Curriculum and Teaching Methodology (2022)
Clausius Scientific Press, Canada

DOI: 10.23977/curtm.2022.050701
ISSN 2616-2261 Vol. 5 Num. 7

A Corpus-based Analysis on Distributional Patterns of Collocations in Chinese High School English Textbooks

Liu Nuan[1], Liu Xiumin[2,*]

[1]School of English Language and Literature, Kookmin University, Seoul, 02707, Republic of Korea
[2]School of Foreign Language, Longyan University, Longyan, Fujian, 364000, China
**Corresponding author*

Keywords: Distributional patterns, collocations, Chinese high school English textbooks

Abstract: The paper explores the distributional patterns of collocations identified in Chinese high school textbooks in comparison with collocations found in the native reference corpus. Six series of Chinese high school English textbooks from the six publishers were compiled as the textbook corpus. The British National Corpus (BNC) was selected as the native norm. Three collocation types, adjective + noun collocation (ANC), verb + noun collocation (VNC) and noun + noun collocation (NNC), were investigated in the paper. The distribution of the collocations was compared between the two corpora in terms of four statistical measures: density, diversity, repetition and association strength. The results showed that whereas the diversity of collocations was higher in the native reference corpus than in the textbook corpus, the other measures showed the opposite pattern. This may partly be due to the pedagogical nature of the textbooks. Moreover, the findings revealed that the textbooks represent a comparable number of VNCs, over-represent ANCs and under-represent NNCs in comparison with the native reference corpus. The findings suggest that textbook authors take account of incorporating more diverse collocations and more NNCs to model more native-like texts in the teaching materials targeting higher-grade students.

1. Introduction

Corpus linguistics provides a powerful tool for exploring a variety of language-related issues, ranging from discovering patterns of actual language use by theoretical linguists and developing teaching materials by language teaching professionals (Reppen & Simpson-Vlach, 2010). One of the characteristics of corpus-based analyses of language is that "it utilizes a large and principled collection of natural texts, known as a 'corpus', as the basis for analysis" (Reppen & Simpson-Vlach, 2010, p.91).

Collocations as one type of formulaic language refer to the sequences or sets of words that commonly co-occur with greater probability than random chance (Reppen & Simpson-Valch, 2010). The statistic conceptualization of collocations naturally leads to the use of corpora for the identification and analysis of collocations. In light of the important role that formulaic language plays in language acquisition and use (Granger, 1998; Lewis, 1993; Nattinger & Decarrico, 1992), collocations have attracted increasing attention in corpus-based studies of language (Gablasova,

Brezina, & McEnery, 2017). The resulting body of corpus-based research into collocations provides valuable insights into not only the characteristics of collocations per se but also the vital role that collocations play in language learning and use.

Given the importance of collocational knowledge in L2 development, Tsai (2014), for example, examined verb + noun collocations in the EFL textbooks used in Taiwan and the Chinese learners' writing in comparison with the native speakers' essays (Louvain Corpus of Native English Essays). It is found that the textbooks were comparable to the native speakers' essays regarding collocational density and diversity, but the former did not repeat collocations as frequently as the latter. There are few studies that specifically targeted the revised Chinese high school textbooks and wordlists of the 2017 revised Chinese national curriculum of English. The present study aims to investigate whether collocations used in Chinese high school English textbooks reflect the native-like patterns of collocations. It specifically attempts to explore the distributional patterns of collocations in Chinese high school English textbooks.

2. Literature review

Frequency-based approach is one of the theoretical approaches to collocations. In the frequency-based approach, collocations are identified by quantitative criteria such as simple frequency of occurrence and statistics that measure the strength of the tendency for individual words to co-occur in a collocation. Collocations are sequences of words that "have a statistical tendency to co-occur" in a corpus (Durrant, 2014, p. 446). Firth, Halliday, and Sinclair are the representatives of this tradition. Firth (1957, p.11) brought the term "collocation" into prominence and emphasized that collocations played a crucial role in establishing the meaning of a word, which could be summarized by his well-known statement: "You shall know a word by the company it keeps". According to Firth (1957, p.12), habitual collocation is a type of "mutual expectancy" between words. To put it differently, when one word is found, the other is likely to be found. This idea of mutual expectancy underlies the conceptualization of collocation as "the relationship a lexical item has with items that appear with greater than random probability in its textual context" (Hoey, 1991, p.7).

The density, diversity, repetition, and association strength have been used to describe the distribution of collocations in a corpus. The four measures have been extensively used to study the distributional characteristics of collocations in a text. For example, Tsai (2015) used density, diversity, and repetition rate to investigate the verb + noun collocations in EFL textbooks in Taiwan and Chinese EFL learners' writings in comparison with the native reference corpus. It was found that textbooks were comparable to the native writings in terms of density. But the textbooks under-represented the verb + noun collocation types in comparison to the native corpus. Moreover, the repetition of the majority of collocation types in the textbooks was deemed to be insufficient enough for L2 learners to acquire collocations (Tsai, 2005).

Likewise, Kim (2020) examined VNCs, ANCs, and NNCs used in Korean EFL textbooks compared with the native reference corpus in terms of density, diversity, repetition, and association strengths. Findings showed that the textbook corpus, by and large, exhibited higher collocation density and diversity, less repetition, and stronger association strength. Kim suggested that collocation density and diversity are at the cost of less repetition, and teachers should prepare supplemental materials to make up for less repetition of collocations in the textbooks. Moreover, more highly probable collocations were found in the textbooks in comparison to the native corpus. Kim (2020) thus suggested "incorporating less-than-typical collocations into the materials".

The two studies successfully showed the detailed distributional characteristics of collocations in L2 corpora using the four measures. The findings provide useful implications for curriculum

developers and textbook writers.

3. Methodology

3.1 Corpora

The British National Corpus (BNC), which is one of the largest English corpora, was chosen as the native reference corpus in the present dissertation. It contains 100 million words and has been considered one of the most representative corpora of general English currently available. The corpus consists of 90% written language and 10% spoken language from a wide range of sources and covers a wide cross-section of British English from the late 20th century. The written texts in the BNC are extracted from "regional and national newspapers, specialist periodicals and journals for all ages and interests, academic books and popular fiction, published and unpublished letters and memoranda, school and university essays" (http://www.natcorp.ox.ac.uk/corpus/), with a total of 86,097,791 words.

The BNC has been referred to by many researchers to determine the frequency and MI of NNS (non-native speakers) and NS (native speakers) collocations (Durrant & Schmitt, 2009; Siyanova & Schmitt, 2008). It is assumed that combinations which are frequently used in the BNC are representative of common usage in English (Durrant & Schmitt, 2009).

A textbook corpus consisting of six series of textbooks was compiled for collocation analysis. These EFL textbooks followed the Chinese 2017 revised national curriculum and were published by six publishers. They have been widely used in high schools in China. Most series of textbooks contain seven or eight volumes designed for the six terms of the high school English curriculum while some series of textbooks comprise six volumes. For the sake of convenience, the textbook series, based on the publishers, are hereafter referred to as BU (Bejing Normal University Press), CU (Chongqing University Press), YL (Yilin Press), SH (Shanghai Education Press), PE (People's Education Press) and FL (Foreign Language Teaching Research Press).

The electronic textbooks were either downloaded from the official website of the publishers or purchased online. The Optical Character Recognition (OCR) feature in Google Drive was used to convert the PDF version of these textbooks into plain text. Since errors were inevitable in the process of automatic converting, errors were corrected by a manual check.

In Kim's (2020) study, reading passages from middle and high school English textbooks were included. In a similar vein, all the reading passages in the main textbooks and workbooks were included in the present study. In Tsai's (2015) study, instructions, and vocabulary exercises were included in the textbook corpus because it was believed that all collocations in the textbooks, regardless of the sections where they occur should be examined given the fact that textbooks serve as the major source of the language input for EFL learners. However, in the present study, instructions and exercises were excluded in that instructions only consist of limited lexical items that occur frequently and are massively repeated in the textbook, and exercises are usually derived from the previous reading materials.

The textbooks under investigation were published based on the 2017 revised national curriculum. As Table 1 shows, the textbook corpora were compiled into six separate sub-corpora by the publisher. The single word refers to the words appearing in the textbooks, while the node word refers to the nouns in the wordlist specified by the national curriculum. Each of the six textbook sub-corpora included on average 49,410 tokens of 6,038 different types. Among these 6,038 word types, an average of 1,210 (20%) was found to match with the nouns from the curriculum, occurring 8,686 times in each textbook corpus. To put it another way, each series of textbooks may present students with, on average, 1,210 noun types from the curriculum, on average 8,686 tokens in total throughout the high school curriculum.

Table 1 General Profile of Chinese High School English Textbook Corpus

Corpus	Number of books		Single-word		Node word	
			Tokens	Types	Tokens	Types
Reference			86,097,791	534,643	5,398,434	1,784
Textbook	BU	7	53,938	6,329	9,708	1,252
	CU	8	59,631	6,593	10,608	1,253
	FL	7	43,718	5,904	7,767	1,217
	PE	7	63,956	6,910	12,061	1,350
	SH	6	22,576	3,985	3,973	913
	YL	6	52,641	6,507	7,996	1,273
Total		41	296,460	36,228	52,113	7,258
Average (S.D)		7	49,410	6,038	8,686	1,210
			14,829	1,059	2,816	152

It is noted that the textbook corpus is a collection of six sub-corpora each of which is hypothesized to constitute the language input to learners. The distribution statistics of collocations in the textbook corpus (i.e., density, diversity, repetition rate and association strength) are averages of those from each sub-corpus consisting of the textbooks from a particular publisher.

3.2 Measures and tools

3.2.1 Statistical criteria

The logDice measure reveals the extent of collocational bonding. It does not favor frequent word combinations nor downgrade the high-frequency combinations. Therefore, the logDice score and frequency of cooccurrence were employed in the present study to identify collocations from the native reference corpus.

Frankenberg-Garcia et al. (2018) found that a threshold of logDice score greater than or equal to 5 works well for identifying lexical collocations. The logDice with a minimum score of 5 is therefore used in this study (Frankenberg-Garcia et al., 2018; Kim, 2020). And the minimum frequency of occurrence for a word combination to be taken to be a collocation was set at 5. The word combinations over both of these cut-off points were identified as collocations, while those below the threshold levels were excluded from the analysis.

3.2.2 Distributional measures

To compare the overall distributional patterns of collocations between the corpora under analysis in this study, four aspects of distribution were analyzed: collocation density, collocation diversity, repetition rate, and association strength. Each of these distributional measures is briefly reviewed below, together with a specific statistic used to quantify it in this study.

Firstly, collocation density reflects in a relative way how many tokens constitute collocations in a corpus. Since the corpora were of different sizes, the collocation density was operationalized as the relative frequencies of collocation tokens per 1,000 words to indicate how many collocations occur within a text of the same length (Laufer & Waldman, 2011).

Secondly, collocation diversity is used to measure the number of collocation types in the target corpus. To minimize the corpus size effect, the formula “collocation diversity = the number of collocation types/the square root of word tokens” was used (Kim, 2020).

Thirdly, the repetition rate is calculated to measure how many times the individual collocation types are repeated throughout the corpus. The formula “Root type-token ratio (RTTR) = total

collocation type counts/the square root of the total collocation token counts" was employed to reduce the size effect (Paquot, 2018). It indicates the degree to which the same collocation types are repeated in an inverse way. The higher the RTTR score is, the less repetitive an individual collocation type is.

Lastly, the logDice score indicates the probability that the words occur together. The median of logDice scores based on the collocation types was used as the measure for association strength in the present study. The median of logDice scores is a central tendency measure to indicate the overall association strength of all the collocations. A higher median logDice score is indicative of the overall tendency towards strongly associated collocations. A lower median score, on the other hand, indicates the overall tendency towards less strongly associated collocations. To put it differently, words in such collocations also frequently occur with many other words.

3.2.3 Software

The extraction of word combinations of interest from the reference corpus and the textbook corpus was conducted with SketchEngine, which is a web-based corpus analysis tool (https://www.sketchengine.eu/). It includes a set of software tools to analyse patterns of language use in a corpus. A range of functions is offered on the website, among which the function of 'word sketch' returns typical collocates for a given word. This function produces a simple frequency of occurrence of collocations and the association strength between the node word and the collocate in terms of the logDice score.

A Python script was written and combined with Application Programming Interface (API) requests sent to the SketchEngine web server to search the reference corpus for word combinations including the node nouns. The Python script automatically extracted all the co-occurring words for each of the 1,784 node nouns. The data retrieved by the Python script contain the co-occurring words for each of the node nouns, together with the frequency of occurrence and logDice score for each identified word combination.

3.3 Collocation identification in target corpora

To identify collocations, the present study adopted the approach used by Tsai (2015) and Kim (2020). A reference collocation list was first generated from the native reference corpus and used to identify collocations in the textbook corpus. Collocations in the present study were identified through three steps.

Firstly, a reference collocation list was generated from the BNC. Collocations in the reference collocation list meet the minimum criteria with logDice score set at 5 and the co-occurrence frequency set at 5.

Secondly, as was done with the BNC, the Python script was used to retrieve the co-occurring words for 1,784 node nouns from the six textbook corpora respectively. Only verb + noun combinations, adjective + noun combinations, and noun + noun combinations were retrieved.

Thirdly, all the items extracted from the target corpora were checked against the reference collocation list. Those items which were found in the reference collocation list were identified as collocations.

4. Results

4.1 Overall results

The token counts of candidates and collocations are presented in Table 2. It should be noted that

candidates refer to the VN combinations, AN combinations, and NN combinations extracted from the native reference corpus while collocations are those chosen from candidates based on the frequency of occurrence and association measure criteria. In the table, "Token counts of collocations" refers to the total number of collocations occurring in the corpus, with all the instances of a collocation counted as the number of tokens for that collocation.

Table 2 Token Counts of the Candidates and Collocations in the Target and the Reference Corpora

		VNC		**ANC**		**NNC**	
		Candidate	Collocation	Candidate	Collocation	Candidate	Collocation
Textbook Corpus	Average	2,342	1,277	2,174	1,271	910	375
	(S.D)	719	397	764	438	295	126
	Percent (%)	55%		58%		41%	
	Per one node word	1.9	1.1	1.8	1.1	0.8	0.3
Learner Corpus	Total	68,892	41,795	50,224	33,668	17,931	7,667
	Percent (%)	61%		67%		43%	
	Per one node word	39.9	24.2	29.1	19.5	10.4	4.4
Reference Corpus	Total	2,229,763	1,659,707	2,078,475	1,880,156	1,096,196	1,046,843
	Percent (%)	74%		90%		95%	
	Per one node word	1249.9	930.3	1165.1	1053.9	614.5	586.8

Table 3 presents the type counts of the candidates and collocations in each corpus. "Type counts of collocations" refer to the number of unique collocations occurring in the target corpus.

Table 3 Type Counts of the Candidates and Collocations in the Target and the Reference Corpora

		VNC		**ANC**		**NNC**	
		Candidate	Collocation	Candidate	Collocation	Candidate	Collocation
Textbook Corpus	Average	1846	883	1706	894	766	282
	(S.D)	518	234	537	259	240	84
	Percent (%)	48%		52%		37%	
	Per one node word	1.5	0.7	1.4	0.7	0.6	0.2
Learner Corpus	Total	13750	4288	10579	3873	6618	1369
	Percent (%)	31%		37%		21%	
	Per one node word	8.0	2.5	6.1	2.2	3.8	0.8
Reference Corpus	Total	75,727	39,369	66,937	42,745	44,436	34,607
	Percent (%)	52%		64%		78%	
	Per one node word	42.4	22.1	37.5	24.0	24.9	19.4

4.2 Distribution of Collocations in Chinese EFL Textbooks

4.2.1. Collocation Density

The top half of Table 4 shows the total number of collocation tokens in each corpus. There are 2,923 collocation tokens (1,277 VNCs, 1,271 ANCs, and 375 NNCs) on average in each textbook series.

The lower half of the table shows the total number of collocation tokens per 1,000 words, which controls for differences in corpus size and represents the density of collocations in each corpus for the present dissertation. The textbook corpus provides 59.16 collocation tokens (average of collocation counts of each sub-corpus) per 1,000 words while the native reference corpus used 53.27 collocations per 1,000 words. Among these tokens, VNCs (25.84) and ANCs (25.72) are

presented more frequently in the textbook corpus than in the native reference corpus (19.24 VNCs, 21.84 ANCs). On the other hand, there are fewer NNCs (7.59) in the textbooks than in the reference corpus (12.16).

A statistically significant difference was found in the density of ANCs and NNCs between the textbook corpus and the reference corpus, but not in VNCs (ANCs: χ^2=7.9376, p<.05; NNCs: χ^2=8.3001, p<.05; VNCs: χ^2=2.5306, p=0.1116). It implies that the textbooks over-represent ANCs and under-represent NNCs in comparison to the native reference corpus.

Table 4 Collocation Density in the Textbook and Reference Corpora

	Subtype	Textbook Corpus	Reference Corpus
Collocation Token Counts[1)]	VNC	1,277	1,659,707
	ANC	1,271	1,880,156
	NNC	375	1,046,843
	Total	2,923	4,586,706
Collocation Token Counts per 1, 000 Words[2)]	VNC	25.84	19.28
	ANC	25.72	21.84
	NNC	7.59	12.16
	Total	59.16	53.27

[1)]Average collocation tokens per each textbook corpus by publishers
[2)]Collocation tokens/Word counts * 1,000

4.2.2. Collocation Diversity

The top half of Table 5 shows the total number of collocation types in each corpus, which is taken to indicate the diversity of collocations used in a corpus. The textbook corpus consists of 2, 059 collocation types on average with 1,210 curriculum-based noun types as its node word. The lower half of the table presents collocation diversity rates which take account of corpus size. It is found that the diversity rates of VNC types (3.97), ANC types (4.02), and NNC types (1.27) in the textbook corpus are lower than those in the reference corpus (4.24, 4.61, and 3.73, respectively). It is little surprising that the overall number of collocation types in the textbooks (9.26) is smaller than in the reference corpus (12.58). Given the pedagogical purpose of the textbooks and the limited class time, textbooks cannot include all the types of collocations which are found in the native reference corpus.

Table 5 Collocation Diversity in the Textbook and Reference Corpora

	Subtype	Textbook Corpus	Reference Corpus
Collocation Type Counts (Raw Frequency)	VNC	883	39,369
	ANC	894	42,745
	NNC	282	34,607
	Total	2,059	116,721
Collocation Diversity Rates to Text Length[1)]	VNC	3.97	4.24
	ANC	4.02	4.61
	NNC	1.27	3.73
	Total	9.26	12.58

[1)] Collocation diversity rates to text length=Collocation types/ the square root of word token counts

The chi-square test results imply that textbooks present significantly fewer collocation types than

the native reference corpus in terms of all the three subtypes of collocations, as a statistically significant difference was found in the number of types of VNCs, ANCs, and NNCs between the two corpora (VNCs: χ^2=12.4616, p<.05; ANCs: χ^2=94.2688, p<.05; NNCs: χ^2=721.1259, p<.05).

4.2.3. Repetition Rate

The repetition rate in the present study was measured with the Root Type-Token Ratio (RTTR). It is calculated by dividing the number of collocation types by the square root of the total number of collocation tokens. It indicates the degree to which the same collocation types are repeated in an inverse way. In other words, a higher RTTR score indicates less repetition of the collocation types found in a corpus, while a lower RTTR score means more recurrences of the same collocation types (Kim, 2020).

Table 6 presents the degrees to which the same collocations are recycled in the textbook corpus and the reference corpus. Overall, the collocations of all subtypes in the textbooks are more repetitive than their counterparts in the reference corpus. The result is conflicting with previous studies which have shown that textbooks present insufficient repetition of target words or collocations (Kim, 2020; Koya, 2004; Tsai, 2015; Yu & Renandya, 2021). It is noticed that NNCs are recycled to a greater extent than VNCs and ANCs in the textbook corpus (24.71 for VNCs, 25.08 for ANCs, and 14.56 for NNCs). On the contrary, NNCs in the reference corpus were repeated less often than were VNCs and ANCs (30.56 for VNCs, 31.17 for ANCs, and 33.82 for NNCs). This difference in the relative repetition rates for NNCs between the two corpora may be because that noun + noun phrases are typical of formal and academic writing. They are usually technical terms or the embodiment of specific meanings (e.g., *balance sheet*). The topics in the textbooks hinge on pedagogical purpose while the themes in the reference corpus vary across different fields. In other words, the topics selected in the textbooks are those that the textbook authors consider useful and helpful to the students' L2 learning and are believed to satisfy the learners' needs. The written texts of the BNC, on the other hand, include academic books, specialist periodicals, and journals, etc., and thus cover a wider range of topics than the textbooks. It may be one of the reasons why NNCs repeat themselves less often in the reference corpus than in the textbooks.

Table 6 Repetition Rate by RTTR* in the Textbooks and Reference Corpora

Subtype	Textbook corpus	Reference corpus
VNC	24.71	30.56
ANC	25.08	31.17
NNC	14.56	33.82

RTTR*= Collocation types / the square root of Collocation tokens

4.2.4. Association Strength

The logDice is used for determining how strong the association is between the words in a collocation. A high score means that the collocate is more often found together with the node word as compared with other words.

Table 7 displays the median logDice score of collocations in the textbook and reference corpora. The median is the middlemost value of the distribution, that is, the value that splits the data into halves: approximately half largest and half lowest (Mackey & Gass, 2016). It is commonly used when the data contain extreme scores (Mackey & Gass, 2016). In the present study, the median is more appropriate because the data contain collocations with very high logDice scores (greater than 11). The median logDice score for all the sub-types of collocations in the textbook corpus (6.85 for

VNCs, 6.99 for NNCs, and 7.44 for ANCs) is higher than those in the reference corpus (6.1 for VNCs, 6.18 for NNCs, and 6.37 for ANCs).

Table 7 Median LogDice Score of Collocations (types) in the Textbook Corpus and Reference Corpus

Corpus	VNCs	ANCs	NNCs
Textbook	6.85	6.99	7.44
Reference	6.1	6.18	6.37

The association strength of collocations was further divided into five range bands of logDice scores: lower-mid (logDice=5~6.5, not including 6.5), mid (logDice=6.5~8, not including 8), upper-mid (8~9.5, not including 9.5), high (9.5~11, not including 11), and very high (over 11).

Figure 1 illustrates visually the distribution of all collocation types across logDice score bands in the textbooks and the native reference corpus. As shown in the Figure, collocations with the lower-mid level of association strength take up the largest proportion in the textbook corpus and the native reference corpus (37.9% and 55.4% respectively). The reference corpus covers relatively more unique collocations with lower-level association strength ranging from 5 to 6.5 (55.4%) than the textbook corpus (37.9%), while the textbook corpus presents relatively more unique collocations toward the higher end of the association strength in comparison to the reference corpus. The result is in line with the median logDice scores of collocations in the two corpora. The median logDice score in the textbook corpus is higher than 6.5 while that in the reference corpus is lower than 6.5.

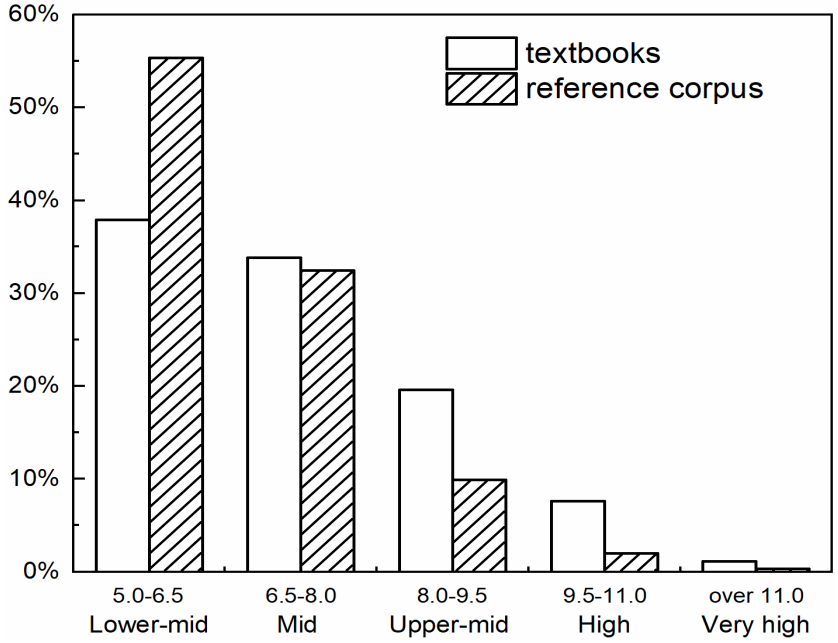

Figure 1 Proportions of collocation types across logDice score bands in the textbook and the native reference corpora (proportions for each corpus add up to 100%)

The divergence between the textbook corpus and the reference corpus is that a majority of collocation types in the textbook corpus (53.4% in total) are distributed at mid-level (6.5-8.0) and upper-mid level (8.0-9.5) of association strength, which is higher than that in the BNC (42.3% in total). Meanwhile, the number of collocation types at the high and very high level of logDice scores over 9.5 take up 8.8% in the textbook corpus, which is much more than that in the BNC (2.4% in total).

The overall findings agree with Kim's (2020) results that a large majority of collocations in the Korean high school textbooks are distributed from the mid to high level (logDice scores over 6.5).

In addition, more types of collocations with association strength from high to very high levels are found in the textbooks than in the reference corpus. It indicates that the EFL textbooks present a majority of collocations which are more strongly associated while the reference corpus prefers collocations with relatively weak associations.

5. Discussion and conclusion

The result shows that collocations in the textbook corpus, on the whole, are much denser, more repetitive, and more strongly associated but less diversified in comparison to those in the native reference corpus.

In terms of the normalized frequency with differences in corpus size controlled for, overall, more collocations occur in the textbook corpus than in the native reference corpus. The main reason may reside in that the number of words in the textbooks is much smaller than that in the native reference corpus. Since the textbooks are purposely designed and used as the main source of L2 input, the textbook writers would present a much denser distribution of collocations within limited texts.

In spite of the denser use of collocations, by and large, the textbooks under-represented NNCs than did the native reference corpus. In other words, textbooks covered comparable tokens of VNCs and a lot more tokens of ANCs, which might lead to an unbalanced acquisition between NNCs on the one hand and VNCs as well as ANCs on the other hand. One plausible reason for the under-representation of NNCs is that NNCs are characteristic of formal or academic writing. Although the texts in the high school EFL textbooks are a kind of academic writing, the topics in the textbooks are selected to cater for the high school students while the topics in the BNC range from newspapers, and specialist journals to academic books. In addition, more proficient learners have been shown to use noun + noun phrases more frequently than less proficient learners (Parkinson & Musgrave, 2014). NNCs might thus have been considered by textbook writers and publishers to be less useful or less important for secondary school students with relatively low proficiency levels.

Collocation diversity was calculated by the number of collocation types divided by the square root of total word tokens to minimize the corpus size effect. The collocations in the textbooks are not as diversified as in the native reference corpus, which is contrary to previous studies that have shown higher collocation diversity in textbooks (Kim, 2020; Koya, 2004; Tsai, 2015). Kim (2020) reported that English textbooks for Korean middle and high school students contained more collocation types of VNCs, ANCs and NNCs than the native reference corpus. More collocation types were found in the English textbooks used in Japan than in the native history textbooks in Koya's (2004) study. In a similar vein, Tsai (2015) found that VNCs in the textbooks were more diversified than in the native reference corpus. However, it does not necessarily imply that the textbooks should present more various collocations. Romer (2004, p.161) claims that "learning a variety of English they will rarely encounter in real-life situations is very unlikely to help learners communicate successfully with competent speakers of English". In other words, useful and important vocabulary items and collocations should be presented in the textbooks. Therefore, it is unrealistic and unnecessary to include all collocation types used by native speakers in textbooks.

The highly repetitive co-occurrence of the same collocations may account for the higher collocation density and less diversity in the textbooks. In other words, the textbooks seem to achieve enhanced repetition at the cost of diversity in terms of collocations. This might be beneficial for L2 learners to the extent that repetition plays a conducive role in language learning. According to Ellis (2001), repetition of sequences allows their consolidation and retention in the long-term memory. It remains inconclusive about how many exposures are required for collocation learning, however. It is claimed in some research that vocabulary learning needs more than six or seven instances of repetition (Peters, 2014; Webb, 2007). Durrant and Schmit (2010) propose that

collocations need to be recycled at least 8–10 times to be acquired. More repetition of collocation types in the textbooks would be conducive to the learners' mastery of collocations.

The spread of collocational strength shows that strongly associated collocations with higher logDice scores are dominant in the textbooks while collocations with comparatively lower logDice scores dominate in the reference corpus. The reason for this difference may be that textbook writers have ascribed greater importance to these strongly associated collocations probably due to greater salience in the input.

The divergences of the four distributional properties between the textbooks and the native reference corpus do not imply that there are many disadvantages in the textbooks. The textbooks are designed and written to equip students with the L2 language at their disposal. The pedagogical purpose of textbooks should be borne in mind. Textbooks do not need to resemble the native texts in all aspects, e.g., the distribution of VNCs, ANCs and NNCs that could be found in the native reference corpus. However, there is still some room for improvement in textbooks. It may be more helpful to the learners if textbook writers could incorporate more diverse collocations to model more native-like texts in the textbooks. Gitaski (1996) contended that inadequate exposure to specific types of collocations in the textbooks has contributed to the learners' avoidance of these types. Therefore, more coverage of NNCs in the textbooks targeting higher-grade learners may benefit their writing.

References

[1] Durrant, P. (2014). Corpus frequency and second language learners' knowledge of collocations. International Journal of Corpus Linguistics, 19(4), 443–477.

[2] Durrant, P., & Doherty, A. (2010). Are high-frequency collocations psychologically real? Investigating the thesis of collocational priming. Corpus Linguistics and Linguistic Theory, 6(2).

[3] Durrant, P., & Schmitt, N. (2009). To what extent do native and non-native writers make use of collocations? IRAL International Review of Applied Linguistics in Language Teaching, 47(2).

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

[5] Ellis, N. C., Simpson-Vlach, R., Römer, U., Brook O'Donnell, M., & Wulff, S. (2015). Learner corpora and formulaic language in second language acquisition. In S. Granger, G. Gilquin, & Meunier (Eds.), The Cambridge handbook of learner corpus research (pp. 357– 378). Cambridge, UK: Cambridge University Press.

[6] Firth, J. (1957). A Synopsis of Linguistic Theory, 1930-55. In Studies in Linguistic Analysis (pp. 1-31).

[7] Foster, P. (2001). Rules and routines: A consideration of their role in the task-based language production of native and non-native speakers. In M. Bygate, P. Skehan, M. Swain (Eds.), Researching pedagogic tasks: Second language learning, teaching and testing (pp. 75–93). Harlow: Longman.

[8] Frankenberg-Garcia, A., Lew, R., Roberts, J. C., Rees, G. P., & Sharma, N. (2018). Developing a writing assistant to help EAP writers with collocations in real time. ReCALL, 31(01), 23–39.

[9] Gablasova, D., Brezina, V., & McEnery, T. (2017). Collocations in Corpus-Based Language Learning Research: Identifying, Comparing, and Interpreting the Evidence. Language Learning, 67(S1), 155–179.

[10] Gitsaki, C. (1996), The development of ESL collocational knowledge. (Doctoral dissertation). University of Queensland.

[11] Granger, S. (1998). Prefabricated patterns in advanced EFL writing: Collocations and formulae. In A.P. Cowie (Ed.), Phraseology: Theory, analysis, and applications (pp. 79–100). Oxford: Oxford University Press.

[12] Hoey, M. (1991). Patterns of lexis in text. Oxford: Oxford University Press

[13] Howarth, P. (1998a). The phraseology of learners' academic writing. In A. P. Cowie (Ed.), Phraseology: Theory, analysis and applications (pp. 161–186). Oxford: Oxford University Press.

[14] HOWARTH, P. (1998b). Phraseology and Second Language Proficiency. Applied Linguistics, 19(1), 24–44.

[15] Kim, Y.S. (2020). A corpus-based analysis of collocations in Korean middle and high school English textbooks and Korean EFL learner writing. (master's thesis). Seoul National University.

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

[17] Laufer, B., & Waldman, T. (2011). Verb-Noun Collocations in Second Language Writing: A Corpus Analysis of Learners' English. Language Learning, 61(2), 647–672.

[18] Lewis, M. (1993). The Lexical Approach. Hove, Brighton: Language Teaching Publications.
[19] Nattinger, J. R., & DeCarrico, J. S. (1992). Lexical Phrases and Language Teaching (Oxford Applied Linguistics) (Illustrated ed.). Oxford University Press.
[20] Parkinson, J., & Musgrave, J. (2014). Development of noun phrase complexity in the writing of English for Academic Purposes students. Journal of English for Academic Purposes, 14, 48–59.
[21] Peters, E. (2014). The effects of repetition and time of post-test administration on EFL learners' form recall of single words and collocations. Language Teaching Research, 18(1), 75-94.
[22] Reppen, R. & Vlach-Simpson, R (2010). Corpus linguistics. In N. Schmitt (Ed.), An introduction to applied linguistics 2nd edition. 89–107. London: Arnold
[23] Schmitt, N. (2012). Formulaic Language and Collocation. The Encyclopedia of Applied Linguistics.
[24] Tsai, K. J. (2014). Profiling the collocation use in ELT textbooks and learner writing. Language Teaching Research, 19(6), 723–740.
[25] Webb, S. (2007). The Effects of Repetition on Vocabulary Knowledge. Applied Linguistics, 28(1), 46–65. https://doi.org/10.1093/applin/aml048
[26] Wray, C. L. C. R. A., Wray, C. F. L. A. C. R. A., & Cambridge University. (2002). Formulaic Language and the Lexicon. Cambridge University Press.
[27] Yu, M., & Renandya, W. A. (2021). A corpus-based study of the vocabulary profile of high school English textbooks in China.