**“What are you talking about?”**

An analysis of lexical bundles in Japanese junior high school textbooks

In a communicative approach to language teaching, students are presented with ‘authentic’ language, which is thought to allow them to produce it in a nativelike way. The current study explores whether the lexical bundles in communicative Japanese junior high school textbooks are representative of conversational English. To do this, we use a corpus-based approach that compares the most frequent lexical bundles in the textbooks to those in an English reference corpus. The study finds that although lexical bundles are very frequent in the textbooks, and conform relatively well to English patterns at shorter lengths (3-word lexical bundles), they deviate considerably at longer ones (4-, 5- and 6-words). This has important implications for the communicative utility of the language in the textbooks.

**Keywords:** lexical bundles, junior high school textbooks, corpus analysis, usage-based theories

**1. Introduction**

Many schools and educational agencies advocate taking a communicative approach in foreign language classrooms. In this approach, students encounter and are asked to produce authentic language in the classroom, which is thought to give them the ability to reproduce it naturally in the real world. The cornerstone of the communicative approach is providing students with authentic language materials. This view of foreign language teaching aligns well with usage-based accounts of language acquisition, which maintain that language develops in reaction to use and exposure (e.g. Barlow & Kemmer, 2000; Bybee, 2006; Ellis, O'Donnell, & Römer, 2013; Tyler, 2010). Crucially, this means that the quality of input defines the quality of language learned. In the case of English as a Foreign Language (EFL) students, this input is primarily from classroom instruction and materials (Meunier, 2012; Nunan, 1991). Despite this, studies focusing on secondary school textbooks have found the dialogues in them to be inauthentic (e.g. Römer, 2004; Römer, 2005; Nguyen & Ishitobi, 2012; also see Gilmore, 2004), in that they do not accurately represent the situations students need to be prepared for. Relevant to the current study, Ogura (2008) found that although the Japanese Ministry of Education (MEXT) places an emphasis on communicative competence, and textbook designers for upper-secondary school claimed their books are communicative*,* the books did little to promote communicative competence.

To ensure that textbooks are providing authentic, real-life opportunities for communicative practice, we need to look at whether students are exposed to frequent, high utility language relevant to general everyday communication. To assess the authenticity of the teaching materials in MEXT approved junior high school textbooks, the present study examines the use of lexical bundles (e.g. *you know what*). More specifically, the study looks at whether the highly frequent lexical bundles that are present in the textbooks occur frequently in real language and similarly whether low-frequency ones appear and occur less frequently. To assess this, we have used a modified version of the framework outlined by Stubbs and Barth (2003). The authors created three corpora to represent distinct genres of English – academic language, fiction, and what they label ‘learned’ language (consisting of texts from sources such as natural and social sciences and humanities). They then extracted lexical bundles from these corpora and analysed the resulting lists for differences between the type and frequency of the items found. They found that each of the text types contained many lexical bundles; however, the lexical bundles were qualitatively different. Lexical bundles have also been used to differentiate between television genres (Bednarek, 2012), legal genres (Breeze, 2013) and academic and conversational genres (e.g. Biber, Conrad, & Cortes, 2004; Conrad & Biber, 2004; Gray & Biber, 2013). In the same way, an investigation of lexical bundles should be able to demonstrate how close the language used in communicative textbooks is to the language used in real communicative situations.

**2. Lexical bundles in second language learning**

Speakers tend to express the same things again and again, using the same words each time, resulting in repetitive language use (e.g. Biber, 2009; Erman & Warren, 2000; Nattinger & DeCarrucio, 1992; Pawley & Syder, 1983; Wray, 2002). Lexical bundles are a type of repetitive language described as ‘recurring sequence[s] of three or more words’ (Biber, Johansson, Leech, Conrad, & Finegan, 1999 p. 990) and are usually defined according to a frequency threshold and criterion for dispersion. For example, Biber et al. (1999) set a minimum of ten times per million words for 3- and 4- word lexical bundles, and at least five times per million words for 5- and 6- word lexical bundles (cf. p. 990-991). However, this number is somewhat arbitrary, and studies tailor the selection criteria to fit their individual needs, particularly when working with small corpora (see Biber, 2006 for a detailed discussion of this). Importantly, unlike some other kinds of formulaic language such as idioms or phrasal verbs, lexical bundles are not necessarily linguistic units with idiomatic, grammatical, semantic or pragmatic status (Biber, 2009; Stubbs & Barth, 2003). That is, they are often incomplete fragments (*in the middle of the*), and often span grammatical categories (*yes i do i*).

Lexical bundles, and formulaic language in general, are important for second language learners for two key reasons. First, by knowing chunks of language, learners can express themselves beyond their current level (Myles, Hooper, & Mitchell, 1998; Myles, Mitchell, & Hooper, 1999; Nattinger & DeCarrucio, 1992; Wray, 2002). For example, one does not need to have learned about post-head adverbial modification to use the phrase *thank you very much* in conversation, or to even consider that technically speaking the related phrase *thanks very much* is not even ‘grammatical’ (although it is very much a standard, accepted expression). Second, knowledge of formulaic language helps learners to free up memory resources (Kuiper, 2004; Pawley & Syder, 1983) and sound more natural (Boers, Eyckmans, Kappel, Stengers, & Demecheleer, 2006; Dickinson, 2012; Pawley & Syder, 1983).

Seminal studies such as those from Pawley and Syder (1983), Biber (2009), Erman and Warren (2000), Nattinger and DeCarrucio (1992), among others, have demonstrated the pervasiveness and importance of formulaic language, including lexical bundles, and a great deal of research has been carried out exploring the phenomenon across multiple disciplines. Studies from corpus linguistics have aimed to categorise the function and structure of lexical bundles (e.g. Biber et al., 2004; Biber et al., 1999; Hyland, 2008). In second language acquisition, studies have tried to identify the best lexical bundles to include in teaching (e.g. Simpson-Vlach & Ellis, 2010), and how best to use lexical bundles to improve fluency (e.g. Boers et al., 2006; Dickinson, 2012; Taguchi, 2007). Others have also looked at lexical bundles in learner output (e.g. Cortes, 2004; Yaoyu & Lei, 2011). Further, behavioural studies have investigated the psychological reality of lexical bundles, among other kinds of formulaic sequences (see for example Conklin & Schmitt, 2012; Siyanova-Chanturia & Martinez, 2014 for good overviews).

In terms of studies within corpus linguistics which have looked at lexical bundles in textbooks, the majority are in academic and university contexts (Biber, 2009; Biber et al., 1999; Chen, 2010; Simpson-Vlach & Ellis, 2010; Wood, 2010). Only a few studies have focused specifically on multi-word units in materials for secondary school students, though none of them are *lexial bundle* studies per-se, and fewer still on secondary school materials in the Japanese context. Koya (2004) compiled a list of collocations based on collocation dictionaries, and examined their usage in Japanese senior-school textbooks. Only a small number of the collocations appeared in the textbooks, and over half only appeared once with only eight items appearing more than five times. In a similar study, Tsai (2014) compiled a list of collocations based on the prescribed wordlists for Taiwanese, Ministry approved textbooks. Only a small number of collocates were presented in the textbooks, and similar to Koya (2004) these were not repeated often.

Although only a few studies have looked specifically at lexical bundles in secondary school materials, there have been studies looking more generally at formulaic sequences explicitly taught in textbooks (Koprowski, 2005; Mcaleese, 2013), as well as one looking at teacher talk in Korean classrooms (Kwon & Lee, 2014), and several examining lexical bundles in academic teaching materials. Like the studies above, these studies primarily found item selection to be insufficiently representative of naturally produced language. Koprowski’s (2005) analysis of three English as a Second Language course books found many lexical bundles in the books did not appear in a reference corpus, and many were deemed not useful due to their low-frequency in the reference corpus. Mcaleese (2013) compared formulaic sequences taught in a commercially available course book to the Bank of English corpus. Formulaic sequences were common in the textbook, but a large proportion were very low-frequency. The author concluded that the textbook failed to accurately represent real language usage, and was therefore of limited value.

In academic contexts, Biber et al., 2004 compared lexical bundles in four separate genres of English: conversation, classroom teaching, textbooks and academic prose. They found lexical bundles to be very common in conversation and university teaching, and to a lesser extent, in textbooks and academic prose. Importantly, it appears that textbooks and academic prose match in terms of the number of lexical bundles (see also Biber, 2009; Biber et al., 1999; Simpson-Vlach & Ellis, 2010). This is an encouraging finding and could be taken as an indication that academic teaching materials are authentic. However, the authors did not compare the exact lexical bundles in each genre. Studies by Wood (2010), focusing on EAP textbooks, and Chen (2010), focusing on lexical bundles in textbooks for electrical engineering students, however, showed that formulaic language in textbooks was not representative of language found in authentic reading materials. Wood extracted and analysed lexical bundles from a corpus of six commercially available textbooks finding that most items were in the instructional components of the material but not in the reading texts, meaning that although students did get exposure to lexical bundles, it was not in the texts that were supposed to represent authentic academic reading. Importantly, the lexical bundles they did encounter were not the same as the ones they could expect to encounter outside of the textbooks. Similarly, Chen (2010) found that only a third of the lexical bundles from actual electrical engineering books appeared in preparatory textbooks and the ones that did appear covered only a narrow range of functional categories.

These studies show that, although students are exposed to lexical bundles and formulaic language, there is a mismatch between the lexical bundles in the input and what students might encounter in the real world. Learners are given the potential to acquire lexical bundles and enjoy the benefits which come with them – however, the lexical bundles presented are not necessarily the ones they need. This is problematic because per usage-based accounts of language learning, *students become fluent in the actual language they are exposed to*. Take for example a phrase such as *in the middle of the*. The first time thisis encountered, it will be processed in a word-by-word manner. However, it will leave an imprint in memory that is strengthened by subsequent exposures (Logan, 1988). Thus, when reading for the first time there would be no difference between *in the middle of* and any other sequence of words. However, when reading itfor the second time, there should be a ‘boost’ in processing speed because we already have an imprint stored in memory (see for example Conklin & Schmitt, 2008; Isobe, 2011; Jiang & Nekrasova, 2007; Northbrook & Conlin in press). Where lexical bundles in materials are nativelike, in that frequent items occur frequently and less frequent items less frequently, this boost will match that of native speakers. Where language is not representative of everyday language, students experience a boost in language they are unlikely to use. Further, because of the imprint left in memory, formulaic sequences such as lexical bundles are highly predictable and tend to be considered more natural sounding (Ellis, 2001; Pawley & Syder, 1983). This means that acquiring lexical bundles that are unnativelike will make it harder to judge the appropriateness of a given phrase, a problem which potentially could impact on a student’s confidence using the language (Wray & Fitzpatrick, 2008). Unfortunately, very little research has looked at materials designed for more general English, such as those used for beginners and/or in primary and secondary school programmes. The current study aims to fill this gap by exploring lexical bundles in a corpus of MEXT approved junior high school textbooks, to determine if they are representative of naturally produced language. In other words, the study examines the following research questions:

1. Do the textbooks contain lexical bundles and provide opportunities for their acquisition?
2. Are the lexical bundles that appear frequently in the English textbooks also frequent in native-speaker English?

If the answer to these questions is *yes*, it will be taken as an indication that the communicative textbooks are representative of naturally produced language, and are preparing students for real-life communication.

**3. Methodology**

A corpus of Japanese junior high school English textbooks (referred to as the JHSETC – the Junior High School English Textbook Corpus) was constructed and consists of 152,966 word tokens, and 6718 word types. Although small, the corpus can be considered exhaustive because it contains all of MEXT’s currently approved textbooks (six series, with three books in each making a total of 18 textbooks). The six series consist of: Columbus 21 English Course (2013), 2279 types and 24,071 tokens; New Crown English Series (2013), 2523 types and 22,142 tokens; New Horizon English Course (2013), 2751 types and 26,598 tokens; One World English Course (2013), 2514 types and 24,845 tokens; Sunshine English Course (2013), 2864 types and 31,481 tokens; and Total English New Edition (2013), 2131 types and 23,829 tokens. It must be noted that transcripts for listening components were not included in the corpus because they were not uniformly available across the different series. This is not considered to be a problem because the listening activities are based on the contents of the textbooks and therefore only add to the existing data, but do not add any new language or lexical bundles **1**. Further, elements such as wordlists, textbook titles and section headings were not included in the corpus.

The language presented in these textbooks is, for the most part, presented as oral language in the form of dialogues and monologues. The SUBTLEXus was selected as the reference corpus. It is made up of scripted language from film and television, that although is not spontaneous speech, it has been written to approximate it. Crucially, the SUBTLEXus has a number of advantages over other spoken corpora such as the COCA’s spoken component, the Santa Barbara Corpus of Spoken American English, the Switchboard corpus or the Buckeye Speech Corpus: 1) it contains both conversations and monologues; 2) it is modern; 3) it includes a very wide range of sources (8,388 files of subtitles from American films and television series); 4) it is the least idiosyncratic of the alternatives mentioned above; and 5) it has the highest correlation with word processing variables (Brysbaert & New, 2009). The SUBTLEXus is also a good approximation of the kind of language American native speakers are regularly exposed to; the Statistic Brain Research Institute found that 99% of American households have TVs, 65% of Americans regularly watch TV while eating dinner, and on average American children spend 1480 minutes per week, or around 3.5 hours a day watching television (Statistic Brain Research Institute, 2017).

One drawback of the SUBTLEXus corpus, however, is that the size of the corpus, at 51 million words of English, makes it problematic as a reference for one of just over 150,000 words. Different approaches are used to deal with size differences amongst corpora. One method is to normalise each frequency into a value per million words (see for example McEnery, Xiao, & Tono, 2006), or alternatively a ‘zipf score’ can be used (van Heuven, Mandera, Keuleers, & Brysbaert, 2014) which ranks items on a logarithmic scale between 1 and 7 (where 7 is considered extremely high-frequency). However, both approaches are problematic when using corpora as small as the JHSETC. For example, taking a 5-word lexical bundle and normalising it would mean that one occurrence in the 150,000 word JHSETC would indicate a frequency of roughly 7/million, or a zipf score of over 4. However, this is out of line with what we know about 5-word lexical bundles, which are already thought to be frequent if they occur at a rate of more than 5/million (Biber et al., 1999). An alternative method involves considering a similarly sized, representative sample of the larger corpus. Ideally, this would be achieved using stratified random sampling techniques (e.g. Biber, 1993) to take a sample from multiple elements within the corpus (in this case, film subtitles), preventing over-representation of items that happen to be highly frequent in a single text. Unfortunately, the SUBTLEXus is only available as a single file of randomised lines (due to copyright restrictions), so this was not possible. Instead, a random sample of 19500 lines was taken from the 51 million word SUBTLEXus to create a corpus of 152,964 tokens. This is somewhat crude as a sampling method, but it should be noted that no special selection criteria were applied to the files making up the SUBTLEXus – rather subtitles were downloaded en-masse and anything with an OCR error rate of less than 2.5% (worked out using the Aspell spelling checker) was included.

Lexical bundle extraction

Lexical Bundles three to six words long, which occurred four or more times in at least three of the six textbook series were extracted from the JHSETC using Antconc (Anthony, 2014). Words containing an apostrophe (e.g. *let’s*, *don’t* or *isn’t*) were considered one word. In the JHSETC the character */* appeared because the textbooks contain many exercises which include alternatives separated by the forward slash (e.g. *Yes, I am. / No, I’m not*). This naturally leads to bundles such as *am no i’m*, which cannot be considered valid items. Thus, extracted items containing a forward slash were removed from the analysis. As recommended by Biber et al. (1999), we also removed items from the JHSETC that crossed speaker turn boundaries (e.g. *do you have i have*). A list of lexical bundles was extracted from the sampled SUBTLEXus. Because the SUBTLEXus is a single text file (as described above) we were unable to include range in the extraction. However, as a confirmative measure, following Biber et al. (1999), we removed any 3- and 4-word items that did not reach the threshold of 10/million, and any 5- 6-word items that did not reach the threshold of 5/million in the full SUBTLEXus. This ensured that none of the lexical bundles from the smaller version of the SUBTLEXus were over-representative of the entire corpus. We then conducted a Pearson's product-moment correlation test comparing the frequency scores from the sampled SUBTLEXus lexical bundles with the scores from the full corpus. The results were significant (*r* (1688) = .82, *p <* .0001), so we can assume that lexical bundles from the SUBTLEXus generalise to the full corpus.

The lexical bundles were classified according to the number of words they contain (3-, 4-, 5- and 6-word lexical bundles) and ranked from most to least frequent. In many cases, a longer lexical bundle contains shorter lexical bundles. For example, *in the middle*, and *the middle of* are both 3-word bundles and are contained in the 4-word bundle *in the middle of*. These, in turn, are part of the 5-word lexical bundle *in the middle of the*. For the purposes of the current study, items such as *in the middle* and *in the middle of* are treated as independent lexical bundles. One important issue is that if a shorter lexical bundle does not occur, or is under-represented in the JHSETC, the larger lexical bundles which contain it will also be under-represented. For example, if the 3-word bundle *do you have* was under-represented in the corpus, this would mean that *do you have any*, *do you have a*, *do you have that*, etc. would also be under-represented.

**4. Analysis and results**

Little is known about lexical bundles in textbooks for teenagers and beginner level students, therefore first we examined whether lexical bundles are common in the textbooks and provide an overview of them as they appear in the JHSETC and compare them to the SUBTLEXus. We then analysed the lexical bundles by comparing and examining the JHSTC items side by side with the SUBTELXus lexical bundles. Following Stubbs and Barth (2003), the analysis is limited to the most frequent items from each lexical bundle size in order to make the data set more manageable. We started with the ten most frequent lexical bundles in the JHSETC and SUBTLEXus corpora at each lexical bundle size and labeled them as follows: (1) items that appear the top 10 of the other corpus, (2) items that appear in the top 50 of the other corpus, (3) items that do not appear in the top 50 of the other corpus, and finally (4) items that do not appear at all in complete reference list. We then counted the number of items in each category, and finally examined them closely for differences and similarities.

* 1. Overview of lexical bundles

The extraction technique outlined above generated four sets of lexical bundles. The distribution of these across the word lengths is displayed in Table 1. There are significantly more lexical bundles in the JHSETC across all sizes by both type and token. This is also true of each word length for both type and token. This suggests that the language used in textbooks is considerably more formulaic than that of language contained in the SUBTLEXus – especially at longer lexical bundle lengths.**2**

**Table 1** Number of lexical bundles by type and token for each of the word lengths in the JHSETC and SUBTLEXus

|  |  |  |  |
| --- | --- | --- | --- |
|  | **JHSETC** | **SUBTLEXus** |  |
|  | **type** | **token** | **type** | **token** | **X2 Type** | **X2 token** |
| 3-word | 1661 | 18739 | 1335 | 10288 | *<* .001 | *<* .001 |
| 4-word | 526 | 4816 | 147 | 996 | *<* .001 | *<* .001 |
| 5-word | 175 | 1430 | 18 | 91 | *<* .001 | *<* .001 |
| 6-word | 73 | 554 | 1 | 4 | *<* .001 | *<* .001 |

* 1. Analysis of frequent lexical bundles.

Lexical bundles clearly make up a substantial part of Japanese learners’ experience of secondary school English. The previous section demonstrated that lexical bundles may be more common in the textbooks than in native speaker language – but are they qualitatively similar? In their corpus analysis of three separate text genres, Stubbs and Barth (2003) demonstrated that lexical bundles were highly frequent in all of them – but that the actual lexical bundles were qualitatively very different. In this case, because textbooks are supposed to represent authentic, communicative language, we would expect to see very little difference between the JHSETC and the SUBTEXus.

As a starting point, we considered the top 10 most frequent items for each lexical bundle size. These are displayed in Table 2. Chi-square statistics for lexical bundles discussed in this section can be found in Appendix 1.

**Table 2** The ten most frequent lexical bundles in the JHSETC and SUBTLEXus corpora at each bundle size, with “\*\*” denoting items that appear the top 10 of the other corpus, “\*” in the top 50, “†” do not appear in the top 50, and underlined do not appear at all.

|  |  |
| --- | --- |
| **3-word lexical bundles** |  |
| JHSETC | i want to\*\*, do you like†, what do you\*\*, do you have†, a lot of\*, i have a\*, yes i am, do you know\*\*, i’m going to\*, this is my† |
| SUBTLEXus | i don't know\*, what are you\*, what do you\*\*, are you doing†,you want to\*, you know what†, out of here†, do you think†, i want to\*\*, do you know\*\* |
| **4-word lexical bundles** |  |
| JHSETC | la la la la, are you going to†, do you want to\*, nice to meet you†, i want to be†, want to be a, what time is it, do you have any, have a lot of, what do you want\*\* |
| SUBTLEXus | what are you doing\*, what do you think\*, get out of here†, what do you want\*\*, i don't want to†, what do you mean†, are you talking about†, what are you talking, i don't know what†, how do you know† |
| **5-word lexical bundles** |  |
| JHSETC | la la la la la, i want to be a, what do you want to\*, are you going to do, what are you going to, how many cds do you, many cds do you have, a junior high school student, i’m going to talk about, what time do you get |
| SUBTLEXus | what are you talking about, what are you doing here, you know what i mean, what are you gonna do, i don't know what to†, in the middle of the†, what do you want to \*\*, know what you're talking about, what are we gonna do, what do you know about † |
| **6-word lexical bundles** |  |
| JHSETC | la la la la la la, how many cds do you have, what are you going to do, what time do you get up, called to say i love you, just called to say i love, i'm a junior high school student, i'm going to talk about my, what do you want to do, i just called to say i |
| SUBTLEXus | get the hell out of here |

*4.2.1 Three-word lexical bundles.*

Fifteen items (30%) appear in the top 50 most common items from the JHSETC that do not appear at all (0 frequency) in the SUBTLEXus and 4 items (8%) appear in the SUBTLEXus but not in the JHSETC (0 frequency). Further, 22 (44%) items from the JHSETC do not appear in the top 50 for the SUBTLEXus, and 37 (74%) items from the SUBTLEXus do not appear in the top 50 for the JHSETC.

Notably, the bundle *yes i am*—which does not appear at all in SUBTLEXus (0 frequency)—is in the top 10 most frequent items from JHSETC. There are two items in the top 10 for JHSETC that do not appear in the top 50 of SUBTLEXus (*do you like*, *this is my*) and three items in the top 10 of the SUBTLEX that do not appear in the top 50 of JHSETC (*are you doing*, *out of here*, *do you think*). Three items appear in the top 10 items for both the JHSETC and the SUBTLEXus, but have a significant difference in frequency in the two corpora: *what do you,* *do you know*, and *i want to*.

The top 50 from the JHSETC contains seven yes/no response bundles, four of which, along with the bundle *yes i am*, do not appear in the SUBTLEXus at all. The remaining three do not appear in the top 50 from the SUBTLEXus. In fact, the only such bundle in the SUBTLEXus top 50 is *no no no*, which although qualitatively different to those from the JHSETC in that it does not follow the set *yes/no+pronoun+verb* pattern is relatively common in the SUBTLEXus with 33 occurrences but does not appear at all in the JHSETC. Related to these, *do you~* patterns are also the most common lexio-grammatical patterns in both lists. These are displayed below, in Table 3 along with the four word continuations of these *do you~* patterns –*what do you~* patterns, which we discuss in more detail in the following section.

**Table 3.** Comparison of yes/no fragment bundles, *do you~* and *what do you~* bundles in the JHSETC and the SUBTLEXus corpus. Raw frequency scores are given in brackets for references as (JHSETC – SUBTLEXus).

|  |
| --- |
| **Yes/ No fragments** |
| JHSETC | yes i am (100 – 0), yes i do (88 – 4), no i don't (65 – 15), no i'm not (65 – 6), yes i have (60 – 0), yes it is (46 – 0), yes i did (42 – 0) |
| SUBTLEXus | no no no (0 – 33) |
| **Do you~ bundles in the most common 50 items for each corpus** |
| JHSETC | do you like (171 – 10), do you have (165 – 17), do you know (98 – 45), do you want (87 – 40), do you play (56-0) |
| SUBTLEXus | do you know (98 – 45), do you think (38 – 48), do you want (87 – 40), do you mean (14 – 25) |
| **What do you~ bundles in the most common 50 items for each corpus** |
| JHSETC | what do you want (30 – 20), what do you think (26 – 23), what do you have (23 – 0), what do you do (22 – 0) |
| SUBTLEXus | what do you think (26 – 23), what do you want (30 – 20), what do you mean (10 – 19), what do you know (0 – 8) |

Notably in the yes/no response bundles, only three of the items (*yes i do*, *no i don’t, no i'm not*) appear at all in the SUBTLEXus corpus, and these are very low-frequency and do not appear in the most frequent 50 3-word bundles. The overall difference between these items is significant – *yes i do*, *no i don’t*, and *no i’m not.* This is not, however, to say that question fragments are less frequent in the SUBTLEXus. To elicit a “yes” or “no” response of the type seen in the textbooks, a *do you~* or *are you~* question would be required. Both the JHSETC and the SUBTLEXus contain one *are you~* bundle each (*are you going* and *are you doing* respectively), as well as five *do you~* bundles for the JHSETC and four for the SUBTLEXus. This suggests then that although questions which elicit yes/no responses are equally common in both corpora, there is a large difference in the way these are responded to. Unfortunately, because the lines of the SUBTLEXus are randomised, it is not possible to say exactly how *are you~* and *do you~* bundles are responded to, however this is taken up below, briefly as part of the discussion of 4-word bundles, and in more detail for 6-word lexical bundles.

In both corpora, *do you~* bundles are the most common (see Table 3), but the verbs that end the pattern are quite different. The lexical bundle *do you play* appears in all six of the textbook series in the JHSETC, but does not appear at all in the SUBTLEXus, and in fact, this lexical bundle only appears 136 times in the full SUBTLEXus (2.7/per million). Turning to the most common *do you~* bundles from the SUBTLEXus, we see that they are less frequent than in JHSETC. However, the difference between two of the bundles is not significant (*do you think*, *do you mean*), suggesting that they are well represented in JHSETC. The other two lexical bundles occur more frequently in the JHSETC (*do you know, do you want*) which, while over-representing their frequency of occurrence, should give students the opportunity to acquire them.

*4.2.2 Four-word lexical bundles.*

Forty (80%) of the 4-word lexical bundles in the top 50 from JHSETC do not appear at all in the SUBTLEXus. Of the remaining 10 items, six appear in the top 50 of the SUBTLEXus. Sixteen (32%) items in the top 50 of the SUBTLEXus do not appear at all in the JHSETC. One of these contains a word that does not appear in the textbooks—“bitch” from *son of a bitch*—and would be considered inappropriate for junior high school students, but does represent an example of very casual language, something which seems to be missing from textbooks.

Only one lexical bundle—*what do you want*—appears in the top 10 of both lists. Five items from the JHSETC top 10 do not appear in the SUBTLEXus list. Notable in Table 2 is the phrase *what are you talking*, which appears in the top 10 most frequent items in SUBTLEXus but not at all in the JHSETC. In fact, *what are you* was also in the top 10 for the SUBTLEXus for the 3-word bundles, too, but only in the top 50 for the JHSETC. Interestingly *what do you mean,* arguably a polite synonym for *what are you talking about*,does appear in the textbooks 10 times, once again suggesting that the language contained in the textbooks is politer than that of the SUBTLEXus.

In terms of lexio-grammatical constructions, there is a great deal of variation in the types of bundles present in both corpora. However, in both *what do you~* phrases are the most common and are shown in Table 3. These are all extended versions of *do you~* patterns. The difference in frequency is significant for these items overall (*X*2 (5) = 42.903, *p <* .001), but insignificant for the individual bundles – *what do you think*, *what do you want*, *what do you mean*. However, the difference is significant for the lexical bundles *what do you know, what do you have* and *what do you do,* which all have zero occurrences in SUBTLEXus.

*4.2.3 Five-word lexical bundles.*

The first notable point with the 5-word lexical bundles is that there are only 18 in the SUBTLEXus. Importantly, only three of these appear in the JHSETC and of these only two are in the top 50. There is only one lexical bundle from the top 10 of either corpus that appears in the other: *what do you want to* (3rd most common in the JHSETC and 7th in the SUBTLEXus). The only other item that appears in both corpora is *how long have you been*, ranked at number 30 in the JHSETC and number 19 in the SUBTLEXus.

Unlike the smaller lexical bundle sizes, for 5-word sequences, there is very little repetition in terms of lexio-grammatical patterns. Apart from three patterns which are repeated twice (*are you going to~*, *do you want to~*, and *do you have any~*), all the items only occur only once in the JHSETC. Interestingly, none of the three repeated patterns appear in the SUBTLEXus. Similarly, in the SUBTLEXus only two patterns appear more than once (*what are you ~ing ~* and *what are ~ gonna do*). As with the smaller lexical bundles, these are predominantly versions of the *do you~* pattern, however, unlike previous lexical bundle sizes, we have reached a stage where the items cease to be frequent (or even appear) in the SUBTLEXus list.

*4.2.4* *Six-word lexical bundles.*

While there are 73 6-word lexical bundles in the JHSETC, there is only one in the SUBTLEXus (*get the hell out of here)* and it does not correspond to any in the JHSETC, likely because the language is not appropriate for a junior high school textbook.

Three lexical bundles that appear in the top 10 items from the JHSETC are from the longer pattern *I just called to say I love you* which is from the song ‘*Stand by Me’* by Stevie Wonder that appears in four of the textbook series. The most frequent bundle is *la la la la la la*, which again comes from a song (‘*Sing’* by Joe Raposo, appearing in three of the series) and would not normally be considered a lexical bundle. After the song lyrics, the most common 6-word bundles in the corpus are *how many cds do you have* and *what are you going to do*, both with 22 occurrences. Interestingly neither of these appear in the abridged SUBTLEXus, and *how many cds do you have* does not appear in the full SUBTLEXus. We also only see one recurring pattern in the top 50 items – *what do you want to~* (*what do you want to do* ranked at number 9 with13 occurrences on the JHSETC and *what do you want to be* ranked at number 11, with 9 occurrences in the JHSETC), neither of which occur in the sampled SUBTLEXus, however they do appear in the full SUBTLEX (*what do you want to do* is frequent with 7.47/ million occurrences, and *what do you want to be* is quite uncommon with 0.68/million occurrences).

As mentioned above, the lexical bundle *what are you going to do* appears in the textbooks 22 times, while *what are you gonna do* does not appear at all. In the full SUBTLEXus, the lexical bundle occurs 801 times (16/million), but significantly less than the 5-word bundle *what are you gonna do* containing the less formal *gonna* (1438 occurrences, or 28/million – *X*2 (1) = 181.23, *p <* .001). This indicates that the less formal, arguably more “conversational” language is more common in the SUBTLEXus reference corpus than in the JHSETC. If we look at this more closely and assume that contracted forms such as *i’m*, *we’re*, *gonna* and *wanna* are indicative of casual language, whereas *i am*, *we are*, *going to*, and *want to* are indicative of more formal language, we would expect casual language to contain a larger number of lexical bundles that include a contracted form. This is indeed the case – the JHSETC contains 176 lexical bundle types (7.7% of the list) including a contraction, whereas the SUBTLEXus contains 259 bundle types (17.24%). The difference is significant *X*2 (1) = 15.84, *p <* .001.

There is only one 6-word item in the SUBTLEXus, and this is something most would consider inappropriate for a secondary school audience; therefore, it is fair to say that nothing is *missing* as such from the textbooks in terms of 6-word lexical bundles – although it is arguably another example of the more casual register used in the SUBTLEXus.

**5. Discussion**

The overview of lexical bundles in the JHSETC revealed that recurrent strings of words are very frequent in the textbooks – more so than in conversational English. Further, this is true of all bundle sizes. The fact that the JHSETC has more repetitions of lexical bundles could be considered beneficial, as repetition is an important component in sequence acquisition (Ellis, 1996; Webb, Newton, & Chang, 2013). In other words, if certain lexical bundles are an important aspect of conversational competence, repeating them will help them be learned.

An important follow up question is whether the lexical bundles in the JHSETC are representative of conversational English. The analysis demonstrated that to an extent the 3-word lexical bundles found in the JHSETC are like those found in the SUBTLEXus corpus. As the lexical bundles become longer, they seem to deviate from the patterns seen in actual conversational English. Four-word bundles only show a small amount of overlap in the corpora, and 5- and 6- word bundles are almost completely different. Further, the high-utility items that we see for the 3-word bundles appear to be, for the most part, embedded in considerably lower utility longer lexical bundles. A good illustrative example is the phrase *how many CDs do you have?*, which is likely of low utility. However, it contains high utility bundles such as *do you have*. This example, despite being the most frequent 6-word lexical bundle in the JHSETC, does not appear in the full SUBTLEXus corpus, the COCA, or the BNC. Neither is the pattern *how many ~ do you have?* particularly frequent in the full SUBTLEXus (88 occurrences or 1.72/million), with the most frequent filler being *children* and *kids* (both with just 10 occurrences each). With “*How many children do you have?”* being the most frequent bundle in English, the inclusion of *how many ~ do you have?* is hard to justify for a teenage audience. This example points to a limitation of the current study; we did not systematically examine the relationship between lexical bundles embedded into larger lexical bundles. This would be an interesting question to explore in future research.

Although we see promising results at small lexical bundle sizes, we must conclude that the language presented in junior high school English textbooks is not sufficiently representative of language outside the classroom. This finding aligns with previous research looking at other kinds of textbooks for more advanced learners (Chen, 2010; Mcaleese, 2013; Nguyen & Ishitobi 2012; Wood, 2010) as well as findings from Koya (2014) and Tsai (2014). Perhaps a part of this is the tendency for textbook designers to script and tailor dialogues to specific goals – generally the illustration of a certain grammar point – without reference to real language (see Gilmore, 2015 for a discussion). Put another way, the lexical bundles in the JHSETC are examples of grammatically correct, possible language – however, they are not examples high-frequency probable language (Lewis, 1993). A good example of this is the lexical bundle *what time do you get up.* This is one of only three 6-word lexical bundles that appears in all of the six-textbook series, yet it only appears twice in the full SUBTLEXus (0.03/million). Unlike the instances in the JHSETC, both are used for a very specific purpose:

1. What time do you get up? Why? Because we wanna cook you breakfast.
2. We're civilized people. Thank you. What time do you get up? About 8, I guess. I'll fix you breakfast.

In both instances from the SUBTLEXus, the question is being used to inform the interlocutor that breakfast will be ready, and therefore we would like to know what time you want it. In the textbooks, however, the item seems to be included purely as an example of a question used in relation to talking about time, usually as a simple question–answer pair (“*What time do you get up?*”, “*I get up at ~”*). Ishihara and Cohen (2010) point out that in many cases textbooks fail to take pragmatics into account. This certainly seems to be the case with the illustrative example here, as well as throughout the junior high school textbooks analysed in the current study.

More generally, the question/answer fragments identified in the corpora reveal several patterns. First, there are many formulaic question and answer patterns in the JHSETC which do not appear in the SUBTLEXus (e.g. *yes i am*, *no i’m not*). Second, the way native speakers answer yes/no questions is complex, and often answers do not explicitly state “yes” or “no” (Hockey, Rossen-Knill, Spejewski, Stone, & Isard, 1997). Corpus studies, such as Biber et al. 2004, have found that although yes-no question fragments are very common, the units of analysis are generally a fragment of the questions themselves and not the responses – presumably because no specific responses to these questions are frequent enough to appear in their lists of lexical bundles. Arguably, the inclusion of so many formulaic yes/no response fragments is beneficial for production, as it gives learners a simple way to construct their responses. However, it indicates to learners that this is what they should produce in the real world, and they will be listening for a “yes” or “no” that in many cases will not actually come. Third, there is a difference in the level of formality. High-frequency question fragments in the SUBTLEXus, such as *what are we gonna do,* are not present in the textbooks. The SUBTLEXus bundles contain significantly more contractions. The analysis suggests that the language contained in textbooks is far politer than standard, everyday conversation. An in-depth analysis of register within the two corpora goes beyond the scope of this study but would be an interesting area for further research.

The current findings have implications for “conversational” approaches and textbooks. There is a large literature in the Japanese context trying to explain the often poor level of achievement in English (see for example Hagerman, 2009). Could it simply be that students do get good at English, but it is in a genre that is misaligned with English that is encountered in the real world? Given some of the research discussed in the Introduction (Koya, 2004; Romer, 2004, 2005; Tsai, 2014) and the findings from the current study, this seems like a possibility. Recent research has assessd the extent to which junior high school students have a sensitivity to the lexical bundles that are present in their textbooks. Northbook and Conklin (in press) presented Japanese junioir high school students with 3-word lexical bundles identified in this corpus study (*do you play*) and matched ones that were frequent in the SUBTLEXus but not their textbooks (*do you hear*). They found a processing advantage for the textbook bundles over the SUBTLEXus ones that was graded by textbook frequency. This demonstrates the powerful role that the textbook input has on the learners’ processing. Thus, we can theorise that replacing language in the textbooks with language that is 1. attuned to the needs of junior high school students, 2. represents the language they are likely to use and encounter and 3. incorporates frequent lexical bundles, as derived from corpora, should be beneficial. More precisely, there may be frames like *how many ~ do you have?*, which are not frequent in everyday language, but which are deemed to be important for a particular grammar point. A corpus could be used to inform how such language is presented in textbooks so that it most closely matches the real world. The SUBTLEXus shows us that this frame most often comes up in the context of *how many kids/children do you have*. An appropriate adaptation of this for junior high school students might be *how many brothers/sisters do you have*. Additionally, there are instances where the utility of textbook language can be increased simply by making a lexical bundle less formal; for instance, replacing *what are we going to do* with *what are we gonna do*. However, in other cases, there is not a more frequent equivalent. This is the case with *what time do you get up,* simply because this is not something we often say. In fact, the pattern *what time do you~* itself is rather low-frequency with only 131 (2.56/million) occurrences in the full SUBTLEXus.

This does not mean that language like *what time do you~* should not be included in the textbooks per-se. Rather it could perhaps be better demonstrated with something more open in meaning, allowing for it to be more easily combined with other more frequent language. For example, *what time do you get up?* could be replaced by *what time do you get off?* (27 occurrences in the SUBTLEXus – 0.53/million), although this is potentially unrelated to junior high school students in that they generally do not have jobs. The similar *what time do you finish?* (seven occurrences in the SUBTLEXus, or 0.14/million), or *what time do you want to ~* (six occurrences in the SUBTLEXus, or 1.2/million) are other alternatives. None of these examples are what would be considered “high-frequency” language, but are more open in meaning and can more realistically be incorporated into dialogues that comprise a range of topics. For example, one activity that most junior high school students can be expected to engage in is that of after-school club activities, and it is not difficult to create a dialogue that contains *“what time do you finish? I want to talk to you”* (5.14/million in the full SUBTLEXus). Or we can easily imagine a situation where a character, upon hearing his classmate has basketball/football/piano/etc. practise, asks, *“What time do you get off? Let’s go to* (15.80/million) *the mall after*.” Another very common 5-word bundle in the JHSTC is *do you have any pets* which only occurs 5 times (0.09 /million) in the full SUBTLEXus, and thus cannot be considered high-utility language. Yet by changing the word *pets* to *idea(s),* the frequency becomes 631 – 12.57/million. This is a large difference, and perhaps more importantly, it seems unlikely that most junior high school students are interested in whether someone has pets. However, for most of them – especially in their third and final year – topics such as what high school their friends are planning to go to almost certainly will be top of mind. It is not difficult, then, to imagine conversations where students ask each other *“Do you have any idea what high school you want to go to?”* or “*What are you gonna do next year?”.* These are very simple examples but by considering the things students are likely to say, and combining that with data from corpora, coming up with textbooks full of high-utility language attuned to students’ needs should not be difficult. Improving the quality of textbook content requires a top-down process of starting with overall topics that teenagers are likely to talk about, and working from there to engineering realistic dialogues that contain high-frequency language, distributed in proportion to what we would expect in a real conversation. Put simply, textbook designers are recommended to ask *what are you talking about?* and base materials on that.

**6. Conclusion**

Our study, working on the framework laid out by Stubbs and Barth (2003), compared lexical bundles in foundational, junior high school ‘communicative’ EFL textbooks and a corpus of spoken English and has highlighted major differences between the two. Crucially, to follow the precepts of the communicative approach, such textbooks should: 1) present and model authentic conversational language; and 2) the relative frequency of the input should mirror that of actual spoken English (i.e. high-frequency language should be encountered more frequently than low-frequency language). Our findings suggest that both of these elements are lacking. While studies of advanced learners in EAP contexts (Chen, 2010; Biber et al., 2004; Wood, 2010) have identified similar issues, this study gives insights into the kind of input *very early* learners receive. This is important because, as Northbrook and Conklin (in press) demonstrate, early learners do develop a processing advantage for the lexical bundles in their textbooks (as would be expected form a usage-based approach). We conclude that the textbooks series under investigation are not representative of anything other than themselves; that is, they are a separate genre of English, not the representation of general conversational English that they claim to be. Thus, despite the claims that designers may make (Ogura, 2008), the textbooks fail in their role of providing students with authentic samples of the language they are likely to need in the real world.

**Notes**

**1** To demonstrate that including listening scripts to the corpus makes little difference to the overall data set, we transcribed the listening exercises from the “New Crown” series. In total, it would have added 5019 words to the corpus. On the assumption that the listening activities in the other series are of a similar in size, this would have added around 30,000 words to the overall corpus to take it from around 150,000 words to 180,000 words. Once we had added the listening activities from the New Crown textbooks series, we generated a list of lexical bundles from the listening ‘texts’. The listening scripts had 41 lexical bundles, 10 of which did not appear in the textbook version, but three of which are actually part of a larger lexical bundle: *does she play, high will be, the high will, the high will be, an old man, does he play, he doesn't play, kind of bird, something that has, who am i.* We compared the top 10 lexical bundles from the New Crown minus listening scripts and the New Crown plus listening scripts. There were no differences and therefore we conclude that including the listening scripts does not change the overall findings of our study. That being said, if the listening scripts were available for all the textbook series, it would be more ecologically valid to include them.

**2** We also considered whether there were differences in lexical bundle types across the six-textbook series. When frequency scores are normalized per 1000 words, there was no difference in the number of lexical bundles at any size (χ2 > 0.43). However, when looking at the actual lexical bundles themselves, there is very little overlap. At 3-words there is only 4.65% overlap in lexical bundles across the series, 2.06% at 4-words, 0.99% for 5-words, and 0.37% for 6-words, yielding a total of 7% overlap in lexical bundles across the series.

**References**

Anthony, L. (Producer). (2014). AntConc (Version 3.4.3) [Computer Software]. Retrieved from http://www.laurenceanthony.net/

Barlow, M., & Kemmer, S. (2000). *Usage Based Models of Language*. United States: CSLI.

Bednarek, M. (2012). “Get us the hell out of here”: Key words and trigrams in fictional television series. *International Journal of Corpus Linguistics, 17*(1), 35-63.

Biber, D. (1993). Representativeness in Corpus Design. *Literary and Linguistic Computing, 8*(4), 243-257.

Biber, D. (2006). *University Language: A corpus-based study of spoken and written registers*. Amsterdam: John Benjamins Publishing Company.

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.

Biber, D., Conrad, S., & Cortes, V. (2004). If you look at ...: Lexical Bundles in University Teaching and Textbooks. *Applied Linguistics, 25*(3), 371-405.

Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *Longman grammar of spoken and written English*. Essex: Pearson Education Limited.

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.

Breeze, R. (2013). Lexical bundles across four legal genres. *International Journal of Corpus Linguistics, 18*(2), 229-253.

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.

Bybee, J. (2006). From Usage to Grammar: The Mind's Response to Repetition. *Language, 82*(4), 711-733.

Chen, L. (2010). An Investigation of Lexical Bundles in ESP Textbooks and Electrical Engineering Introductory Textbooks. In D. Wood (Ed.), *Perspectives on formulaic language* (pp. 107-125). London: Continuum.

Columbus 21 English Course. (2013). Tokyo: Mitsumura Tosho.

Conklin, K., & Schmitt, N. (2008). Formulaic Sequences: Are They Processed More Quickly than Nonformulaic Language by Native and Nonnative Speakers? *Applied Linguistics, 29*(1), 72-89.

Conklin, K., & Schmitt, N. (2012). The Processing of Formulaic Language. *Annual Review of Applied Linguistics, 32*, 45–61.

Conrad, S., & Biber, D. (2004). The Frequency and Use of Lexical Bundles in Conversation and Academic Prose. *Lexicographica, 20*, 56-71.

Cortes, V. (2004). Lexical bundles in published and student disciplinary writing: Examples from history and biology. *English for Specific Purposes, 23*(4), 397-423.

Dickinson, P. (2012). Improving second language academic presentations with formulaic sequences. *Bulletin of Niigata University of International and Information Studies Department of Information Culture, 15*, 25-36.

Ellis, N. C. (1996). Sequencing in SLA: Phonological Memory, Chunking, and Points of Order. *Studies in Second Language Acquisition, 18*, 91-126.

Ellis, N. C. (2001). Memory for Language. In P. J. Robinson (Ed.), *Cognition and Second Language Acquisition* (pp. 33-68). New York: Cambridge University Press.

Ellis, N. C., O'Donnell, M. B., & Römer, U. (2013). Usage-Based Language: Investigating the Latent Structures That Underpin Acquisition. *Language Learning, 63*, 25-51.

Erman, B., & Warren, B. (2000). The idiom principle and the open choice principle. *Text, 20*(1), 29-62.

Gilmore, A. (2004). A comparison of textbook and authentic interactions. *ELT Journal, 59*(4), 363-374.

Gilmore, A. (2015). Research into practice: The influence of discourse studies on language descriptions and task design in published ELT materials. *Language Teaching,* 48(04), 506-530.

Gray, B., & Biber, D. (2013). Lexical frames in academic prose and conversation. *International Journal of Corpus Linguistics, 18*(1), 109-136.

Hagerman, C. (2009). English language policy and practice in Japan. *Osaka Jogakuin University, Departmental Bulletin Paper, 6*, 47-64.

Hockey, B. A., Rossen-Knill, D., Spejewski, B., Stone, M., & Isard, S. (1997). Can you predict answers to Y/N questions? Yes, No and Stuff. *Proceedings of Eurospeech 1997*.

Hyland, K. (2008). As can be seen: Lexical bundles and disciplinary variation. *English for Specific Purposes, 27*(1), 4-21.

Ishihara, N., & Cohen, A. D. (2010). *Teaching and Learning Pragmatics: where language and culture meet*. Edinburgh: Pearson.

Isobe, Y. (2011). Representation and processing of formulaic sequences in L2 mental lexicon: How do Japanese EFL learners process multi-word expressions. *JACET Kansai Journal, 13*, 38-49.

Jiang, N., & Nekrasova, T. M. (2007). The Processing of Formulaic Sequences by Second Language Speakers. *The Modern Language Journal, 91*(3), 433-445.

Koprowski, M. (2005). Investigating the usefulness of lexical phrases in contemporary coursebooks. *ELT Journal, 59*(4), 322-332.

KOYA, T. (2004). Collocation research based on corpora collected from secondary school textbooks in Japan and in the UK. *Dialogue, 3*, 7-18.

Kuiper, K. (2004). Formulaic performance in conventionalised varieties of speech. In N. Schmitt (Ed.), *Formulaic Sequences: Acquisition, processing and use* (pp. 37-54). Amsterdam: John Benjamins.

Kwon, Y.-E., & Lee, E.-J. (2014). Lexical Bundles in the Korean EFL Teacher Talk Corpus: A Comparison Between Non-native and Native English Teachers. *The Journal of Asia TEFL, 11*(3), 73-103.

Lewis, M. (1993). *The lexical approach*. Hove: Language Teaching Publications.

Logan, G. D. (1988). Toward an Instance Theory of Automatization. *Psychological Review, 95*(4), 492-527.

Mcaleese, P. (2013). *Investigating Multi-Word Items in a Contemporary ELT Course Book.* Paper presented at the JALT2012 Conference Proceedings, Tokyo.

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

Meunier, F. (2012). Formulaic Language and Language Teaching. *Annual Review of Applied Linguistics, 32*, 111-129.

Myles, F., Hooper, J., & Mitchell, R. (1998). Rote or Rule? Exploring the Role of Formulaic Language in Classroom Foreign Language Learning. *Language Learning, 48*(3), 323-363.

Myles, F., Mitchell, R., & Hooper, J. (1999). Interrogative chunks in French L2: A Basis for Creative Construction? *Studies in Second Language Acquisition, 21, 40-80*.

Nattinger, J. R., & DeCarrucio, J. S. (1992). *Lexical Phrases and Language Teaching*. Oxford: Oxford University Press.

New Crown English Series. (2013). Tokyo: Sanseido.

New Horizon English Course. (2013). Tokyo: Tokyo Shoseki.

Nguyen, H., & Ishitobi, N. (2012). Ordering Fast Food: Service Encounters in Real-Life Interaction and in Textbook Dialogs. *JALT Journal, 34*(2), 151-186.

Northbrook, J., and Conklin, K. (in press). Is what you put in what you get out? – Textbook-derived lexical bundle processing in beginner English learners, *Applied Linguistics.*

Nunan, D. (1991). *Language Teaching Methodology: a textbook for teachers*. Hertfordshire: Prentice Hall.

Ogura, F. (2008). Communicative competence and senior high school oral communication textbooks in Japan. *The Language Teacher, 32*(12), 3-8.

One World English Course. (2013). Tokyo: Kyoiku Shuppan.

Pawley, A., & Syder, F. H. (1983). Two puzzles for linguistic theory: nativelike selection and nativelike fluency. In J. C. Richards (Ed.), *Language and Communication* (pp.191-225). London: Longman.

Römer, U. (2004). Comparing real and ideal language learner input: the use of an EFL textbook corpus in corpus linguistics and language teaching. In G. Aston, S. Bernardini & D. Stewart (Eds.), *Corpora and Language Learners* (pp. 151-168). Amsterdam: John Benjamins.

Römer, U. (2005). *Progressives, Patterns, Pedagogy: a corpus-driven approach to english progressive forms, functions, contexts and didactics*. Amsterdam: John Benjamins Publishing Company.

Simpson-Vlach, R., & Ellis, N. C. (2010). An Academic Formulas List: New Methods in Phraseology Research. *Applied Linguistics, 31*(4), 487-512.

Siyanova-Chanturia, A., & Martinez, R. (2014). The Idiom Principle Revisited. *Applied Linguistics*, 1, 1-22.

Statistic Brain Research Institute (2017). Television Watching Statistics. Retrieved from https://www.statisticbrain.com/television-watching-statistics/

Stubbs, M., & Barth, I. (2003). Using recurrent phrases as text-type discriminators: A quantitative method and some finding. *Functions of Language, 10*(1), 61-104.

SUBTLEXus. Universiteit Gent. Available from http://expsy.ugent.be/subtlexus/

Sunshine English Course. (2013). Tokyo: Kairyudo.

Taguchi, N. (2007). Chunk learning and the development of spoken discourse in a Japanese as a foreign language classroom. *Language Teaching Research, 11*(4), 433-457.

Total English New Edition. (2013). Tokyo: Gakko Tosho.

Tsai, K.-J. (2014). Profiling the Collocation Use in ELT Textbooks and Learner Writing. *Language Teaching Research, 19*(6), 723-740.

Tyler, A. (2010). Usage-Based Approaches to Language and Their Applications to Second Language Learning. *Annual Review of Applied Linguistics, 30*, 270-291.

van Heuven, W. J., Mandera, P., Keuleers, E., & Brysbaert, M. (2014). SUBTLEX-UK: A new and improved word frequency database for British English. *The Quarterly Journal of Experimental Psychology, 67*(6), 1176-1190.

Webb, S., Newton, J., & Chang, A. (2013). Incidental Learning of Collocation. *Language Learning, 63*(1), 91-120.

Wood, D. (2010). Lexical Clusters in an EAP Textbook Corpus. In D. Wood (Ed.), *Perspectives on formulaic language* (pp. 88-106). London: Continuum.

Wray, A. (2002). *Formulaic Language and the Lexicon*. Cambridge: Cambridge University Press.

Wray, A., & Fitzpatrick, T. (2008). Why can't you just leave it alone? Deviations from memorized language as a gauge of nativelike competence. In F. Meunier & S. Granger (Eds.), *Phraseology in Foreign Language Learning and Teaching* (pp. 123-147). Amsterdam: John Benjamins.

Yaoyu, W., & Lei, L. (2011). Lexical Bundles in the Academic Writing of Advanced Chinese EFL Learners. *RELC Journal, 42*(2), 155-166.

**Appendix 1**

**Table 4.** Chi-square scores testing the difference between JHSETC frequency and SUBTLEXus requency (ranked in the order they are discusssed in the paper).

|  |  |  |  |
| --- | --- | --- | --- |
| **Lexical bundle** | **DF\*** | **χ 2** | **p** |
| what do you | 1 | 17.966 | < .001 |
| do you know | 1 | 19.643 | < .001 |
| i want to | 1 | 80.161 | < .001 |
| yes i do | 1 | 76.696 | < .001 |
| no i don’t | 1 | 31.25 | < .001 |
| no i’m not | 1 | 49.028 | < .001 |
| do you like | 1 | 143.21 | < .001 |
| do you have | 1 | 120.35 | < .001 |
| do you want | 1 | 76.696 | < .001 |
| do you play | 1 | 56 | < .001 |
| do you think | 1 | 1.1628 |  = 0.28 |
| do you mean | 1 | 3.1026 |  = 0.08 |
| what do you think | 1 | 0.18367 |  = 0.67 |
| what do you want | 1 | 2 |  = 0.16 |
| what do you mean | 1 | 2.7931 |  = 0.09 |
| what do you know | 1 | 8 |  = 0.004 |
| what do you have | 1 | 23 | < .001 |
| what do you do | 1 | 22 | < .001 |
| DF\*, degree of freedom |