/files/EffectSize_KELES31.pdf

Nation, I. S. P. (2022). *Learning vocabulary in another language* (3rd ed.). Cambridge University Press.

Nattinger, J. R. (1980). A lexical phrase grammar for ESL. *TESOL Quarterly, 14*(3), 337–344. https://doi.org/10.2307/3586598

Nattinger, J. R., & DeCarrico, J. S. (1992). Lexical phrases and language teaching. Oxford University Press.

Negishi, M., Takada, T., & Tono, Y. (2012). A progress report on the development of the CEFR-J. *Studies in Language Testing, 36*, 137–157.

Nelson, M. (2023). Corpora for English language learning textbook evaluation. In R. Ratkaine Jablonkai & E. Csomay (Eds.), *The Routledge handbook of corpora and English language teaching and learning* (pp. 147–160). Routledge.

Northbrook, J., & Conklin, K. (2018). What are you talking about? *International Journal of Corpus Linguistics, 23*(3), 311–334. https://doi.org/10.1075/ijcl.16024.nor

Northbrook, J., & Conklin, K. (2019). Is what you put in what you get out? Textbook-derived lexical bundle processing in beginner English learners. *Applied Linguistics, 40*(5), 816–833. https://doi.org/10.1093/applin/amy027

Pan, F., Reppen, R., & Biber, D. (2020). Methodological issues in contrastive lexical bundle research. *International Journal of Corpus Linguistics, 25*(2), 215–229. https://doi.org/10.1075/ijcl.19063.pan

Pawley, A., & Syder, F. H. (1983). Two puzzles for linguistic theory: Nativelike selection and nativelike fluency. In J. C. Richards & R. W. Schmidt (Eds.), *Language and communication* (pp. 191–225). Longman.

R Core Team (2022). *R: A language and environment for statistical computing* (Version 4.2.0) [Computer software]. R Foundation for Statistical Computing. https://www.R-project.org/

Simpson-Vlach, R., & Ellis, N. C. (2010). An academic formulas list: New methods in phraseology research. *Applied Linguistics, 31*(4), 487–512. https://doi.org/10.1093/applin/amp058

Sinclair, J. (1991). *Corpus, concordance, collocation*. Oxford University Press.

Staples, S., Egbert, J., Biber, D., & McClair, A. (2013). Formulaic sequences and EAP writing development: Lexical bundles in the TOEFL iBT writing section. *Journal of English for Academic Purposes, 12*(3), 214–225. https://doi.org/10.1016/j.jeap.2013.05.002

Szudarski, P. (2018). *Corpus linguistics for vocabulary: A guide for research*. Routledge.

Takesue, T. (2019). Collocation research based on junior-high school English textbooks in Japan. *The Bulletin of the Graduate School of Language & Literature of Prefectural University of Kumamoto, 12*, 125–137. http://rp-kumakendai.pu-kumamoto.ac.jp/dspace/handle/123456789/1969

Wray, A. (2002). *Formulaic language and the lexicon*. Cambridge University Press.

Yan, H. (2019). I think we should…: Investigating lexical bundle use in the speech of English learners across proficiency levels. *International Journal of Translation, Interpretation, and Applied Linguistics (IJTIAL), 1*(2), 1–16. http://doi.org/10.4018/IJTIAL.2019070105

Appendix

1. Python script for sampling from original SUBTLEXus

```
# Import a Python module called "random"
import random
# Read the oroginal SUBTLEXus
with open('Subtlex.US.txt') as f:
    lines = f.readlines()
# Sampling the original SUBTLEXus
sample = random.sample(lines, 49520)
SampleText = []
for line in sample:
    line = line.strip()
    SampleText.append(line)
SampleText_str = '\n'.join(SampleText)
# Write the results to an empty file prepared in advance
f = open('Subtlexus_Sampled2.txt', 'w', encoding='UTF-8')
f.write(SampleText_str)
f.close()
```

2. Python script to count how many textbook series lexical bundles are used in

```
# import Pandas, a data analysis library of Python
import pandas as pd
# Read a 3-word bundle file that compiles data gathered from six sub-corpora
# The '3-gram_all.csv' file includes all 3-word bundles from six sub-corpora
all_3grams = pd.read_csv('3-gram_all.csv', encoding='utf-8')
# Count the number of lexical bundles recorded in the column named 'grams' and assign it to a variable 'vc'
vc = all_3grams['grams'].value_counts()
# Convert counted data to data frame type in Pandas
all_3grams_counted = pd.DataFrame(vc)
```

Note. The script above is an example only for 3-word bundles.