Asian Journal of English Language Teaching, 32(1), 131–154

Measuring Vocabulary Use in Chinese Tertiary Textbooks: Potentials for Incidental Vocabulary Learning

Chen WANG
Guangdong University of Foreign Studies, China

Yuhua LIU
Jiangxi Normal University, China
Guangdong University of Foreign Studies, China

This article explores the underresearched area of single words and collocations in English textbooks for Chinese tertiary students. The present study is based on a corpus of English tertiary textbooks consisting of texts from the three most widely used materials in Chinese universities covering two years of English learning. By examining a range of indexes of single words and collocations, this study reveals interesting findings for EFL learners. Our results show that textbooks might not provide enough opportunities for students' incidental learning of vocabulary. In this case, students' current level of vocabulary size would have to be enough to achieve acceptable comprehension of the texts. Our analyses showed that the majority of single words and collocations appeared only once in the textbook series. In addition, the results showed that the overlap between the single words and the required vocabulary list issued by the Ministry of Education in China is relatively modest. Relevant implications are discussed based on the findings.

Introduction

Vocabulary use occupies an important position in the field of applied linguistics. The available current literature suggests that the word use of language learners in a foreign language context, compared to native speakers, has a greater number of errors, an overreliance on high-frequency single words and collocations, and a smaller number of collocations (e.g., Laufer & Waldman, 2011; Siyanova-Chanturia & Spina, 2015). These results have fueled increasing research into ways to enhance the efficacy of vocabulary education in an English as a Foreign Language (EFL) setting. However, compared to the interest in intentional and incidental vocabulary learning, studies in vocabulary in textbooks are relatively few.

In a foreign language learning context, textbooks play a central role in language input both inside and outside classrooms, acting as the pivot of language learning and the teaching process. These materials prescribe the vocabulary items that should be taught in the classroom and outline the sequence and manner in which they need to be presented to students (Nordlund, 2015). Textbooks are an indispensable carrier of new knowledge and ideas and ensure the realization of learning outcomes (Ministry of Education, 2020). It is thus important for second language researchers to look into the vocabulary in textbooks to see if they provide sufficient learning opportunities for students.

To better understand the vocabulary in Chinese tertiary textbooks, this study attempts to profile the single words and collocations in the reading texts of these materials to see if they provide enough learning opportunities for incidental vocabulary learning. It is anticipated that the results of this study could inform textbook and syllabus design and enhance the learning efficacy of vocabulary.

Lexical Coverage

Lexical coverage is a widely used concept in studies of vocabulary use. It refers to the percentage of known words in the text (Nation, 2013). Previous studies have found that certain lexical coverage thresholds are informative for reading comprehension. Hu and Nation (2000) estimated that to achieve an adequate understanding of the text, a lexical coverage ranging between 95% and 98% is needed. In a more recent study, Schmitt et al. (2011) used a fine-grained analysis to examine the scenario

and suggested that a lexical coverage of at least 98% is needed to achieve optimal comprehension of academic texts.

These two lexical thresholds have also been applied to other studies of vocabulary size and vocabulary in textbooks. In their research focused on academic texts, Dang and Webb (2014) found that 5,000-word families plus proper nouns and marginal words could provide adequate comprehension of academic texts of all types (96.95% for arts and humanities and 98.12% for social science). Similarly, Hsu (2014) showed that 5,000-word families plus proper nouns, abbreviations, and compounds would be sufficient to achieve a coverage rate of 95% for business and engineering textbooks.

This study profiles the vocabulary in tertiary textbooks that is used as the major language input for EFL learners. It would be reasonable for these textbooks to be challenging for students who enter this stage of learning with their existing vocabulary knowledge acquired in their secondary education. Therefore, this study adopted 95% and 98% coverage as benchmarks to analyze the lexical demands of tertiary textbooks for general purposes.

Repetition

Different scholars have presented similar conclusions on this matter and concluded that repetition plays a role in learning new single words and collocations in incidental and intentional learning (e.g., Boers et al., 2014; Peters, 2014; Peters & Webb, 2018; Teng, 2020; Webb, 2007). The relationship between repetition and vocabulary learning is affected by several variables, including students' personal traits, word and text features, and the types of treatment during the study journey (Uchihara et al., 2019).

It is believed that more encounters with words and collocations increase the likelihood of better gains in various aspects of vocabulary knowledge. Relevant studies have found that one encounter can lead to learning gains in receptive knowledge of words from reading (40–67% depending on the types of knowledge tested by Webb, 2007; 19–52% depending on the types of test and treatment conditions, according to Teng, 2020).

Other related studies found that the learning gains of the receptive orthographic knowledge of words were higher than those of meaning (43%

for orthographic form and 1% for meaning in Chen & Truscott, 2010; 67% for orthographic form and 58% for meaning in Webb, 2007). These studies suggested that three to seven encounters would show significantly higher learning gains than one encounter in receptive knowledge in some aspects of vocabulary knowledge (Chen & Truscott, 2010; Teng, 2020; Webb, 2007).

Recent studies focusing on the incidental learning of vocabulary found that frequency of occurences (1–6 occurrences) is one of the many variables that would affect vocabulary learning. According to Peters and Webb (2018), some potential variables could be the cognateness and relevance of the target items to the comprehension of the texts in a video. Other related studies into collocations have found that more encounters are needed to learn collocations than single words. Peters (2014) looked into the learning of single words and collocations in reading together and found that for one, three, and five encounters, learning gains in single words are higher than those in collocations in the receptive knowledge test (49–80% for single words and 42–67% for collocations). Similarly, Webb et al. (2013) showed that, assisted by listening, the learning gains of collocations from reading-graded readers yielded significantly better results at all four thresholds of encounters in both receptive and productive knowledge tests (3–54% in 1 encounter, up to 55–82% in 15 encounters).

These research outcomes suggest that to experience real learning gains, students would need 1 to 10 encounters for single words and 1 to 15 encounters for multi-word units. Uchihara et al. (2019) pointed out that the relationship between frequency of occurences and vocabulary knowledge is much more complex than previously understood. They stressed that "the role of frequency in lexical learning is greatly complicated by the presence of numerous other variables, including individual differences and word characteristics" (p. 31). Since our current study aims to profile the single words and collocations in tertiary textbooks, both learning modes, that is, incidental and intentional, are involved in the process.

The learning goal of vocabulary textbooks is more than building form-meaning links between the items, as often tested in the abovementioned studies. These materials are also designed to build partial to full knowledge of the nine aspects of knowledge identified by Nation (2013). Repetition is the aspect of vocabulary knowledge most often examined in textbook analyses, and many scholars have examined with different methodologies

how textbooks provide enough exposure to the target single words and collocations for learning to occur (e.g., Liu & Zhang, 2015). Considering what has already been done in this study field, this research analyzes the repetition of vocabulary items at different numbers of encounters together with the frequency levels of words.

In addition, we have identified that previous studies have not examined the learning opportunities of vocabulary considering different modes of learning, that is, intentional and incidental learning, separately from textbooks. Our study addressed this issue by looking into how incidental learning of vocabulary would take place in learning from textbooks. The Chinese tertiary textbooks included in this research provide one text for intentional learning of words for each unit, in which the teacher would explicitly explain new vocabulary to the students. Additionally, these materials provide another one or two texts for further reading, in which the primary goal is the comprehension of the general meaning. The robust evidence provided above suggests that the frequency of occurences of unknown words in written texts would contribute to vocabulary learning (e.g., Uchihara et al., 2019). Our goal is to understand whether the written texts in textbooks provide enough opportunities for incidental learning of vocabulary in terms of the number of encounters.

Previous Studies on Vocabulary in Textbooks

Different scholars who have engaged with second language research and educational research (e.g., Abello-Contesse & López-Jiménez, 2010; Biber et al., 2004; Harwood, 2014; Matsuoka & Hirsh, 2010) have shown interest in studying vocabulary in textbooks. Some researchers who focused on this phenomenon adopted the corpus-based approach to examine the frequency level of single words and collocations in textbooks compared to base word lists (e.g., Bi, 2020; Criado & Pérez, 2009; Hsu, 2018). Researchers could find out the level of difficulty of words in the textbooks and verify whether the textbooks have provided enough learning opportunities for learners (e.g., Nordlund, 2015, 2016). Tsai (2015) generated a list of verb + noun collocations based on single word frequency in the British National Corpus (BNC) and identified the coverage of the generated list in the three textbooks used in Taiwan. He found that only a small number of collocations in the list were used in

textbooks and that they did not recur frequently enough to sustain learning.

The other two studies on English textbooks published in China used the bottom-up approach, whereby the lists of collocations were generated from the textbooks. Ren (2014) analyzed the presentation of lexical bundles in the textbooks, considering the total number, types of grammatical combinations, and pragmatic functions in four textbooks of college English. He concluded that there were more verb, noun, and prepositional phrases in the texts and more referential bundles and discourse organizers than stance bundles. Liu and Zhang (2015) also analyzed the single words and lexical bundles in textbooks, with a particular focus on the overlap between the lexical bundles and the vocabulary list in *The College English Curriculum Requirement* issued by the Higher Education Department of the Ministry of Education (2007). They found that the overlap rate between both was quite high, while the frequency of occurences of the words and bundles that appeared in both the textbooks and the vocabulary list was quite low.

The results of these studies have pointed out the inefficiency of the textbooks in providing conducive materials for vocabulary learning. These studies combined all the texts in the textbooks as a single corpus but neglected the fact that the textbooks are used by learners over two or three years. This type of analysis might overlook the nuance in a series of textbooks, so portraying the lexical difficulty of words in the series of textbooks can lead to a better understanding of vocabulary learning from textbooks.

Chinese Tertiary Textbook Development

Previous studies have looked into the principles and guidelines in the compilation of college English textbooks in China (e.g., Yang, 2018). The development of tertiary textbooks is supposed to strictly follow *The College English Curriculum Requirement* issued by the Higher Education Department of the Ministry of Education (2007) (hereafter, the Requirement). The Requirement also specifies the vocabulary items to be included in textbooks and classroom teaching (hereafter, the Requirement List). All textbooks have to be examined to see if the vocabulary in the books is covered by the official list in the Requirement (Jin et al., 2016).

According to the Requirement, all tertiary students in China have to attend English courses in the first two years of their college education.

Upon entering universities, the students must have mastered at least 3,500 words, following the requirements of the National Entrance Examination Board. The latest *Guideline for College English Education* (Ministry of Education, 2020) (hereafter, the Guideline) specified that learners should add 2,000 to 3,000 new words to their existing vocabulary repertoire when they finish their tertiary study. Therefore, we could set the vocabulary size of 3,500-word families as the starting point to analyze the vocabulary in textbooks.

This research aims to address the vocabulary in the textbooks used by first- and second-year tertiary learners in China and examine how the textbooks provide learning opportunities for this group of students. The framework designed for this study is also helpful to verify whether, 13 years after its first issue, the Requirement List is still a useful yardstick of vocabulary in tertiary textbooks. Therefore, the following research questions were outlined to address the underresearched areas in the textbook study:

1. What are the vocabulary profiles of the English textbooks used in Chinese universities?
2. Do these textbooks provide enough learning opportunities for vocabulary?

The Corpus

According to the Requirement, tertiary English courses should last for two to four semesters. We built a corpus comprising three English textbooks that are widely used in Chinese universities: *New College English* (2nd ed.; hereafter NCE) published by Shanghai Foreign Language Education Press; *New Standard College English* (2nd ed.; hereafter NSCE); and *New Horizon* (hereafter NH), by Foreign Language Teaching and Research Press. These materials were chosen based on their representation of tertiary English textbooks and their compliance with the Requirement. Each set consists of eight to twelve volumes designed for four semesters (four for reading and writing, four for listening and speaking). The texts extracted from the reading and writing sections of these textbooks constitute the corpus of this study.

All the reading texts were examined, including the two pieces extracted from each unit in the textbooks. However, the exercise sections of the textbooks were eliminated since this study concerns the vocabulary

embedded in the reading materials, which could be learned explicitly through instruction or implicitly through reading (Webb et al., 2013). In total, our textbook corpus consists of 284 texts with 249,360 running words (Table 1).

Table 1. The Textbook Corpus

Textbook	NCE	NSE	NH
Texts	120	64	80
Words	129,009	59,843	60,508

Data Processing

Lexical Difficulty and Lexical Coverage

To explore the use of single words, the present study used word families as counting units. Word families include the headword, its inflected forms, and its related derived forms (Nation, 2013). For example, *teach*, *teaching*, *taught*, and *teacher* all belong to the same word family under the headword *teach*. Some researchers have expressed concerns about using word families as a counting unit and stated that it cannot be assumed that learners know all the derivational forms and that not all members in a word family share similar meanings (e.g., Lei & Liu, 2016; Ward & Chuenjundaeng, 2009). However, we chose this approach because the textbook corpus consists of texts for tertiary learners who have a background of at least eight years of English education before they enrolled. Therefore, they already have accumulated some basic knowledge of both inflected and derivational forms of a word. Second, it is a general practice that teachers present a new word along with its derivational forms to promote the study of related words, so working with word families makes the design of our research more practical.

To measure lexical difficulty, VocabProfile (Laufer & Nation, 1995) was used to verify the frequency levels of all running words in the corpus. This online tool developed by Tom Cobb (2021) is embedded with the 20 frequency lists of the British Nation Corpus (Nation, 2004) and can show the proportion of words at each frequency level, providing a clear profile of lexical difficulty in terms of word frequency.

In addition, this measurement tool can be used to examine lexical coverage. Measuring lexical coverage would help us see the vocabulary size learners need to understand the texts and reveal the extent to which

they may need external help to comprehend the texts. Based on the abovementioned studies, we set the two lexical thresholds at 95% and 98% to analyze the texts included in this study.

Word List and Collocations

The textbook corpus was run on AntConc (Anthony, 2020) to retrieve word lists of single words and n-grams for the analysis of collocations. *Wordlist* in AntConc was used to retrieve the word lists of each textbook for the study of repetition. Finally, the software Range (Nation, 2005) was used to compare the word list of the textbook corpus to those published by the Chinese Ministry of Education.

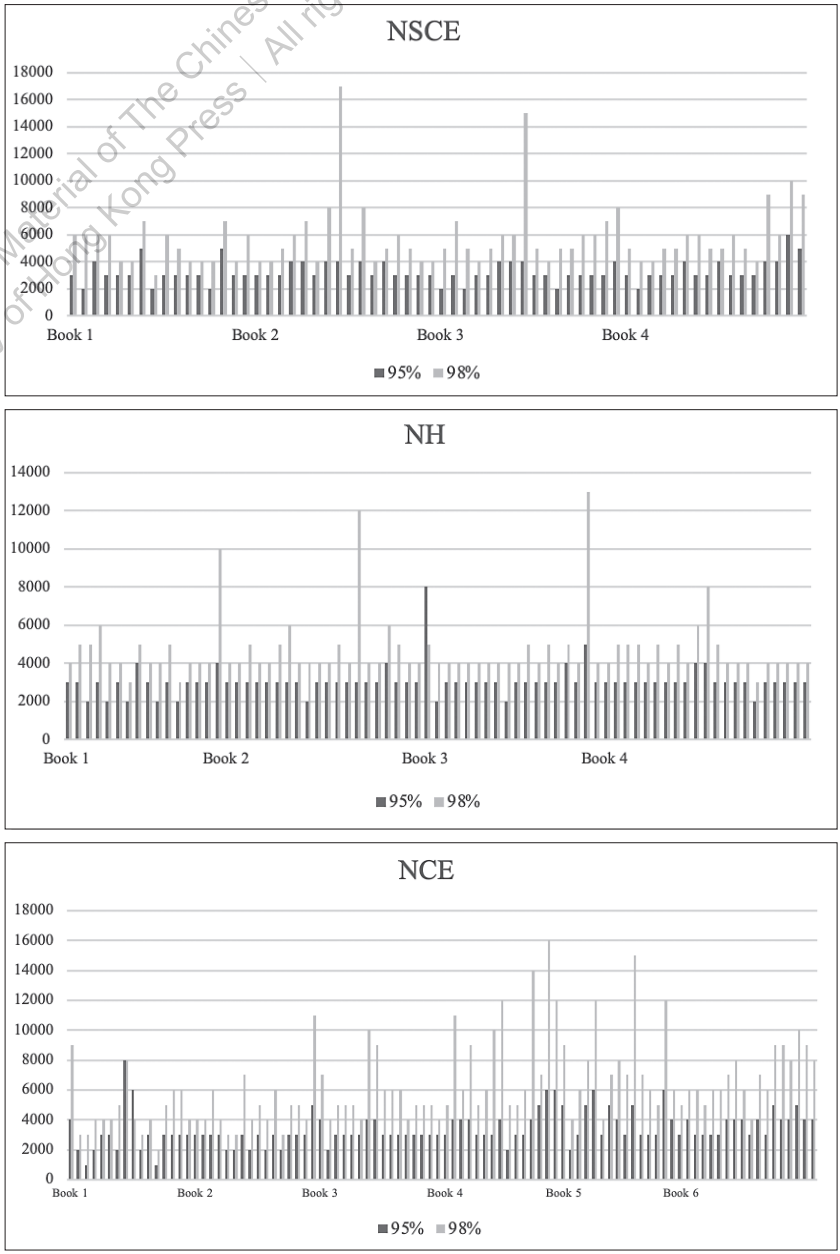
N-gram in AntConc was used to retrieve multiword units from the corpus. All bigrams were collected from the textbook corpus. A bigram is likely to be frequently used not because of its formulaic status but because its constituents are high-frequency words, such as *of the*, *in the*, *it was*, *to be*, and so on. Bigrams such as these would not be of much value for the present study. Considering this, a further step was taken to identify the two-word collocations in the bigrams retrieved from the corpus. We manually identified the verb + noun and adjective + noun collocations from the bigram lists. Collocations that include numbers (e.g., three students), proper nouns (e.g., Chinese people), pronouns (e.g., my writing), and idioms (e.g., on the same boat) were not included in the study. Previous research on collocations (e.g., Peters, 2016; Wang, 2019; Yamashita & Jiang, 2010) has shown that these two types pose some learning challenges for students, so it can be interesting and valuable to look into how they are presented in the textbooks.

Results

Research Question 1: Lexical Difficulty and Coverage

To examine the lexical coverage of the texts, we retrieved the percentage of known words in the texts at two thresholds (95% and 98%). Figure 1 shows the lexical coverage of three textbooks with proper nouns and acronyms removed. Proper nouns and acronyms were removed because they would not create much difficulty in the understanding of the texts since the textbooks provide explanations for each proper noun in the footnotes and culture notes.

Figure 1. Lexical Coverage at Two Thresholds of the Three Textbooks



On average, for all the textbook series, around 3,200 words (M = 3,250, SD = 950) would be necessary to achieve 95% coverage, with a variance of 2,000 to 8,000 words. Also, 3,600 words (M = 3,600, SD = 2,400) would be necessary to achieve 98% coverage in the textbooks, varying from 3,000 to 16,000 words. Among the three textbooks, NH has the lowest demand for vocabulary size at both 95% and 98% coverage (M = 3,000, SD = 700; M = 4,700, SD = 1,600). NSCE has the highest demand of vocabulary size at 95% coverage (M = 4,600, SD = 780; M = 5,600, SD = 2,000), and NCE has the highest demand for vocabulary size at 98% coverage (M = 3,400, SD = 1,100; M = 6,300, SD = 2,700). These results suggest that learners could easily achieve an acceptable understanding of the texts with their current vocabulary size of 3,500 words and might need additional help to achieve optimal understanding (Laufer & Ravenhorst-Kalovski, 2010; Nation, 2006).

The vocabulary size needed to achieve the two thresholds of coverage of the three textbooks was not normally distributed according to the results of the Kolmogorov-Smirnoz analysis. The nonparametric analysis of Kruskal-Wallis H was used to detect variances in the textbook series (Table 2). It can be seen from Table 2 that, among the three textbooks, NCE demonstrates a more variant picture of lexical requirements, and the table shows significant differences in the demand for vocabulary size between the six books.

The post-hoc Mann-Whitney U analysis with Bonferroni-corrected probability value shows significant differences between groups (Table 3). As shown, we identified significant differences between the first two books and the fourth, fifth, and sixth books at two thresholds of lexical coverage. On the contrary, NSCE and NH showed steady demand for vocabulary size in the eight textbooks, regardless of the level of study. These results suggest that as students keep progressing, NCE has a higher requirement for vocabulary size to comprehend the text. Differently, NSCE and NH maintained the same level of lexical demand for text comprehension.

Table 2. Statistical Results of Kruskal-Wallis *H* Analysis of the Lexical Coverage of Textbooks

95% coverage	Chi-square	*df*	Sig.	98% coverage	Chi-square	*df*	Sig.
NSCE	3.550	3	.314	NSCE	1.904	3	.593
NH	3.033	3	.387	NH	0.352	3	.950
NCE	31.151	5	.000	NCE	20.930	5	.0002

Table 3. Statistical Results of Mann-Whitney *U* Test Of NCE

95% coverage	Book 4	Book 5	Book 6	98% coverage	Book 4	Book 5	Book 6
Book 1	–2.684 (.007)	–2.547 (.011)	–2.841 (.004)	Book 1	–4.046 (.000)	–3.208 (.001)	–3.590 (.000)
Book 2	–3.053 (.004)	–2.916 (.004)	–3.210 (.001)	Book 2	–3.821 (.000)	–2.983 (.003)	–3.365 (.001)
				Book 3	–2.591 (.010)		–2.591 (.033)

Note. Z value followed by probability value (sig.) in parentheses.

The results suggest that the three textbooks could help learners achieve the basic vocabulary learning goal of 2,000 words specified by the Guideline (Ministry of Education, 2020). However, students would need additional input if the higher goal of 3,000 words had to be achieved. By looking at the number of repetitions of single words in the texts, it is evident that the textbook alone might not provide enough exposure for vocabulary acquisition to occur.

Research Question 2: Repetition and Overlapping

Single Words

Based on the study of word repetition on incidental vocabulary learning, the frequency of one, three, seven, and ten times could witness sizable learning on the receptive and productive knowledge of new words (Webb, 2007). We retrieved word lists for each set of textbooks and worked out the number of word types that have occurred at the designated frequency levels (Table 4). Of all three sets of textbooks, around 50% of the word types happened only one to two times, which means that those words could leave only a very vague impression in students' minds. According to Webb (2007), sizable productive knowledge could be witnessed after seven encounters. We identified that around 5% of the word types repeat seven to nine times in the textbooks. The remaining word types (around 10%) occur more than ten times in the textbooks, which could represent a sizable learning gain in both receptive and productive knowledge.

Table 4. Repetition of Single Words in the Textbooks

	≥1	≥3	≥7	≥10
NSCE	3,914 (51%)	1,393 (18%)	336 (4%)	714 (9%)
NH	3,401 (44%)	1,583 (22%)	425 (5%)	830 (10%)
NCE	5,690 (45%)	2,650 (21%)	664 (5%)	1,542 (12%)

Note. Percentages of all word types are in parentheses.

In addition to knowing the number of occurrences of the word types, it would also be interesting to know the words that appear at these four thresholds to see if the words unknown to the learners are repeated frequently enough for learning to take place. We analyzed the types of words in two ways. First, the word types at each threshold were analyzed using the Vocabprofilers (Cobb, 2021) to identify the lexical difficulty. Second, we compared the word list with the Requirement List to explore how the required words were repeated in the textbooks.

First, we looked at the types of words that have been repeated more than ten times in the textbooks. The Vocabprofilers (Cobb, 2021) results show that of the three textbooks, more than 90% of the word types that appear more than ten times are high-frequency words, ranging from 1,000-word level to 3,000-word level. This means that, at this stage of learning, students are well acquainted with these words. The remaining word types, 9 for NSCE, 31 for NCE, and 10 for NH, are above the 4,000-word level; in other words, they are likely to be unknown. This finding suggests that very few word types occur frequently enough for sizable learning gains in receptive and productive knowledge of words to take place.

The results regarding the overlap range between the words in the textbooks and the Requirement List show that the three textbooks have a different degree of overlap (Table 5). The number of overlaps between the word types in textbooks and the Requirement List at levels 1–4 is strikingly low in NSCE and NCE. Only 20% of word types in NSCE and 15% in NCE overlap with the Requirement List. Generally speaking, the overlap between the textbooks and the Requirement List is higher at levels 5–6, which means that more word types in the textbooks come from words with higher learning demands. In total, NH shows the greatest degree of overlap, which means that 67% of the word types in this material are from the required word list. NSCE has the lowest degree of overlap, with only 44% of the word types in the textbooks appearing in the Requirement List.

Table 5. Overlap Between Textbooks and the Required Word List Published By the Ministry of Education

	NSCE	NH	NCE
Word list (levels 1–4)	20%	36%	15%
Word list (levels 5–6)	24%	31%	36%
Total	44%	67%	51%

Collocations

We retrieved bigrams from three textbooks and manually collected verb + noun and adjective + noun collocations for analysis. The results show that the total number of collocations constitutes less than 10% of the bigrams (e.g., 1,066/34,737 for NSCE) (Table 4). To further understand the collocations in the textbooks, we analyzed the number of repetitions of the two different types of collocations. Based on Webb et al.'s (2013) study on the number of encounters, we categorized them accordingly at four thresholds of encounter (1, 5, 10, and 15). The overwhelming majority of the collocations occur fewer than five times (Table 6).

Table 6. Number of Collocations at Four Thresholds in the Three Textbooks

	≥15	≥10	≥5	≥1
NSCE		1 (1,066)	9 (1,066)	1,056 (1,066)
NH			3 (1,232)	1,229 (1,232)
NCE	1 (1,893)	1 (1,893)	21 (1,893)	1,870 (1893)

Note. The total numbers of collocations are shown in parentheses.

In addition to profiling the two types of collocations in the textbooks, it would also be interesting to learn if the new words in the textbooks are presented with collocates, since precious studies have shown that new words presented with frequent collocates show better retention (Wang & Yang, 2020). To find out the collocations of the new words, we selected new words that appear more than ten times in the three textbooks as examples, based on the notion that these words are most salient to the learners and that their high exposure would lead to better learning gains.

Twenty-one new words from the 4,000 level to the 12,000 level in the BNC were searched for collocates with AntConc, and later these collocates were checked with the Contemporary Corpus of American English for their association strength to identify strong collocations (mutual information score >3) (Stubbs, 2001). Ten new words were presented with strong collocates: *bolt*, *urban*, *ethical*, *couch*, *myth*, *greenhouse*, *latitude*, *longitude*, *scholarship*, and *misconduct.* For example, for the noun node word *bolt,* the textbooks presented the words with their adjective and verb collocates, which are *drop* and *dead*. For the adjective node word *ethical*, it appears with noun collocates (*issue*, *behavior*, *person*, *question*, *standard*, *reservation*, *conduct*, and *foundation*) in the texts. For the noun node word *couch*, it appears every time with its noun

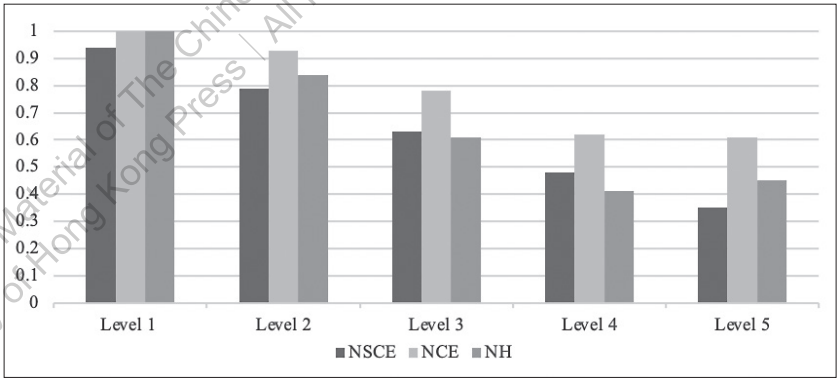
collocate *potato* in the textbooks. We see from the results that, at least for new words that appear more than 10 times in the textbooks, learners could associate them with strong collocates.

The results on collocations in textbooks show that the number of occurrences for two-word collocations is not sufficient to support sizable learning gains. However, at least for high-frequency new words in the textbooks, there is a good chance that students could learn them with the collocates. This result could explain the lack of adequate knowledge and use of collocations in Chinese tertiary learners' writing previously observed by other scholars (e.g., Wang, 2019). As the major source of input for this group of learners, the textbooks fail to provide enough scaffolding for the learning of collocations to successfully occur.

The last perspective on exploring collocations in the textbooks is to identify the extent to which the textbooks cover useful collocations that are essential for learning. The baseline collocation list used in this study was the PHRASal Expressions List (hereafter, PHRASal List) (Martinez & Schmitt, 2012). We chose this list because it is a multiword unit list compatible with the list of single words that could be used to identify lexical difficulty. Furthermore, it has a precise categorization of phrases for three styles of language: multiword units for spoken, general written, and academic language. For the present study, we chose the multiword units that are most commonly used in general written language in the PHRASal List, which is the language style that best describes the language in the three textbooks.

We retrieved 342 multiword units from the PHRASal List, and they were all marked with three asterisks, in other words, units most commonly used in the general written genre. These units consist of five frequency bands based on the most common word families in BNC. The chosen multiword units were made up of two to four words, and therefore the textbook corpus was analyzed by AntConc to retrieve n-grams accordingly. Figure 2 reports the percentage of multiword units that appeared in the three textbooks. Among them, over 50% of the multiword units from the PHRASal List appeared in the textbooks (56% for NSCE, 75% for NCE, and 55% for NH). We can interpret from the figure that the number of multiword units that appeared in the textbooks decreased with the increase in frequency levels. NCE outperformed the other two textbooks in terms of the number of multiword units overlapping with the PHRASal List at all five levels.

Figure 2. The Percentage of Multiword Units That Appeared in the Three Textbooks



Another perspective to look at the overlapping multiword units in the textbooks and PHRASal List is the number of repetitions. We explored four thresholds (1, 5, 10, and 15) according to the categorization of occurrence mentioned above. Table 7 shows the summary of the number of repetitions of the overlapping multiword units. We can see from the table that the majority of the multiword units appeared fewer than five times in the textbooks, which means that it is unlikely that one can witness significant gains in both receptive and productive knowledge of these units. Additional help is needed for substantial progress to be witnessed.

These results suggest that the overlapping between the multiword units most commonly used in the general written genre and the three textbooks falls from 100% at the most frequent level to around 40% at the most infrequent level.

Table 7. Number of Multiword Units at Four Thresholds That Appeared in Both the Textbooks and the PHRASal List

	≥15	≥10	≥5	≥1
NSCE	15 (193)	9 (193)	28 (193)	140 (193)
NH	12 (188)	8 (188)	28 (199)	140 (188)
NCE	20 (256)	19 (256)	54 (256)	162 (256)

Note. The total numbers of multiword units are shown in parentheses.

Discussion

Our study provides a comprehensive and detailed profile of the single words and collocations found in reading texts from English textbooks used by Chinese university students. Based on the corpus created through widely used textbooks, it reports the single words and collocations in the three series of textbooks holistically and also the changes in lexical difficulties in these series of classroom resources.

Our examinations pointed out several major findings. First, we concluded that the lexical difficulty of single words and collocations remains quite static in the series. The results from the statistical analysis suggest that only NCE requires different vocabulary sizes to achieve 95% and 98% lexical coverage. The remaining two textbook series show quite consistent demand for vocabulary size to achieve these two thresholds of lexical coverage. It would require around 3,000 words to achieve an acceptable understanding of the reading texts (i.e., 95% of lexical coverage), and 5,600 words for optimal understanding (i.e., 98% of lexical coverage).

Although the demand for vocabulary size seems quite high in some textbooks, these high-demanding words are closely related to the topic of the text. For example, infrequent words above the 10,000-word level in BNC, such as *counter-attack, permafrost,* and *heredity* in NCE, are closely related to the topics of the text (the Normandy landing during World War II, global warming, and globalization). These sets of words could be guessed based on the context and would not pose many difficulties for learners at the tertiary level of study.

According to the requirements of the National College Entrance Examination, learners are supposed to have 3,500-word families to pass the test and graduate from high school. This means that, with their vocabulary size, learners could achieve an acceptable understanding of all the reading texts in their textbooks without additional help from teachers. In this context, teachers could step in to help learners acquire the remaining 2,400 words to achieve an optimal understanding of the texts. This result is in line with the latest Guideline (Ministry of Education, 2020), which specifies that the basic learning goal for tertiary learners is to master 2,000-word families in two years of study. However, the lexical difficulty of the textbooks would not suffice for the higher learning goal of 3,000 words, and if it is quite static across the two years of study, it could mean that similar sets of words are recycled in the

textbooks. Although this is good news for vocabulary learners, when we look at the repetition of single words and collocations in the textbooks, this rarely happens.

Second, the results suggest that reading texts in textbooks would not provide enough learning opportunities for lexical items for incidental learning. The results on single words and two-word collocations in textbooks show that the number of occurrences for these vocabulary items is not sufficient to support sizable learning gains, according to the results of research on incidental learning of vocabulary. Similar findings have also been reported in studies on vocabulary in textbooks (e.g., Abello-Contesse & López-Jiménez, 2010; Matsuoka & Hirsh, 2010; Nordlund, 2015). In her study of vocabulary in primary textbooks, Nordlund (2015) found that more than 40% of the words (nouns, adjectives, and verbs) occur only once in the textbook series, and the frequency is even lower for higher levels of study (e.g., around 60% for adjectives for grade 6). In a similar study of vocabulary in textbooks for young learners, Nordlund (2016) found that 67% to 90% of words (nouns, adjectives, and verbs) occur from one to four times in the two textbook series. Over 50% of single words and 99% of collocations appear only once in the reading texts. This might not be a problem for intentional learning, since the teacher would spend time explaining, practicing, and assessing the items. However, in incidental learning, learners might not even notice the new items, not to mention gradually building up knowledge of them.

The vocabulary items that appear more than ten times in the texts are high-frequency items (e.g., for single words, from the 1,000- to the 3,000-word level) with which the learns are well acquainted. For example, when we look at the collocations that appear more than five times in the textbooks, we see that they are already well-known by tertiary learners, for example, *high school* (18 times), *global warming* (12 times), *other people* (11 times), *next morning* (9 times), and *young people* (5 times). This evidence suggests that single words and collocations that are unknown and should be learned do not appear frequently enough to allow sizable learning gains to occur.

Our results also point out that the overlap between the single words in the textbook series and the Requirement List is quite modest. This result differs from the ones presented by Liu and Zhang (2015), who found that the overlap between the single words and the Requirement

List ranges from 78.8% to 98.5% for the three series of textbooks. The discrepancy between these two outcomes is possibly due to the version of the Requirement List. Liu and Zhang used a list categorized into three levels with no overlap between them. On the contrary, our study used two lists of 1–4 and 5–6 levels that included some words of higher learning demand, so the results of the overlaps are a bit confusing. The three textbooks were supposed to be compiled following the instructions in the Guideline and Requirements.

However, the results of the present study show that the overlap between the lists is around 50%. The discrepancy between the lists could be for two reasons. First, considering that the overlap is higher in the 5- to 6-word levels, the Requirement List could reflect a lower learning demand according to the uneven regional education in China (Liu & Zhang, 2015). Second, the Requirement List does not distinguish between genre types, so the reading texts include mainly words that appear in the written genre. Comparatively, the Requirement List may cover words from spoken language.

Pedagogical Implications

This study has strong implications for vocabulary teaching and learning in China. The results suggest that textbooks alone do not provide enough opportunities for vocabulary learning, so extensive support is needed to achieve sizable learning gains. Teachers could design a teaching process to increase the exposure of vocabulary items, for example, by going through the word lists before analyzing items in the texts and reviewing these items at certain periods to reduce attrition. Enhancement techniques could also be applied in textbooks, for instance, with visual enhancements that include underlining new vocabulary items, using bold letters, or providing a glossary next to the texts (e.g., Boers et al., 2014).

In addition, the small overlap between the textbooks and the Requirement List implies that it is time to build a new and more refined vocabulary list. The reference information on the list was based on the Collins and Longman dictionaries developed in 1995. This undermines the time efficiency of the list as a compulsory guideline more than 25 years later. Also, it might be useful to categorize the list based on genres to increase its applicability.

Conclusion

This article reports a vocabulary study of Chinese tertiary textbooks to explore the use of single words and collocations covering two years of college English education. The three textbooks consist of 284 reading texts and 249,360 running words. For single words, we explored the lexical coverage in the three sets of textbooks and the number of repetitions. Our results showed that to achieve 95% and 98% of lexical coverage in the reading texts, around 3,200- and 5,600-word families would be needed, respectively. In addition, only NCE demonstrates a significantly higher demand for vocabulary size in textbooks later in the series.

We also explored the overlap between the word list retrieved from the reading texts and the Requirement List provided by *The College English Curriculum Requirement* (Ministry of Education, 2007) and found that the overlap rate was around 50% for the two levels of lists. To understand the use of collocations in textbooks, we further examined the repetition of collocations and the overlap between the multiword units and the PHRASal List. The results of the repetition of verb + noun and adjective + noun collocations were considered alarming. In three textbooks, fewer than 20 collocations appeared more than 10 times and 60 appeared more than 5 times. Although the 10 high-frequency words examined in the study appeared with fixed collocates to facilitate learning, the overwhelming collocations were not salient enough. The overlap between the multiword units in the textbooks and the PHRASal List diminished with the increase of frequency level, falling from 100% for the most frequent 1,000 words to around 40% for the fifth 1,000 words.

The results presented here can strengthen the awareness of vocabulary teaching and learning sections in the textbooks. Also, they can link to the international literature on this specific phenomenon. Future studies need to consider the recycling of words (Nordlund, 2015). In addition, it should not be taken for granted that the repetition of items is evenly distributed across textbook series to strengthen the learning outcomes. It would be interesting for other scholars in the field to look at the information of repetition to see the proportion of vocabulary items that appear at a certain span in specific textbook series.

Although this study has revealed interesting findings about vocabulary use in Chinese tertiary textbooks, it is important to acknowledge its limitations. First, our study examined the number of repetitions to check if

the textbooks have provided enough learning opportunities for students. However, other aspects of vocabulary, such as the dispersion, regularity, length, polysemy, and contextual distinctiveness of single words and collocations (Hashimoto & Egbert, 2019), were not included. Future studies might focus on these aspects to examine the lexical difficulty of vocabulary items in textbooks. Second, our research focused only on reading texts, leaving out the practical exercises. Exposure to exercises might also contribute to the learning of vocabulary items and, more importantly, to understanding how the type of exercises has an impact on the way students learn vocabulary items (e.g., Boers et al., 2014).

Acknowledgments

This was financially supported by University Research Grant 299-GK19G055 from Guangdong University of Foreign Studies, Guang Zhou.

References

Abello-Contesse, C., & López-Jiménez, D. (2010). The treatment of lexical collocations in EFL textbooks. In M. Moreno Jaén, F. Serrano Valverde, & M. Calzada Pérez (Eds.), *Exploring new paths in language pedagogy: Lexis and corpus-based language teaching* (pp. 95–109). Equinox.

Anthony, L. (2020). *AntConc (Version 3.5.9)* [Computer software]. Waseda University.

Bi, J. (2020). How large a vocabulary do Chinese computer science undergraduates need to read English-medium specialist textbooks? *English for Specific Purposes*, *58*, 77–89.

Biber, D., Conrad, S., & Cortes, V. (2004). If you look at…: Lexical bundles in university teaching and textbooks. *Applied Linguistics*, *25*(3), 371–405.

Boers, F., Demecheleer, M., Coxhead, A., & Webb, S. (2014). Gauging the effects of exercises on verb-noun collocations. *Language Teaching Research, 18*(1), 54–74.

Chen, C., & Truscott, J. (2010). The effects of repetition and L1 lexicalization on incidental vocabulary acquisition. *Applied Linguistics, 31(5)*, 693–713.

Cobb, T. (2021). *The Vocabprofilers* [Computer program]. https://www.lextutor.ca/vp/.

Criado, R., & Pérez, A. S. (2009). Vocabulary in EFL textbooks: A contrastive analysis against three corpus-based word ranges. *A Survey of Corpus-Based Research*, 862–875.

Dang, T. N. Y., & Webb, S. (2014). The lexical profile of academic spoken English. *English for Specific Purposes*, *33*(1), 66–76.

Harwood, N. (2014). Content, consumption, and production: Three levels of textbook research. In N. Harwood (Ed.), *English language teaching textbooks: Content, consumption, production* (pp. 1–41). Palgrave Macmillan.

Hashimoto, B. J., & Egbert, J. (2019). More than frequency? Exploring predictors of word difficulty for second language learners. *Language Learning, 69*(4), 839–872.

Hsu, W. (2014). The most frequent opaque formulaic sequences in English-medium college textbooks. *System*, *47*, 146–161.

Hsu, W. (2018). The most frequent BNC/COCA mid- and low-frequency word families in English-medium traditional Chinese medicine (TCM) textbooks. *English for Specific Purposes*, *51*, 98–110.

Hu, M., & Nation, I. S. P. (2000). Unknown vocabulary density and reading comprehension. *Reading in a Foreign Language, 13*(1), 403–430.

Jin, T. A. N., Li, Y. T., & Li, B. C. (2016). Vocabulary coverage of reading tests: Gaps between teaching and testing. *TESOL Quarterly, 50*(4), 955–964.

Laufer, B., & Nation, I. S. P. (1995). Vocabulary size and use: Lexical richness in L2 written production. *Applied linguistics, 16*(3), 307–322.

Laufer, B., & Ravenhorst-Kalovski, G. C. (2010). Lexical threshold revisited: Lexical text coverage, learners' vocabulary size and reading comprehension. *Reading in a Foreign Language*, *22*(1), 15–30.

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

Lei, L., & Liu, D. (2016). A new medical academic word list: A corpus-based study with enhanced methodology. *Journal of English for Academic Purposes*, *22*, 42–53.

Liu, Y. H., & Zhang, J. (2015). A corpus-based study of lexical coverage and density in college English textbooks. *Foreign Language Education in China, 8*(1), 42–50.

Martinez, R., & Schmitt, N. (2012). A phrasal expressions list. *Applied Linguistics*, *33*(3), 299–320.

Matsuoka, W., & Hirsh, D. (2010). Vocabulary learning through reading: Does an ELT course book provide good opportunities? *Reading in a Foreign Language*, *22*(1), 56–70.

Ministry of Education. (2007). *The college English curriculum requirement.* Higher Education Press.

Ministry of Education. (2020). *Guideline for college English education.* Higher Education Press.

Nation, I. S. P. (2004). A study of the most frequent word families in the British National Corpus. In P. Bogaards & B. Laufer (Eds.), *Vocabulary in a second language: Selection, acquisition and testing* (pp. 3–13). Amsterdam: John Benjamins.

Nation, I. S. P. (2005). *Range (version 3.2)* [Computer software]. Victoria University of Wellington.

Nation, I. S. P. (2013). *Learning vocabulary in another language* (2nd ed.). Cambridge University Press.

Nordlund, M. (2015). Vocabulary acquisition and the textbook. *ITL—International Journal of Applied Linguistics*, *166*(2), 199–228.

Nordlund, M. (2016). EFL textbooks for young learners: A comparative analysis of vocabulary. *Education Inquiry*, *7*(1), 47–68.

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.

Peters, E. (2016). Learning German formulaic sequences: The effect of two attention-drawing techniques. *Language Learning Journal, 40*(1)*,* 65–79.

Peters, E., & Webb, S. (2018). Incidental vocabulary acquisition through viewing L2 television and factors that affect learning. *Studies in Second Language Acquisition, 40*(3)*,* 551–577.

Ren, X. H. (2014). A study of lexical chunks in college English textbooks. *Language Education, 2*(3), 41–47.

Schmitt, N., Jiang, X., & Grabe, W. (2011). The percentage of words known in a text and reading comprehension. *Modern Language Journal*, *95*(1), 26–43.

Siyanova-Chanturia, A., & Spina, S. (2015). Investigation of native speaker and second language learner intuition of collocation frequency. *Language Learning*, *65*(3), 533–562.

Stubbs, M. (2001). Texts, corpora, and problems of interpretation: A response to Widdowson. *Applied Linguistics*, *22*(2), 149–172.

Teng, F. (2020). Retention of new words learned incidentally from reading: Word exposure frequency, L1 marginal glosses, and their combination. *Language Teaching Research, 24*(6), 785–812.

Tsai, K.-J. (2015). Profiling the collocation use in ELT textbooks and learner writing. *Language Teaching Research*, *19*(6), 723–740.

Uchihara, T., Webb, S., & Yanagisawa, A. (2019). The effects of repetition on incidental vocabulary learning: A meta-analysis of correlational studies. *Language Learning*, *69*(1), 1–41.

Wang, C. (2019). Profiling collocations in EFL writing of Chinese tertiary learners. *RELC Journal, 50*(1), 53–70.

Wang, C., & Yang, J. J. (2020). Revisiting the explicit learning of vocabulary of Chinese EFL learners. *English Language Teaching, 13*(2), 86–96.

Ward, J., & Chuenjundaeng, J. (2009). A basic engineering English word list for less proficient foundation engineering undergraduates. *English for Specific Purposes*, *28*(3), 170–182.

Webb, S. (2007). The effects of repetition on vocabulary knowledge. *Applied Linguistics*, *28*(1), 46–65.

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

Yamashita, J., & Jiang, N. (2010). L1 influence on the acquisition of L2 collocations: Japanese ESL users and EFL learners acquiring English collocations. *TESOL Quarterly*, *44*(4), 647–668.

Yang, H. (2018). Research on digital college English textbooks in information environment. *Helongjiang Science, 2018*(9), 82–88.

Chen WANG, Ph.D., is currently a Senior Lecturer at Guangdong University of Foreign Studies. Her research interests include vocabulary learning and teaching, and second language learning.

Yuhua LIU is currently a Senior Lecturer at Jiangxi Normal University. She is also a doctoral candidate at Guangdong University of Foreign Studies. Her research interests include language assessment and second language learning.