A Corpus-based Analysis of Collocations in Korean Middle and High School English Textbooks

Young Shin Kim[1] & Sun-Young Oh[2†]
[1]Hansan Middle School, [2]Seoul National University

ABSTRACT

This study analyzed collocations in Korean middle and high school English textbooks based on the 2015 revised national curriculum. All 1,718 nouns from the curriculum wordlist were selected as node words and paired with their collocates statistically verified using a billion-word reference corpus to better represent the existing lexical syllabus. The analysis revealed that collocation density was higher in the textbooks, with readers encountering one collocation per 16–17 words. However, collocations in the textbooks showed insufficient repetition and a narrower range of association strength. Fewer repetitions and a limited collocational repertoire led to a weaker correlation between the two variables. This suggests that Korean learners may not benefit enough from frequent encounter to consolidate their lexical knowledge or distinguish different collocational strength levels. These findings call for considering the level of repetition and association strength of collocations in developing the English curriculum and materials.

Keywords: collocation, density, repetition, association strength, English textbooks

1. Introduction

The past few decades have seen research in (applied) linguistics establish the significance of collocations in language use and acquisition. Collocation, or "semi-preconstructed phrases that constitute single choice" (Sinclair, 1991, p. 110), forms a basic psychological unit that is stored as a whole in the mental lexicon of language users (Hoey, 2005; Schmitt, 2010; Stubbs, 2001). Native speakers seem to make extensive use of collocations because they can retrieve them as "chunks" (Ellis, 1996, 2002) from their memory without having to generate or analyze the sequences in segments (Cowie, 1992; Erman & Warren, 2000; Howarth, 1998; Sinclair, 1991;

* Acknowledgement: The authors would like to express deep gratitude to three anonymous reviewers for useful suggestions, and to Hyun Soo Kim for his invaluable help with retrieval and processing of corpus data through a customized program.
† Corresponding author: sunoh@snu.ac.kr

Wray, 2005). Abundant evidence has been accumulated on the pervasiveness of formulaic language in English. Cowie (1992), for example, measured collocational density in native writings and found that more than 40% of verb-noun pairs were already well-established collocations.

As a result of the extensive and repeated exposure to word associations in natural input, formulaic language is believed to help speakers reduce working memory storage and offer the advantages of automated processing, fluent language production, and native-like idiomaticity (Pawley & Syder, 1983; Segalowitz, 2010; Wray, 2005). Furthermore, Ellis (2002) observed that children's grammar acquisition is a gradual process, which begins with picking up frequent formulas, through low-scope patterns, to ultimately generalizing more abstract constructions. He hypothesized that the ready-made units may reduce cognitive demands and facilitate further language development. Since the early 1980s, the critical role of collocation and formulaic language in native-like production has been recognized in second language acquisition (Cowie, 1992; Nattinger & DeCarrico, 1992; Pawley & Syder, 1983) and EFL coursebooks development (Harmer & Rossner, 1997; McCarthy & O'Dell, 1994). For example, Erman and Warren (2000) suggest that raising awareness of the abundance of existing prefabrications would improve students' learning strategies and command of English and that teaching materials should be adapted to more precisely represent the native-like use of language.

While many coursebook designers have recognized the significance of collocation in language learning, the near absence of a unified lexical syllabus focusing on the multi-word unit has prevented them from putting these ideas into practice systematically. Collocation items are often chosen by the subjective judgment of the publishers without consistent and reliable criteria. The lack of a preselected list of collocations may lead to the overuse of items with little pedagogical usefulness. Koprowski (2005) investigated the usefulness of lexical phrases in different coursebooks written by major international publishers and found that a quarter of collocations presented in the textbook were not as useful in natural language. More recent studies have statistically verified collocations by native reference corpus and analyzed their distribution in the textbook materials. By exploiting British National Corpus (BNC) to identify collocations compatible with the national English curriculum wordlist in Taiwan, Tsai (2015) explored the representation of collocations in the textbook series and native writings. He found that a majority of collocations in the textbooks were not recycled in a principled manner to facilitate learning, while the recurrent collocations did not appear to be worth the class time.

Collocation learning has also not been fully incorporated into Korea's lexical syllabus in its national English curriculum. Unlike the curriculum wordlist, which has been continuously revised to improve its representativeness based on highly objective criteria, such as word frequency, range, and teacher ratings for item familiarity (Lee & Shin, 2015), only marginal attention has been paid towards the development of a curriculum-related, statistically verified collocation list exploiting a large-scale reference corpus data. A few researchers (Lee, 2009, 2015; Shin, 2019) have thus called for the need to establish curriculum guidelines and objective criteria to select pedagogically meaningful collocations. Research has also pointed out the restricted collocation repertoire in textbooks, for example, limited in their diversity of Part of Speech (POS) types (Kim, 2004; Shin, 2019), complexity (Shin, 2019), or the level of repetition (Lee, 2015; Tsai, 2015). These findings suggest that distributing collocations in an appropriate and principled way in pedagogical materials is as important as providing a sufficient number of target items for their effective acquisition.

Since few empirical studies have yet thoroughly examined the recently-developed textbooks and wordlists of the 2015 revised national curriculum of English in Korea, the present study investigates how effectively the textbooks represent collocations in line with the curriculum wordlist. The aim is to assess the textbook data compared to the native language model in the use of collocations. To this end, all nouns are derived from the curriculum wordlist to be used as node words, and the reference collocation list is developed based on the data from the native reference corpus. To model the language input given to the first language (L1) and the second language (L2) learners, distributional patterns of collocations are analyzed in the native reference corpus and the textbook corpus. Specifically, the collocations from each corpus are examined in terms of distributional variables, such as collocation density, repetition, and the association strength. By profiling these variables in each corpus, we hope to make meaningful suggestions for developing guidelines on selecting and presenting collocations in English textbooks.

2. Literature Review

2.1. Concept of collocation

Collocation, the habitual word association, has been thought of as a large and

significant component of native speakers' language production. The concept, however, has yet to be firmly established given its ambiguity as a component with a moderate level of fixedness. According to Nesselhauf (2003), collocations are not entirely fixed but are subject to some degree of "arbitrary restriction" (p. 225) in the choice of components with which they can co-occur. For example, collocations (i.e., *take a photograph* or *take a picture*) are distinguished from other types of highly restricted formulaic expressions (i.e., *sweeten the pill*), or free combinations (i.e., *want a car*) by their semantic/structural unity and fixedness of form. Researchers have noted that collocations constitute "the large and complex middle ground" (Howarth, 1998, p. 42), halfway between "the extreme ends of the spectrum, free combinations, and idioms" (Cowie, 1998, p. 186). However, these theoretical notions may lack the objectivity required to judge whether a word pair belongs to the collocational category (i.e., collocability).

As a way of establishing reliable criteria for collocability, the present study follows a frequency-based approach to collocations. With its theoretical ground credited to Firth (1957), this approach conceptualizes collocation as recurrent word combinations, and the probability of co-occurrence of their constituent words is central to distinguishing collocations from free associations. Firth (1957) argued that the meaning and behavior of each word is, to some degree, determined by its collocates, stating that "a word is known by the company it keeps" (p. 179) and that each word has a different level of "mutual expectancy" (*ibid*, *p*. 181) to the other. This indicates that word choice in natural language is not entirely random, but has some recurrent tendencies.

The frequency-based approach has drawn scholarly interests to statistical modeling of associative relations and has become one of the major trends in corpus-based research of collocations. Collocation is now identified by the level of association strength, which indicates the intensity of the word-pairings, ranging from the strongest association to no association at all (completely independent combinations). This measure quantifies the attraction between words by comparing the observed co-occurrence frequency against the independent frequencies of the constituent words (Bartsch & Evert, 2014). It ranks the probability of two words co-occurring together against the likelihood of them each occurring separately (Schmitt, 2010). By ranking the target items based on the continuum of the probabilistic scale, instead of categorizing them into the dichotomy of collocation/non-collocation, the association scores could provide a more detailed profile of stronger/weaker collocations (Chen, 2019). Over the years, several association measures and statistical

formulas have been developed to best estimate the probability of the co-occurrence (e.g., log-likelihood ratio, t-score, Dice- coefficient, and Mutual Information).

2.2. Assessment of collocation in ELT materials

Studies on collocation have tapped into the significant determiner of successful intake of collocational knowledge. Boers and Lindstromberg (2009), for example, contend that extensive exposure to many different types of collocations could be related to fluent processing, or reception, of collocational input, whereas intensive usage is crucial in durably entrenching the word pairings into the memory, and increasing the fluency of production.

When it comes to the extensive use of collocations, density is one of the most commonly investigated distributional features. It represents how many collocation tokens are presented in the corpus. A higher density indicates that the text contains a large number of collocational pairs, which is believed to develop the learners' sensitivity to native-like collocational patterns (e.g., Durrant & Schmitt, 2010; Ellis, 2002). Interestingly, several research findings have shown that native texts and English language teaching (ELT) materials are indistinguishable in the total frequency of collocations. For example, Koya (2004) found that Verb-Noun collocation token and type counts did not differ between English textbooks used in Japan and history textbooks used in the UK. Similarly, Tsai (2015) found an even higher collocation density in textbooks published in Taiwan than in the native writings.

Repetition is another distributional feature that represents collocational idiomaticity. While density counts the total frequency of collocations in the text, repetition shows the individual frequency of each item. Repetitive exposure to collocations is critical in developing fluency in their production and processing (Pawley & Syder, 1983) and achieving the basic communicative purpose (Wray, 2005). Ellis (1996) suggested that through repeated encounters, sequences of words that were previously independent come to be processed as a single unit or "chunk," and the memory trace of the word association is formulated in the language learners' minds. Some researchers even maintain that the repetitive exposure to the target items may be a more significant determiner of a successful intake than the total number of items. Thus, teaching materials for EFL learners also need to be assessed on the repetition and frequency they give to individual collocation types. Results of previous research, however, vary in such assessments; some researchers observed

that textbooks tend to provide a limited amount of collocation repetition (Tsai, 2015; Lee, 2015), while others report the opposite trend (Shin, 2019).

The last variable to be examined is association strength, which examines the probabilistic nature of collocation and ranks the combinations according to the likelihood that two-component words would co-select each other over the rest. The level of association strength has often been used as criteria to assess collocations. In their statement of the principles behind the selection and the presentation of collocations in the *Oxford Collocations Dictionary for Students of English*, Lea and Runcie (2002) maintained that the most frequent and useful collocations are of "medium-strength," and this "slightly less fixed/fairly open category"(p. 823) should be included in the dictionary. While the significance of medium-strength associations in the native language has been acknowledged by many researchers (Cowie, 1998; Hill, 2000; Howarth, 1998; Lea & Runcie, 2002; Schmid, 2003), it has not been fully addressed in practices of English teaching and learning. In this regard, Howarth (1998) pointed out that learners may be given fewer chances to encounter middle ground restricted collocations and eventually face "challenge in differentiating between combinations that are free and those that are somehow limited in substitutability"(p.42). Likewise, Hill (2000) suggested that learners are generally unsuccessful in using medium strength collocations, and it is necessary to give them good coverage.

Other studies have explored the relationship between the level of association strength and other cognitive or linguistic variables. In psycholinguistic studies, association strength is translated into "predictability," "probability," or "salience" of the formulaic language since stronger collocations allow for processing advantage; the more predictable the association is, the faster the formulaic sequence can be processed (e.g., Conklin & Schmitt, 2012; Durrant & Schmitt, 2010). Meanwhile, in the L2 development research, the association strength tends to be seen as one of the constructs of "lexical complexity," a level of lexical knowledge that reliably predicts language proficiency or text difficulty (e.g., Durrant & Schmitt, 2009; Paquot, 2018, 2019). Generally, the findings suggest that the knowledge of rare collocations is related to higher proficiency. Durrant and Schmitt (2009) found that compared to native writers, L2 writers tend to use more common, weakly associated collocations (e.g., *good example, long way,* and *hard work)*, while native speakers preferred collocations of high association level (e.g., *densely populated, bated breath,* and *preconceived notions*).

Unfortunately, fewer studies have explicitly focused on the association strength of

collocations in the teaching materials. One such attempt was made by Choi and Chon (2012), who analyzed collocation use in 10th grade English textbooks for Korean learners. They found that the textbooks tended to contain more common, weak associations, as shown in the examples of "*good boy, school students, really enjoy*, *volunteer work, good grade, get grades, use cellphone*, etc.". They maintained that free-association collocations with little pedagogical value were over-represented in the textbooks and recommended that more authentic collocations should be incorporated in the materials. A similar pattern was observed by Shin (2019), who investigated the complexity of collocations in textbooks across different countries. Most textbooks were almost identical in using the items with a lower level of complexity, which belongs to the highest 500th frequency band in the Corpus of Contemporary American English (COCA).

Despite the ongoing progress made in collocation studies, there exist some limitations. Firstly, we have little knowledge of the use of collocations, either in the recently-revised English textbooks in Korea or in the authentic native English input. In addition, few studies have considered the national curriculum-wordlist as a base to analyze collocations. English textbooks in Korea are bound to cover a required proportion of words from the curriculum-wordlist, and thus it would be useful to focus on the items which can collocate with the entries from the list. Another limitation of many previous studies lies in the use of the co-occurrence frequency to measure the collocability when the most frequently associated combinations may not necessarily be the authentic, pedagogically meaningful items. Therefore, statistically verifiable association measures need to be employed to identify 'true collocations.' To fill these research gaps, the current study sets out to explore how current middle and high school English textbooks used in Korea present collocations that can be paired with the 2015 revised national curriculum wordlist. Specifically, the study will attempt to answer the following three research questions on distributional patterns of collocations:

RQ1. With how much density do target collocations appear in the textbook and native reference corpora?
RQ2. To what extent are the collocations repeated in the textbook and native reference corpora?
RQ3. What is the association strength of the collocations in the textbook and native reference corpora?

3. Methodology

3.1. Corpora

The present study assesses the use of collocations in the textbook materials with reference to a large-scale native reference corpus to compare the language input that Korean learners and native speakers would be exposed to. As the reference corpus, we used Sketch Engine for Language Learning (SkELL) corpus. This decision was guided by two main considerations. First, the reference corpus should be suitable for language learning purposes, as the current study aims to examine the use of pedagogically meaningful collocations. Second, the corpus should ensure sufficient coverage of the English language to model the language input in a natural environment. SkELL was considered to meet both of these criteria as its compilation reflects pedagogical appropriateness of language, and it contains one billion word tokens from the web-crawled corpora of English, which is now becoming commonplace in recent corpus linguistics.[1] As presented in Table 1, the reference corpus consists of 57,143,446 texts with 1,041,138,575 tokens and 3,602,507 types on average. A total of 122,642,638 tokens of nouns with 1,718 different types were found to match with the curriculum wordlist and thus identified as node words for collocation.

We compiled the textbook corpus with reading sections of the newly published textbooks based on the 2015 revised national curriculum. For comparability with the reference corpus, non-reading sections such as listening scripts and grammar exercises were excluded. The design of the corpus was determined by two criteria. First, the corpus was established to represent the Korean EFL learners' language experience in English classrooms. At the secondary level of education, Korean EFL learners from middle school year 1 (K7) to high school year 1 (K10) use one textbook at each level. While the high school curriculum for students in grades 2 and 3 (K11-12) provides a wider variety of general and elective English courses,

1) Available at Sketch Engine (http://www.sketchengine.eu), a web-based corpus query system, SkELL serves as a large, up-to-date, general-language corpora to meet the various needs for language teaching and learning, research, and lexicography. The SkELL corpus was compiled using an automatic technique which scores sentences according to their appropriateness for pedagogical use. By effectively excluding the highly infrequent words or special terminology, the query outcome from SkELL is considered more desirable for language learners than those from other general reference corpora such as BNC or COCA (Baisa & Suchomel, 2014). The corpus features news script, academic papers, Wikipedia articles, open-source fiction books, webpages, discussion forums, blogs, etc. (Baisa & Suchomel, 2014; Baroni, Kilgarriff, Pomikálek, & Rychlý, 2006)

English 1 and English 2 tend to be the most popular choice at a local school-level curriculum (96% and 58% of high schools in Korea) due to its importance as a main subject of College Scholastic Aptitude Test in Korea (CSAT) (Korean Ministry of Education, 2018). Therefore, we selected 7 out of 12 publishers that provide textbooks of all grade levels of middle school (Middle School English 1-3) and the courses that most high school students are likely to take (Highschool English and English 1·2) and compiled the reading passages from these textbooks into a separate corpus by the publisher (see Appendix for the details). We decided to segment the data by the publisher and not merge them into one corpus in order to model the language input that individual learners may encounter throughout the secondary level of education. As summarized in Table 1, each of the seven textbook corpora contained on average 26,151 words of 4,770 different types. Among these 4,770 word types, 752 items (15.1%) were the nouns from the curriculum-based wordlist, which occurred 3,772 times in each textbook corpus. In other words, learners who use textbooks of each publisher may encounter, on average, 752 types of nouns from the curriculum wordlist a total of 3,772 times throughout the middle and high school curriculum.

Table 1. General profile of corpora

Corpus		Number of texts/books	Single-word		Node word	
			Tokens	Types	Tokens	Types
Reference		57,143,446	1,041,138,575	3,602,507	122,642,638	1,718
Textbook	A	6	27,596	4,533	3,991	755
	B	6	25,684	4,683	3,813	758
	C	6	25,555	4,936	3,697	778
	D	6	20,717	3,783	2,983	652
	E	6	31,682	5,444	4,746	873
	F	6	29,640	5,350	4,461	808
	G	6	22,186	4,664	2,715	640
	Total	42	183,060	14,478	27,087	1,410
	Average (S.D)	6	26,151 (3887.50)	4,770 (557.88)	3,772 (732.4)	752 (82.71)

3.2. Data analysis

To profile curriculum-related collocations in the textbook corpus, we used the curriculum wordlist as node words to form collocations. In particular, nouns from the wordlist were selected for target node words, as nouns are a dominant part of speech carrying most of the semantic component of sentences (Algeo, 2006; Wang & Pei, 2015). With nouns as their node words, three grammatical subtypes of collocates were chosen to be the target of analysis, verbs for VNCs, nouns for NNCs, and adjectives for ANCs. While VNCs have been reported as the primary source of errors for L2 learners (Laufer & Waldman, 2011; Nesselhauf, 2003), the significance of nominal collocations (NNCs and ANCs) in modern English has been acknowledged by many recent studies (Biber & Clark, 2002; Biber & Gray, 2011).

In verifying the collocability of the pairs, we followed the approach taken by Tsai (2015), who first generated a reference collocation list and used the list to statistically verify the collocability of the word pairs in the textbook corpus. Extending the previous research, the current study has compiled the reference collocation database, containing collocate candidates with statistical data on their co-occurrence frequencies and association scores. In generating the reference collocation database, the first step was to query the reference corpus for all possible collocate candidates for each of the 1,718 target nouns as node words within a span of ±4 for VNCs and ±1 for NNCs and ANCs. After retrieving the maximum 1,000 candidates for each node word, all other grammatical categories except verbs, nouns, and adjectives were removed. The remaining candidates were checked against the minimum criteria based on association measures set at 5 for logDice score, 4 for MI-score, and 2 for t-score, with the minimum co-occurrence frequency set at 5.[2] These are more stringent criteria than the commonly cited threshold level for MI of 3 (e.g., Hunston, 2002) in conjunction with a minimum t-score of 2 (Church & Hanks, 1990) or cut-off frequencies set at 3-5 co-occurrences (Church & Hanks, 1990; Stubbs, 1995). In determining the threshold level for logDice score, we followed Frankenberg-Garcia et al. (2019), who noted that collocations with logDice scores below 5 are perceived as free associations rather than collocations. The candidates over these threshold

2) The MI-score has been one of the most popular measures for association strength, but it has limitations in its excessive emphasis on the rarity of combinations by penalizing the high-frequency words. Given the interest of the current study in collocation learning of EFL students, we decided to complement the MI-score with additional association measures. The logDice measure was considered a better alternative to the MI-score since the measure "has a reasonable interpretation, scales well on different corpus size, and the values are in reasonable range" (Rychlý, 2008, p. 7).

levels were verified as collocations, while those below the cut-off points were categorized as free combinations.

Next, using this reference collocation database, collocations from the target textbook corpus were identified. We extracted collocation candidates for 1,718 node words from the textbook corpus. As was done with the reference corpus, the three collocation subtypes were queried within a window of ±4 for VNCs and ±1 for ANCs and NNCs. After extracting all existing combinations from the textbook corpus with a minimum frequency threshold set at 1, they were checked against the reference collocation database and assessed based on the statistical criteria. Once confirmed by the reference corpus data, the word pairs in each textbook series were identified as collocations.

3.3. Measures and tools

In compiling the textbook corpus and searching for collocations, we utilized a web-based corpus analyzing tool, SketchEngine, with Application Programming Interface (API) for efficient data retrieval. SketchEngine provides automatic processing, lemmatization, and part-of-speech tagging of the corpus, which is necessary to specify the grammatical categories and the base form of words to be searched. Additionally, it measures the association strength between the node word and candidates and ranks candidates by the computed association measure. To automize the data retrieval process, a customized program was utilized for API requests. The retrieved data contained the maximum 1,000 collocates for each 1,718 node nouns and the calculated co-occurrence frequency and association measures. The data obtained from each corpus was merged for the additional collocation identification process using the Python script. Finally, SPSS software was used to summarize the data and carry out the test of statistical difference.

3.4. Distributional variables for collocation use

Once collocations were verified, three distributional patterns were quantified: collocation density, repetition, and association strength. Since the textbook corpus had been compiled and analyzed separately by seven publishers, the mean of estimated values of the three variables was calculated.

Following Laufer and Waldman (2011), the collocation density was operationalized as the relative frequencies of collocation tokens per 1000 words to indicate how

many collocations appear within a text of the same length. To measure repetition, the formula "Root type-token ratio (RTTR)" was adopted, which calculates the number of total collocation types against the square root of the total collocation token counts[3] (Guiraud, 1954; Paquot, 2018). To estimate the overall associative strength of collocations, the median of logDice score given to each collocation was computed. Additionally, for finer analysis, the collocations were categorized into the following bands of five association score levels, and the proportions of each band were compared: lower-mid (logDice = 5～6.5), mid (6.5~8), upper-mid (8～9.5), high (9.5～11), and very high strength (over 11). The lowest score band (logDice<5) was deleted from the data, because the items in this band are not considered as collocations.

Lastly, to test the significance of the difference in the collocation density and repetition, non-parametric Chi-square tests were carried out. When comparing the median association scores in each corpus, the Kruskal-Wallis test was run. The correlation between the co-occurrence frequency and the association strength was tested by Kendall's correlation test. Kendall's Tau is a measure of rank correlation (Gries, 2010; Möller, 2017), which calculates how much the ranking orders of the target items differ or agree within the comparing groups.

4. Results

4.1. Token and type counts of collocations

The token and type counts of collocations identified in each corpus were tallied (see Table 2). Collocation token indicates the entire amount of collocations appearing in the corpus. Meanwhile, collocation type counts estimate a total number of different collocations used in the corpus, representing the degree of variety of collocations.

The reference corpus includes the total 14,002,016 collocation candidates (94,666,254 V-N pairs, 19,143,137 N-N pairs, 26,568,942 A-N pairs). The number

3) To operationalize repetition, modified collocation type-to-token ratio (CTTR) is often adopted as it gives the reverse score of how much a writer repeats individual collocations (Durrant, 2008). High CTTR means fewer repetitions made by each collocation type to explain a set number of tokens. Since the measure could penalize the (generally longer) texts with higher collocation token counts, some adjustment was needed to minimize the size effect. To reduce the effect of the denominator, the square root of the total collocation token counts was used as a denominator of RTTR.

of statistically verified collocations is 24,345,386 for VNCs (25.72% of all candidates), 8,498,302 for NNCs (44.39%) and 12,502,402 for ANCs (47.06%), which yields a total of 45,346,090 (32.30%) collocations. The reference corpus exhibits 38,676 type counts for VNC, 22,920 NNC types, and 21,499 ANC types.

In the textbook corpus, 1,571 pairs were verified as collocations by the reference database. The proportion of verified collocations varied on subtypes; the collocations with the highest frequency were VNCs (954), followed by ANCs (425) and NNCs (192) in frequency order. As for type counts, each textbook material contains on average 724 VNC types, 145 NNC types, and 319 ANC types.

Table 2. Token and type counts of collocations

Corpus		VNC		NNC		ANC	
		Tokens	Types	Tokens	Types	Tokens	Types
Reference		24,345,386	38,676	8,498,302	22,920	12,502,402	21,499
Textbook	A	1,026	773	205	141	483	324
	B	981	737	172	143	427	327
	C	979	736	190	155	430	341
	D	723	529	130	104	305	250
	E	1,061	840	252	188	526	403
	F	1,148	854	242	172	501	369
	G	761	598	151	110	303	222
	Average (S.D)	954 (30.89)	724 (156.12)	192 (13.8)	145 (45.1)	425 (20.6)	319 (90.0)
	Total	6,679	5,067	1,342	1,013	2,975	2,236

4.2. Collocation density

Collocation density estimates the proportion of statistically verified collocations in relation to text length (total word counts). Table 3 shows the total number of collocations per 1,000 words in both corpora.

Table 3. Collocation density

	Subtype	Textbook corpus	Reference corpus
Collocation token counts per 1,000 words	VNC	36.49	23.38
	NNC	7.33	8.16
	ANC	16.25	12.01
	Total	60.07	43.55

Inspecting the textbook data more closely, VNCs (36.49) account for the largest proportion of the collocation token counts per 1,000 words, followed by ANCs (16.25) and NNCs (7.33). A total collocation token count is 60.07 per 1,000 words, which is nearly 17 more tokens than in the reference corpus of 43.55 collocations in every 1,000 words (23.38 VNCs, 12.01 ANCs, and 8.16 NNCs). This result indicates that textbook readers would generally encounter one collocation in every 16~17 word counts, which is more frequent than for the reference corpus in the same length (every 22~23 word counts), with the exception of NNCs. While VNCs and ANCs tend to appear more frequently in the textbook materials than in the reference corpus, NNCs appear to be underrepresented compared to the native norms. Statistical significance in the difference was found for VNCs and ANCs (VNCs: χ^2=185.075, *df*=1, *p*<.001; NNCs: χ^2=2.141, *df*=1, *p*=.143; ANCs: χ^2=38.534, *df*=1, *p*<.001).

4.3. Repetition rate

The median number of repetitions for each collocation type shows that all collocation subtypes appear almost only once in the English textbooks.[4] Looking into the detailed distribution (Table 4), 80.71% of VNC types appeared only once throughout the textbook materials, whereas 0.70% of VNCs are presented more than six times. Similarly, the proportion of items repeating over six times remains limited to 0.99% of NNCs and 1.25% of ANCs. Moreover, nearly 82.28% of NNCs and 82.22% of ANCs are never revisited. For example, the co-occurrence frequency of the most frequent collocates of the node word "money," such as *raise* (2), *send* (1.7), *spend* (1.6), *make* (1.6), *save* (1.5), *give* (1.3), and *get* (1.3), were two at most. The rest of the collocates (e.g., *pay, earn, lose, borrow*, *put*) occurred only once. Likewise,

4) Median frequencies are used here because the distribution is not normal.

such collocates of "idea" as *creative* (2), *good (*1.0), *bad* (1.0), *new* (1.0), *great* (1.0), *innovative* (1.0), *general* (1.0), and *brilliant* (1.0) were rarely repeated. In contrast, those collocates have shown distinctively high co-occurrence frequency in the reference corpus. This result indicates that Korean learners are likely to encounter more than 80% of the collocations only once throughout the middle and high school textbook materials, which is far below the number (10 encounters) required to acquire collocations (Webb, Newton, & Chang, 2013).

Table 4. Distribution of collocation (types) according to the number of repetitions in the textbook corpus

Subtype	1	2 to 5	6 to 10	over 10	Total	Mdn	M (SD)
VNC	585* (80.71%)	134 (18.59%)	5 (0.65%)	0 (0.05%)	724	1.0	1.31 (0.852)
NNC	119 (82.28%)	24 (16.73%)	1 (0.76%)	0 (0.23%)	145	1.0	1.32 (1.025)
ANC	263 (82.22%)	53 (16.53%)	3 (1.09%)	1 (0.16%)	320	1.0	1.33 (1.029)

* Average number of collocation types in textbooks from each publisher.

Next, Table 5 compares the repetition rate of the textbook corpus with the reference baseline. With RTTR, the modified type-token ratio gives a reverse score for the number of repetitions made by each collocation type. As indicated by higher RTTR scores (23.43 for VNCs, 10.50 for NNCs, 15.49 for ANCs), collocations in the textbooks are less recursive than their equivalents in the reference corpus (7.84 for VNCs, 7.86 for NNCs, and 6.08 ANCs). Statistical significance of difference was confirmed in RTTR scores of collocation between the two corpora (VNCs: χ^2= 71.279, *df*=1, p<.001; NNCs: χ^2=38.944, *df*=1, p<.01; ANCs: χ^2= 32.892, *df*=1, p<.001).

Table 5. Repetition rate by RTTR*

Subtypes	Textbook corpus	Reference corpus
VNC	23.43	7.84
NNC	10.45	7.86
ANC	15.49	6.08

RTTR* = Collocation type/$\sqrt{}$Collocation token.

4.4. Association strength

The overall association strength of collocations in the textbook and reference corpora was calculated by the logDice score (see Table 6). A higher logDice score indicates a higher probability that component words would be associated strongly enough to be a collocation. Association strength of collocations in the textbook corpus (6.58 for VNCs, 6.94 for NNCs, and 7.11 for ANCs) was higher than their native equivalents, which yield a median score of 5.70 for VNCs, 5.85 for NNCs, and 5.85 for ANCs. The independent Mann-Whitney U test confirmed the statistical significance of the difference between the association measures of the two corpora for VNCs (U= 52,734,541.0, p<.001), NNCs (U=6,402,910.0, p<.001), and ANCs (U= 11,820,027.0, p<.001). Data from the table points to seemingly counterintuitive results: Collocations in the textbooks tend to be those associated with higher probability, whereas the reference corpus contains relatively weaker associations.

Table 6. Median logDice score of collocations (type)

Corpus	VNCs	NNCs	ANCs
Textbook	6.58	6.94	7.11
Reference	5.70	5.85	5.85

Next, to provide a detailed comparison, Table 7 illustrates the proportion of collocations in each range band of logDice scores in the two corpora. Interestingly, the textbook corpus relies to a greater extent on the items at the mid (6.5-8.0) to upper-mid (8.0-9.5) level of association strength than the reference data in which a majority of collocation types have lower-mid (5-6.5) association strength.

To elaborate, the reference corpus presents a larger body of collocation of all subtypes (VNCs 80.1%, NNCs 72.3%, ANCs 72.9%) at the lower-mid level of logDice scores ranging from 5 to 6.5, than the textbook corpus (VNCs 46.8%, NNCs 37.8%, ANCs 35.5%). Meanwhile, more than half of the collocation types in the textbook corpus belong to the mid to high level of logDice scores over 6.5 (VNCs 53.2%, NNCs 62.2%, ANCs 64.5%), which is disproportionately larger than their native counterparts (VNCs 19.5%, NNCs 27.74%, ANCs 27.1%).

Table 7. Distribution of collocations (type) by association strength

Corpus	Subtype	Lower-mid 5-6.5	Mid 6.5-8.0	Upper-mid 8.0-9.5	High 9.5-11.0	Very high over 11	Total
Textbook	VNC	339 (46.8%)	268 (37.1%)	96 (13.2%)	19 (2.6%)	2 (0.3%)	724
	NNC	55 (37.8%)	53 (36.9%)	29 (20.3%)	6 (3.8%)	2 (1.1%)	145
	ANC	113 (35.5%)	107 (33.4%)	77 (24.2%)	20 (6.2%)	2 (0.8%)	319
Reference	VNC	30,997 (80.1%)	6,599 (17.1%)	982 (2.5%)	92 (0.2%)	5 (0.0%)	38,675
	NNC	16,573 (72.3%)	4,979 (21.7%)	1,184 (5.2%)	168 (0.7%)	16 (0.1%)	22,920
	ANC	15,680 (72.9%)	4,639 (21.6%)	1,018 (4.7%)	140 (0.7%)	21 (0.1%)	21,498

To provide a detailed description of the different collocational strengths, Tables 8 and 9 exemplify the collocates associated with the node words "*money*" and "*idea,*" respectively. The textbook corpus presents a narrower range of collocations, most of which are relatively strong, common associations ranked at the top of the logDice

Table 8. VNC collocates for "money"

logDice score	Textbook corpus	Reference corpus
High over 9.5	spend, save,	spend, save
Upper-mid 8.0-9.5	raise, pay, earn, make, borrow	raise, pay, borrow, earn, invest, make
Mid 6.5-8.0	buy, lose, put, get, donate, give, cost, collect, need, steal, receive, send	buy, receive, lose, lend, waste, send, put, get, donate, give, owe, cost, collect, need, steal
	want, go, help, ask	want, ask, go, offer, keep, sell, help, deposit, throw, transfer, fund, demand
Lower-mid 5-6.5	allocate	withdraw, generate, purchase, launder, accept, hand, print, finance, pour, refund, allocate, repay, manage, charge, obtain, distribute, loan, refuse, contribute, guarantee, flow, bet, extort, count, recover

scale (e.g., *spend, raise, save, earn* for the node word "money," and *good, bad, new, great* for the node word "idea"). Meanwhile, they do not present as many lower-mid strength items below logDice score 6 (e.g., *offer, loan, obtain* for "*money*," *original, bright, interesting* for "*idea*") as reference corpus does.

Table 9. ANC collocates for "idea"

logDice score	Textbook corpus	Reference corpus
Upper-mid 8.0-9.5	good	good
Mid 6.5-8.0	bad, new, great	bad, new, basic, great
Lower-mid 5-6.5	innovative, general, creative, brilliant	whole, original, very, innovative, general, clear, creative, abstract, brilliant, interesting, fresh, bright

We further examined if stronger collocations appear more repetitively by measuring the extent to which association strength corresponds to the number of repetitions (co-occurrence frequency) given to the item. The test results of the correlation between the two variables, association strength, and co-occurrence frequency, are presented in Table 10. The table reveals that the rank order by co-occurrence frequency and association strength of each collocation is less concordant in the textbook corpus than in the reference corpus. In other words, collocation use indicated by the frequency of the target items in the textbooks is less likely to match the level of association strength, and vice versa.

Table 10. Correlation between association strength and co-occurrence frequency of collocations (type)

Subtype	Index	Textbook corpus			Reference corpus		
		Mdn	Corr.	N	Mdn	Corr.	N
VNC	logDice	6.579	.192***	5,067	5.700	.321***	38,676
	Co-occurrence	1			277		
NNC	logDice	6.935	.069**	1,013	5.852	.316***	22,920
	Co-occurrence	1			153		
ANC	logDice	7.110	.184***	2,236	6.110	.338***	21,499
	Co-occurrence	1			581.53		

**p<.01, ***p<.001

To explain the details concerning the textbook corpus, the correlation coefficient between the logDice score and co-occurrence frequency of individual VNCs is τ =.192 (p<.001). The correlation of two variables in the textbook is lower than in the reference corpus, τ=.321 (p<.001), which means that the frequency of collocations in the textbook corpus is less likely to reflect the association strength than in the reference corpus. In other words, the extent of collocational input in the textbook materials has little relation to the association strength. For instance, strongly associated collocates of the head-noun *"money"* (e.g., *spend, raise, save, pay, make*) tend to occur only once or never appear in most of the textbook materials, of which frequencies are not distinguishable from that of weaker associations (e.g., *allocate, ask, cost, steal*). In contrast, in the reference corpus, the former group occurs almost six times more frequently (median co-occurrence frequency=10,714) than the latter one does (1,644). The potential problem indicated by this lower correlation between the two variables in the textbook materials will be discussed in the following section.

5. Discussion and Conclusion

The present study set out to investigate the use of collocations in the recently developed middle and high school English textbooks based on the 2015 revised national curriculum. The analysis of the extensive and intensive use of collocations derived from the curriculum wordlist has revealed a higher density and association strength, but less repetition in comparison to the reference corpus. As for collocation density, the result shows that the textbook corpus presents a significantly larger body of VNCs and ANCs than the reference corpus, indicating that Korean learners would be exposed to a relatively higher proportion of collocational input from the textbook materials. This finding is in line with previous studies, which have demonstrated that ELT materials exhibited a relatively denser distribution of collocations than NS productions (e.g., Koya, 2004; Tsai, 2015; Shin, 2019).

While textbook materials have shown some advantages over natural input with their superior coverage of collocations, it remains inconclusive whether textbooks outdo the natural setting as an input. The higher collocation density, in fact, appears to compromise the intensity of its use, which is also central to defining and learning this lexical category. The current data where all three subtypes of collocation are

found to be markedly less repetitive in textbook materials than in the reference corpus reveals that collocational formulaicity may not be fully represented in the materials for EFL learners. Also, the median co-occurrence frequency of one or two exhibited in the textbook corpus is much lower than the level of repetition (at least eight co-occurrences) recommended by previous research (Webb et al., 2013), whereas in native input, collocations are highly recurrent phenomena. Given the significant role that repetition plays in the consolidation of learners' collocational knowledge in their long-term memory (Ellis, 2001, 2002; Wolter & Gyllstad, 2013), the present finding highlights the need for sufficient repetition of collocations in the pedagogical materials. Without repetitive exposure through language input, the extensive coverage of collocations in the materials alone may not ensure the efficient learning of the units. In this regard, current data seems to support Koprowski's (2005) statement, "more is less" (p. 329), suggesting that higher collocation density may not always be advantageous and that the extensive coverage of a larger number of collocations in textbook materials may compromise its idiomatic tendency as 'habitual' co-occurrence of words.

On the other hand, the present study may also contribute to enhancing knowledge of the optimal level of association strength for instructing EFL learners. The association strength of collocations in the textbook corpus was found to be generally higher, while collocations at low-mid level association strength were relatively scarce in comparison to native reference data. The collocational repertoire consisting mostly of stronger collocations implies that learners may be given fewer chances to encounter mid-level strength collocations, contrary to their prominence in native data (Durrant & Schmitt, 2010; Hill, 2000; Howarth, 1998; Lea & Runcie, 2002; J. K. Lee, 2009). Hence, special attention may also need to be paid to less than typical, collocations with lower-mid level association strength to foster learners' sensitivity to a broader range of associative relationships.

The last notable finding was the mismatch between the two aforementioned variables, i.e., co-occurrence frequency of individual collocations and their association strength, in the textbook corpus. With the amount of exposure that reliably predicts the level of collocability, collocational input in native English seems to enable native speakers to develop their intuition of the associative relationship. By contrast, a distinctively weaker correlation between the frequency data and association strength in the textbook corpus suggests that Korean EFL learners are less likely to have such language experiences. By better aligning the frequency of target collocations with their association strengths, learners may be given more

frequency cues and develop the collocational sensitivity to predict 'true collocations.'

The current study is not without limitations, which point to directions for future research. The use of large-sized reference data as a comparison group to textbook corpora has both its benefits and shortcomings. While it provides a look into the native English usage with maximum representativeness, the effects from the different sample sizes were unavoidable issues. Although we attempted to use various mathematic formulae to reduce the sample size effect, they may still be insufficient to standardize the corpus size accurately. Another limitation regarding data collection concerns that the textbook corpus analyzed in this study only contains the reading passages. To include listening scripts in the future analysis may provide a broader picture of collocation use in the textbook materials, including the difference between oral and written input. Notwithstanding these limitations, this study offers useful insights into the design of the lexical syllabus beyond the single word level and the pedagogical issues concerning the expected advantages and disadvantages of collocation learning through the current English textbooks.

References

Algeo, J. (2006). *British or American English?: A handbook of word and grammar patterns*. Cambridge. England: Cambridge University Press.

Baisa, V., & Suchomel, V. (2014). SkELL: Web interface for English language learning. In Horák, A. & Rychlý, P (Eds.), *Proceedings of recent advances in slavonic natural language processing* (pp. 63–70). Karlova Studánka, Czech Republic: Tribun EU.

Baroni, M., Kilgarriff, A., Pomikálek, J., & Rychlý, P. (2006). WebBootCaT: a web tool for instant corpora. In E. Corino, et al. (Eds.), *Vol. 1. Proceeding of the EuraLex conference* (Vol. 1, pp.123–132). Turin, Italy: Alessandria.

Bartsch, S., & Evert, S. (2014). Towards a Firthian notion of collocation. *Vernetzungsstrategien, Zugriffsstrukturen und Automatisch Ermittelte Angaben in Internetwörterbüchern, 2,* 48–61.

Biber, D., & Clark, V. (2002). Historical shifts in modfication patterns with complex noun phrase structures. In T. Fanego, J. Pe´rez-Guerra, & M. J. Lo´pez-Couso (Eds.), *English historical syntax and morphology* (pp. 43-66). Amsterdam: John Benjamins.

Biber, D., & Gray, B. (2011). Grammatical change in the noun phrase: The influence of written language use. *English Language and Linguistics*, *15*(2), 223–250.

Boers, F., & Lindstromberg, S. (2009). *Optimizing a lexical approach to instructed second language acquisition.* Basingstoke, UK: Palgrave Macmillan.

Chen, W. (2019). Profiling collocations in EFL writing of Chinese tertiary learners. *RELC*

Journal, *50*(1), 53–70.

Choi, H. Y., & Chon, Y. V. (2012). A corpus-based analysis of collocations in tenth-grade high school English textbooks. *Multimedia Assisted Language Learning*, *15*(2), 41–73.

Conklin, K., & Schmitt, N. (2012). The processing of formulaic language. *Annual Review of Applied Linguistics*, *32*, 45–61.

Cowie, A. P. (1992). Multiword lexical units and communicative language teaching. In P.J.L. Arnaud, & H. Béjoint (Eds.), *Vocabulary and applied linguistics* (pp. 1–12). London: Palgrave Macmillan.

Cowie, A. P. (1998). *Phraseology: Theory, analysis, and applications*. Oxford, England: Oxford University Press.

Durrant, P. L. (2008). *High frequency collocations and second language learning* (Doctoral dissertation). University of Nottingham, United Kingdom.

Durrant, P. L., & Schmitt, N. (2009). To what extent do native and non-native writers make use of collocations? *International Review of Applied Linguistics in Language Teaching*, *47*(2), 157–177.

Durrant, P. L., & Schmitt, N. (2010). Adult learners' retention of collocations from exposure. *Second Language Research*, *26*(2), 163–188.

Ellis, N. C. (1996). Sequencing in SLA: Phonological memory, chunking, and points of order. *Studies in Second Language Acquisition*, *18*(1), 91–126.

Ellis, N. C. (2001). Memory for language. In P. Robinson (Ed.), *Cognition and second language Instruction* (pp. 33–68). Cambridge, England: Cambridge University.

Ellis, N. C. (2002). Frequency effects in language processing: A review with implications for theories of implicit and explicit language acquisition. *Studies in Second Language Acquisition*, *24*(2), 143–188.

Erman, B., & Warren, B. (2000). The idiom principle and the open choice principle. *Text-Interdisciplinary Journal for the Study of Discourse*, *20*(1), 29–62.

Firth, J. R. (1957). *Papers in Linguistics 1934-1951.* London: Oxford University Press.

Frankenberg-Garcia, A., Lew, R., Roberts, J. C., Rees, G. P., & Sharma, N. (2019). Developing a writing assistant to help EAP writers with collocations in real time. *ReCALL*, *31*(1), 23–39.

Gries, S. T. (2010). Useful statistics for corpus linguistics. In A. Sánchez, M. Almela, (Eds)., *A mosaic of corpus linguistics: selected approaches* (pp. 269-291), Frankfurt, Germany: Peter Lang.

Guiraud, P. (1954). *Les caractères statistiques du vocabulaire: essai de méthodologie*. Paris, France: Presses universitaires de France.

Harmer, J., & Rossner, R. (1997). *More than words: Vocabulary for upper intermediate to advanced students*. Essex: Addison Wesley Longman.

Hill, J. (2000). Revising priorities: From grammatical failure to collocational success. In M.Lewis (Ed.), *Teaching collocation: Further developments in the lexical approach*, (pp. 47–69). Hove, England: Language Teaching Publications.

Hoey, M. (2005). *Lexical priming: A new theory of words and language*. London: Routledge.

Howarth, P. (1998). Phraseology and second language proficiency. *Applied Linguistics*, *19*(1), 24–44.

Hunston, S. (2002). *Corpora in applied linguistics*. Cambridge. England: Cambridge University Press.

Kim, N. B. (2004). Collocational analysis of Korean high school English textbooks and suggestions for collocation instruction. *English Language & Literature Teaching*, *10*(3), 41–66.

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

Korean Ministry of Education. (2018). *An inquiry into the organization and operation of optional subjects following the 2015 revised curriculum system* (No. 11-1342000-000359-01). Retrieved from http://www.prism.go.kr/homepage/entire/retrieveEntireDetail.do;jsessionid=BA5DD2A074E12CDE35BC8FB315538EB3.node02?cond_research_name=&cond_research_start_date=&cond_research_end_date=&research_id=1342000-201900002&pageIndex=3&leftMenuLevel=160

Church, K. W., & Hanks, P. (1990). Word association norms, mutual information, and lexicography. *Computational Linguistics*, *16*(1), 22–29.

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

Laufer, B., & Waldman, T. (2011). Verb-noun collocations in second language writing: A corpus analysis of learners' English. *Language Learning*, *61*(2), 647–672.

Lea, D., & Runcie, M. (2002). Blunt instruments and fine distinctions: A collocations dictionary for students of English. In A. Braasch, & C. Povlsen (Eds.), *Proceedings of the tenth EURALEX International Congress* (pp. 819–829). Copenhagen, Denmark.

Lee, J. K. (2009). Analysis of collocability of word list in the revised national curriculum. *Studies in Modern Grammar*, *58*, 249–271.

Lee, J. K. (2015). The repetition of chunks in Korean middle school English textbooks. *English Language Teaching*, *8*(10), 60–75.

Lee, M. B., & Shin, D. K. (2015). Development of the Korean basic English word list of the 2015 revised national curriculum of English. *Journal of the Korea English Education Society*, *14*(4), 115–134.

McCarthy, M., & O'Dell, F. (1994). *English vocabulary in use: 100 units of vocabulary reference and practice*. Cambridge, England: Cambridge University Press.

Möller, V. (2017). A statistical analysis of learner corpus data, experimental data and individual differences: Monofactorial vs. multifactorial approaches. In P. J. de Haan, C. M. de Vries, & S. V. Vuuren (Eds.), *Language, learners and levels: Progression and variation* (pp. 409–439). Louvain-la-Neuve: Presses Universitaires de Louvain.

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

Nesselhauf, N. (2003). The use of collocations by advanced learners of English and some

implications for teaching. *Applied Linguistics*, *24*(2), 223–242.
Paquot, M. (2018). Phraseological competence: A missing component in university entrance language tests? Insights from a study of EFL Learners' use of statistical collocations. *Language Assessment Quarterly*, *15*(1), 29–43.
Paquot, M. (2019). The phraseological dimension in interlanguage complexity research. *Second Language Research*, *35*(1), 121–145.
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). London: Longman.
Rychlý, P. (2008). A lexicographer-friendly association score. In P. Sojka, & A. Horák (Eds.), *Proceedings of recent advances in Slavonic natural language processing* (pp. 6–9). Brno, Czech Republic: Masaryk University.
Schmid, H. J. (2003). Collocation: Hard to pin down, but bloody useful. *Zeitschrift Fur Anglistik Und Amerikanistik*, *51*(3), 235–258.
Schmitt, N (Ed.) (2010). *Researching vocabulary: A vocabulary research manual*. Basingstoke, England: Palgrave Macmillan.
Segalowitz, N. (2010). *Cognitive bases of second language fluency*. London: Routledge.
Shin, D. K. (2019). A comparative study on the use of single words and collocations in domestic and overseas. *Journal of Language Science, 26*(4), 87–108.
Sinclair, J. (1991). *Corpus, concordance, collocation*. Oxford, England: Oxford University Press.
Stubbs, M. (1995). Collocations and semantic profiles: On the cause of the trouble with quantitative studies. *Functions of Language*, *2*(1), 23–55.
Stubbs, M. (2001). *Words and phrases: Corpus studies of lexical semantics*. Oxford, England: Blackwell Publishers.
Tsai, K. J. (2015). Profiling the collocation use in ELT textbooks and learner writing. *Language Teaching Research*, *19*(6), 723–740.
Wang, L., & Pei, F. (2015). Types and features of noun phrase in Chinese scholars' abstracts. *International Journal of English Linguistics*, *5*(6), 84-94.
Webb, S., Newton, J., & Chang, A. (2013). Incidental learning of collocation. *Language Learning*, *63*(1), 91–120.
Wolter, B., & Gyllstad, H. (2013). Frequency of input and L2 collocational processing. *Studies in Second Language Acquisition*, *35*(3), 451–482.
Wray, A. (2005). *Formulaic language and the lexicon*. Cambridge, England: Cambridge University Press.

Young Shin Kim
English Teacher
Hansan Middle School
251, Pungseong-ro, Gangdong-gu Seoul 05371, Korea

E-mail: wooltrakim@sen.go.kr
Sun-Young Oh
Professor
Department of English Language Education
Seoul National University
1 Gwanak-ro, Gwanak-gu Seoul, 08826, Korea
E-mail: sunoh@snu.ac.kr

Received: October 30, 2020
Revised version received: December 21, 2020
Accepted: December 30, 2020

Appendix

The English textbooks analyzed in the study

Publisher	First Author	Grade Levels
Chunjae	Lee, J.	Middle School English 1, 2, 3 (2017-2019) High School English (2017) High School English 1 (2017) High School English 2 (2018)
Darakwon	Kang, Y.	
Jihaksa	Min, C.	
Kumsung	Choi, I.	
Visang	Kim, J.	
	Hong, M.	
YBM	Park, J.	
	Han, S.	High School English (2017) High School English 1 (2017) High School English 2 (2018)
	Song, M.	Middle School English 1, 2, 3 (2017-2019)