Frontiers of Information Technology & Electronic Engineering
www.zju.edu.cn/jzus; engineering.cae.cn; www.springerlink.com
ISSN 2095-9184 (print); ISSN 2095-9230 (online)
E-mail: jzus@zju.edu.cn

Corpus-based research on English word recognition rates in primary school and word selection strategy*

Wen-yan XIAO[1,2], Ming-wen WANG‡[1], Zhen WENG[1], Li-lin ZHANG[1], Jia-li ZUO[1]

([1]*School of Computer Information Engineering, Jiangxi Normal University, Nanchang 330022, China*)

([2]*Jiangxi University of Science and Technology, Nanchang 330003, China*)

E-mail: wyxiao@jxnu.edu.cn; mwwang@jxnu.edu.cn; 1091013334@qq.com; 1006806747@qq.com; 44124148@qq.com

Received Apr. 6, 2016; Revision accepted Aug. 23, 2016; Crosschecked Feb. 20, 2017

Abstract: Acquiring vocabulary is important when studying English, as it assists in listening, speaking, reading, and writing. In this paper, we develop an English webpage corpus (EWC) and create a word frequency list using web crawler technology. By comparing EWC word lists with the British National Corpus (BNC), we find that the BNC word frequency list possesses the feature of timeliness. We also explore primary school students' English word recognition rates by comparing the word frequency lists of several corpora, including EWC, BNC, SUBTLEX-US, and Subtitle Corpus of Children's BBC (CBBC). The results show that the word recognition rates for primary school children are relatively low in both general language and specific language register. Motivated by the experiment results, we finally propose some word-selection strategies for compiling English textbooks for Chinese primary school students.

Key words: Corpus; Primary English; Recognition rate; Word frequency; Coverage rate
http://dx.doi.org/10.1631/FITEE.1601118 **CLC number:** H313; TP391

1 Introduction

In 2001, the foreign language teaching research group of Beijing Educational Scientific Academy reported that 85% of information on the Internet was presented in English (BESA, 2001). English plays a significant role in the world. In China, English study begins in the third grade of primary school, and sometimes as early as the first grade. However, many people who have studied English for a decade are unable to communicate in English. Research shows that younger learners perform better than older learners, indicating that studying English in primary school is critical. Lenneberg (1967) also suggested that the best time to learn a language is when children are aged 3 to 12.

Vocabulary, one of the three elements of the English language system, is very important in communication. Learners become more proficient in English as they acquire a larger vocabulary (Liu and Sun, 2013). New vocabulary is acquired mainly through textbooks, so the quality of vocabulary selections in textbooks directly affects learners' abilities to communicate. With the development of social life, new things and concepts make their continuous appearances, which may bring about new words and relevant changes in the vocabulary system. Thus, a special note should also be made about updating the vocabulary in textbooks (Zhao, 2007). Vocabulary in primary English textbooks should keep up with the times and be updated accordingly.

To investigate primary school English word recognition, we developed an English webpage corpus (EWC) and obtained a word frequency list using web crawler technology. Then we compared and analyzed the word lists in primary school English

‡ Corresponding author

* Project supported by the National Natural Science Foundation of China (Nos. 61272212, 61462043, and 61462045) and the Graduates Innovation Fund of Jiangxi Normal University, China

ORCID: Wen-yan XIAO, http://orcid.org/0000-0001-6253-2414

textbooks and the word frequency lists of EWC, British National Corpus (BNC), Subtitle Corpus of Children's BBC (CBBC), and SUBTLEX-US. Motivated by the experiment results, we proposed some word selection strategies and word frequency suggestions for compiling primary school English textbooks, to assist Chinese primary schools with teaching and learning English.

2 Related work

2.1 Textbook vocabulary research

Vocabulary, as the smallest element of language, not only correlates with grammar and phonics but also lays the foundation for language. Wilkins (1972) expressed the importance of vocabulary for communication, noting that without grammar very little can be conveyed, but without vocabulary nothing can be conveyed. Nation (1990) pointed out that the problems in language reception and production can be attributed to a deficiency in vocabulary. Cunningsworth (1995) stated that communication cannot last long without a relevant or relatively large vocabulary. English learners, especially those with limited language proficiency, use lexical knowledge rather than grammatical knowledge to communicate effectively. Vocabulary is key to communication. Moreover, Chinese students acquire English vocabulary mainly through textbooks. Thus, pupils' abilities to communicate in English depend largely on the vocabulary selected for textbooks (Shi, 2015). Since the 1980s, vocabulary has been the main concern when establishing a teaching syllabus and compiling textbooks. Which words should be selected for textbooks? So far, a few scholars have conducted research on this question. White (1998) and Thornbury (2006) proposed criteria for word selection, emphasizing word frequency and semantics. Sun (2005) argued that word selection should be based on objective statistics and large corpora, together with contextual diversity, practicality, usefulness, and pedagogical applicability of the word list. Xie and He (2008) compared word lists from a primary textbook and middle school textbook with the New Curriculum Standard and proposed that corpus tools and research results based on corpus, such as word frequency, semantics frequency, and the typical collocations, improve textbook word lists. Corpus tools and research results also make compiling textbooks more reasonable and consistent in terms of the selection, classification, and annotation of the words. Fu (2013) conducted a comprehensive analysis of textbook vocabulary in terms of the total amount of vocabulary, vocabulary distribution, and the recurrence rate of low- and high-frequency words. He suggested paying attention to high-frequency words when selecting textbook vocabulary, as well as designing individualized word lists. Lang and Li (2009) discussed the concept and importance of common words. They also pointed out that if textbook writers develop and select appropriate lists of common words according to the learners' actual needs, the learners' anxiety will be greatly alleviated.

2.2 Corpus studies

As society develops, language research blends with other disciplines. Computer technology and the Internet have become efficient tools in acquiring large amounts of authentic language materials. Language researchers tend to use corpus to conduct all aspects of research. As Svartvik (1996) pointed out, corpus has become mainstream, providing not only a research method, but also a new philosophical way of thinking. In many branches of linguistics, corpora provides core data for survey research and the development and testing of hypotheses (Rietveld *et al*., 2004). The definitions of corpus differ in literal expressions but are the same in essence. According to Sinclair (1996), a corpus is "a collection of pieces of language that are selected and ordered according to explicit linguistic criteria in order to be used as a sample of the language". Similarly, Kennedy (1998) defined a linguistic corpus as "systematic, planned and structure compilation of text" and a database that has been "designed and structured specifically to be used for linguistic description and analysis". In other words, a corpus is a large collection of texts stored on a computer. Corpus serves as a useful research tool and has solved a lot of language research questions, including ones on vocabulary, which are difficult to answer by other means. Schmitt (2010) said that corpus linguistic analysis is an important vocabulary research method. There has been recent corpus-based vocabulary research on a new medical academic word list with enhanced methodology that can meet the

needs of learning academic medical English (Lei and Liu, 2016). In regard to corpus-based textbook vocabulary research, Wang and Wu (2008) analyzed three different primary school English textbooks and explored the use of large-scale balanced corpora to write English textbooks. They stated that vocabulary selection should be based on word frequency, and took into consideration the contextual diversity and usefulness of words, students' cognitive abilities, cultural differences, and national conditions. According to their work, corpus provides a scientific basis for textbook vocabulary selection and sequence arrangement of selected words. Zhang and Ma (2007) employed Range and WordSmith software to analyze the word distribution in the English textbook *Go for It*, and their work shed light on teaching and learning English vocabulary in primary schools.

The following are definitions of important terms that should be made clear:

Token: the total number of words in a text or corpus, regardless of how often they are repeated.

Type: the number of distinct words in a text or corpus. For example, the sentence "Rose is a rose is a rose is a rose" contains 10 tokens, but only 3 types (i.e., rose, is, a) (Liang *et al*., 2010).

Word frequency: the number of times each type occurs, which affects all aspects of lexical processing and acquisition (Schmitt, 2010). Word frequency is arguably the most important variable in word recognition research (Brysbaert *et al*., 2011).

Leech (2001) claimed that frequency plays a role in determining priorities in teaching English. Nation (2001) suggested that learners generally acquire vocabulary more frequently than lexis. It has also been proved that high-frequency words are perceived and produced more quickly and efficiently than low-frequency words (Monsell *et al.*, 1989; Jescheniak and Levelt, 1994; Brysbaert and New, 2009). Hence, primary school students should learn high-frequency words first, and they should be the first words considered for inclusion in textbooks.

To sum up, based on current textbook vocabulary research, we find that corpus-based vocabulary study has become a popular research field. However, since most studies were conducted using the BNC and its word frequency list, it seems that little attempt has been made to connect the webpage corpus with SUBTLEX-US, which contains the latest language data and can authentically reflect the language that people use in daily life. Meanwhile, there is very little macro-level study of China's primary school students' word recognition rates. Therefore, in this study, we focus mainly on primary school students' English word recognition rates on the basis of a large webpage corpus. For the language register, we also make use of the CBBC word frequency list (van Heuven *et al*., 2014), which is extracted from CBBC channel (for primary school children) subtitles. First, we use web crawler technology to develop an English webpage corpus and extract a word frequency list. Then we conduct comparisons between the word frequency lists of EWC and BNC. Furthermore, we explore primary school students' word recognition rates based on EWC, BNC, CBBC, and SUBTLEX-US word frequency lists. Finally, we propose word selection strategies for compiling English textbooks for Chinese primary schools.

3 Methods

3.1 Corpora used in the study

3.1.1 English textbooks used in primary school

In this study, the compulsory education textbooks (grades 3–6) published by the People's Education Publishing House, China are used as the corpus for primary school English textbooks. This set of textbooks is one of the most widely used English teaching materials in China. The textbooks meet the five requirements of learning proposed by the New Curriculum Standard: language skills, language knowledge, emotional attitudes, learning strategies, and cultural awareness (Liu, 2014). We used the following procedure to analyze the word lists and obtained the total vocabulary in the textbooks. We first dealt with each word list using Lemmatizer for such words as desks and swimming, and then deleted the repeated words to form the word list of each textbook. Finally, we filtered the repeated words in the four word lists to establish the total primary school English word list. Some high-frequency words in the 3–6 grade textbooks' word lists occurred more than once, not only because they have several semantic meanings, but also because repetition is conducive to word acquisition. For instance, the word 'teach' appears in every grade's textbook word list, as is the word 'right',

which has two semantic meanings, i.e., 'correct' and 'not left'. Thus, after the second deletion of repeated words, there were 726 words in the final word list. Detailed information of each grade's word list is presented in Table 1.

Table 1 Statistics of each grade's word list

Grade	Number of words	Number of words after lemmatization and duplication deletion
3	141	133
4	208	194
5	307	267
6	247	228
Total	903	822

3.1.2 English webpage corpus

The text data used on the Internet, to a large degree, represent the language people use in daily communication. Take Internet forums and chat rooms, for example. Internet users participate in discussions on a variety of topics without much supervision or editing in a manner that everyday language exposure is approximated (Brysbaert and New, 2009). Therefore, we select three main and popular websites as the corpus sources of our EWC. They are: Delphi Forums (one of the most vibrant forums, serving more than 4 million registered members and 200 million total messages), BBC Network, and America Online (AOL, the primary English language web portal, covering diverse categories of news, sports, weather, earth, lifestyle, art, tourism, nature, culture, business, entertainment, etc.). In developing the EWC, we first used web crawler technology to automatically collect all of the BBC and AOL text data from Jan. 2016, and the Delphi Forums' data from June 2016. Then we cleaned the collected data by deleting non-English characters, such as punctuation, numbers, and irregular characters, eliminating all the person names and lemmatized words with various grammatical forms into their basic forms. For instance, words like 'continues', 'continued', and 'continuing' were lemmatized into the base form 'continue'. Also, markup language not related to language communication was eliminated (e.g., the words 'html' and 'http'). After developing the EWC, we calculated Token and Type, as shown in Table 2, and generated the EWC word list on the basis of frequency.

Table 2 EWC statistics

Number of valid text	Token	Type
63 850	29 918 009	72 571

3.1.3 SUBTLEX-US corpus and CBBC word frequency list

Brysbaert and New (2009) made a critical evaluation of current word frequency norms and introduced a new and improved word frequency measure for American English. As for the language register, they found that frequencies based on television and film subtitles are better than frequencies based on written sources, certainly for the monosyllabic and bisyllabic words used in psycholinguistic research. In their work, a subtitle corpus for American English (i.e., SUBTLEX-US) was assembled by using television and film subtitles (http://subtlexus.lexique.org). Following this work, van Heuven *et al*. (2014) presented word frequency based on subtitles of British television programs (SUBTLEX-UK), including word frequencies in children's programs based on the subtitles of the CBBC channel for primary school children (aged 6–12) (http://www.psychology.nottingham.ac.uk/subtlex-uk/). We downloaded the SUBTLEX-US and SUBTLEX-UK (where the frequency counts in the CBBC subtitles are included). The statistics of SUBTLEX-US and CBBC are summarized in Table 3.

Table 3 SUBTLEX-US and CBBC statistics

Corpus	Token	Type
SUBTLEX-US	43 817 894	54 967
CBBC	11 814 733	45 143

3.1.4 BNC word frequency list

The BNC (http://www.natcorp.ox.ac.uk) is a 100-million-word collection of samples of written and spoken language from a wide range of resources, designed to represent a wide cross-section of British English, both spoken and written, from the late 20th century. It is one of the most representative contemporary English corpora, available on the Internet after free registration. The BNC word frequency list used in this study is a lemmatized 6318-word frequency list (Kilgarriff, 1995). The list creation process replicated

that used at Longman for making dictionary frequencies in LDOCE 3rd Edition (Kilgarriff, 1997). A study by Nation and Waring (1997) found that 2000 word families cover approximately 80% of written text, and the 3000 most frequently used word families represent 84% coverage. Therefore, we can infer that the goal of acquiring 2000–3000 words would ensure the basis for language use. Moreover, according to Piaget's theory of cognitive development, primary school children (aged 7–11) generally can master 2500 common words. Therefore, we focus our experiments on the analysis of the top 3000 words in the BNC list, which would be more targeted and realistic.

3.2 Relevant indexes and terms

Using statistical indexes in Table 4, we analyzed the corpora used in this study to describe and compare word lists from multiple perspectives (Appendix shows the results for the first 20 words in EWC word frequency list).

4 Experiments

4.1 Experiment settings

In this study, we designed three experiments to determine the frequency of commonly used words and the primary school students' English word recognition rates on the basis of EWC, BNC, CBBC, and SUBTLEX-US word frequency lists:

1. Analysis of the EWC word frequency list: A list of the most frequently used 3000 words was generated using the EWC word frequency and contextual diversity rates. Then the coverage rates of the top 1000 words, the top 2000 words, and the top 3000 words were calculated, respectively.

2. Comparison of the word frequency list of EWC and BNC: By calculating the coverage rates of the top 2000 and 3000 EWC words, we figured out which words in the BNC word frequency list are not as commonly used as they once were. We found out which high-frequency words in the EWC word frequency list did not appear in the BNC frequency word list. Changes in the frequency of words commonly used in daily life can be calculated using the Jaccard coefficient of the EWC and BNC word frequency lists.

3. EWC, SUBTLEX-US, CBBC, and BNC word frequency lists comparisons and comparisons of word lists from each grade's textbook: Using statistics from the four grades' vocabulary lists, we calculated the word recognition and increase rates for each grade based on the four above-mentioned word frequency lists. We determined which words used in real-life language deserve more attention, and which words in the textbooks are not often used in the frequency word lists.

Table 4 Statistical indexes

Index	Symbol or formula	Description
Number of valid texts	#(Texts)	The number of valid texts collected by the web crawler
Token	#(Tokens)	The number of tokens in the corpus
Type	#(Types)	The number of types in the corpus
Number of times	t_w	The number of times a certain word appears in the corpus
Number of texts	d_w	The number of texts in the corpus, including a certain word
Word frequency: tf_w	$\frac{t_w}{\#(\mathrm{Tokens})}$	The percentage of the occurrences of a certain word in the corpus
Contextual diversity: df_w	$\frac{d_w}{\#(\mathrm{Texts})}$	The rate of the texts where a certain word occurs. The more the texts containing a certain word, the higher the contextual diversity is
Coverage rate: c_w	$\frac{\sum_{i=1}^{N} t_w}{\#(\mathrm{Tokens})}$	Used to examine the coverage of the top N words in the corpus
The word recognition rate of grade's textbook word list in the corpus: kf_d	$\frac{\sum_{i=3}^{d}\sum_{w} t_w}{\#(\mathrm{Tokens})}$	To calculate the recognition rate in the dth grade ($3\le d\le 6$)
Jaccard coefficient	$J(A,B)=\frcl{\lvert A\cap B\rvert}{\lvert A\cup B\rvert}$	Used to obtain the degree of similarity between two word lists

4.2 Experiment results and analysis

We obtain the following results using the experiments mentioned above.

4.2.1 EWC word list analysis

We analyzed EWC on a statistical basis, computed the frequency of its words, and generated a word frequency list. Then we computed the coverage rates of the top 1000, 2000, and 3000 words, which were 80.64%, 87.40%, and 90.46%, respectively. The results showed that the top 3000 words making up 90.46% of all word occurrences were used frequently in daily life. They should be considered in the selection of target words for vocabulary studies. Additionally, the coverage rate of the top 2000 words was consistent with the finding of Nation and Waring (1997), i.e., 80%, which to some extent indicates that EWC is representative and can reflect the language that people use in their daily lives.

4.2.2 Comparison of EWC and BNC word frequency lists

We can see the similarities and differences in EWC and BNC word frequency lists by looking at the top 3000 words in each list. Table 5 shows the word coverage rates of the two lists. The results of the comparison are illustrated in Fig. 1 and Table 6. Let A be the set of top 3000 words of the EWC word frequency list, A_1 the set of top 2000 words, and A_2 the set of the third 1000 words (i.e., the words ranging from 2001 to 3000). Let B be the set of top 3000 words of the BNC word frequency list, B_1 the set of top 2000 words, and B_2 the set of the third 1000 words.

The experiment results indicated that the coverage rates of the top 1000, 2000, and 3000 words of the BNC word frequency list are lower than those of the EWC word frequency list. To a certain degree, it showed that some high-frequency words in the BNC word list are not as commonly used as they once were. As indicated in Table 6, the intersection of the top 2000 words in the EWC word frequency list (A_1) and

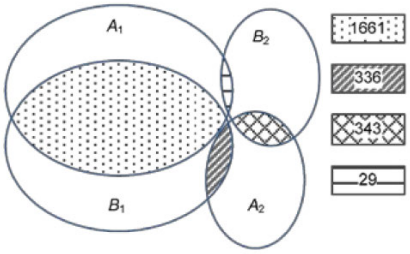

Fig. 1 Comparison of EWC and BNC word frequency lists

Table 6 Intersection of EWC and BNC word frequency lists

Intersection (∩)	A	A_1	A_2
B	2369	1690	679
B_1	1997	1661	336
B_2	372	29	343

the top 2000 words in the BNC word frequency list (B_1) is 1661, and the intersection of A_2 and B is 679. The intersections demonstrate that 310 words of the top 2000 words in the EWC word frequency list are absent from the top 3000 words in the BNC word frequency list, and 321 words of the third 1000 words are not present in the top 3000 words in the BNC word frequency list. In other words, 310 words with high frequency have been newly added in the top 2000 words of the EWC word frequency list, and 321 in the words in 2001–3000. Moreover, 1997 words in the top 2000 words on the BNC word frequency list are included in the top 3000 words of the EWC frequency word list, and there are 372 words in B_2 occurring on the set of A. This points to the fact that three words on the top 2000 of the BNC frequency list and 628 words in the word set from the top 2001–3000 of the BNC word frequency list are not used as much in daily life as they once were.

Statistical analysis reveals that the intersection of the top 3000 words on EWC and BNC word frequency lists is 2369 and the union is 3627. The Jaccard coefficient of the top 3000 words on the EWC and BNC word frequency lists can be computed according to the formula, which is 0.6531. This result not only indicates that there are some similarities

Table 5 The coverage rate of EWC and BNC word frequency lists

Word frequency list	Coverage rate		
	Top 1000 words	Top 2000 words	Top 3000 words
EWC	80.64%	87.40%	90.46%
BNC	73.66%	81.16%	85.85%

between the EWC and BNC word frequency lists, but also reveals the differences between them. From the perspective of language dynamics, some words on the BNC word frequency list are less frequently used in daily life than they once were, and a number of words emerge as new high-frequency words in language communication.

We looked at the register and spelling of the top 3000 EWC and BNC words to further examine the reasons for their differences. First, for spelling differences, we compared the top 3000 EWC words with the Comprehensive List of American and British Spelling Differences (http://www.tysto.com/uk-us-spelling-list.html) and found that 38 words follow American spelling rather than British spelling, such as apologize, behavior, center, check, color, favor, favorite, gray, honor, kilometer, labor, organize, practice, program, realize, and theater. None of the EWC top 3000 words follows British spelling, although EWC contains words of both American and British spellings. This finding also reveals that the American way of spelling is now more frequently adopted in present daily English usage and far more popular. For example, in EWC, the t_w of 'color' is 1142, while that of 'colour' is 148; the t_w of 'meter' is 1397, while that of 'metre' is 78. Moreover, unlike EWC, we found that all the words in BNC follow British spelling, contributing to the differences between EWC and BNC.

Second, since register differences exist between EWC and BNC, the top 3000 words in each corpus are likely to vary too. Halliday and Hasan (1976) interpreted register to be "the linguistic features which are typically associated with a configuration of situational features—with particular values of the field, mode and tenor". From a narrow point of view, register is limited only to the field of discourse, such as the language varieties used in different professions: law, news, medicine (Wardhaugh, 1972; Spolsky, 1998; Verschueren, 1999; Trudgill, 2000). More generally, register is now used to indicate degrees of formality in use. For instance, 90% of BNC is samples of written language use and only 10% is samples of spoken language use. Thus, BNC input is much more formal and standard than EWC, which approximates the language people use in daily life. Unfortunately, register covers such broad dimensions that it would be an arduous task to compare each of them and figure out to what extent the differences in the two corpora are caused by register. Nevertheless, it can be observed that differences between EWC and BNC are related to register.

The fact that the Jaccard coefficient of the top 3000 EWC and BNC words is relatively low indicates that the differences between the two corpora are related to not only time but also spelling and register.

4.2.3 Comparisons of EWC, BNC, SUBTLEX-US, and CBBC word frequency lists and word lists in each grade's textbook

Through comparison and analysis of EWC, BNC, SUBTLEX-US, and CBBC word frequency lists and word lists in each grade's textbook, we can obtain the word recognition and increase rates for each grade, provided that all textbook word lists are mastered. The results are shown in Table 7. The experiment data are analyzed as follows:

1. Compared with the four word frequency lists, the recognition rate of each grade grows over the grade, which demonstrates that pupils are learning more and more words. The final word recognition rates after four-year study are, respectively, 50.33%, 58.94%, 51.34%, and 53.98%, which on the whole indicates that textbook words may lay a solid foundation for primary school students learning English. As shown in Table 7, the BNC word recognition rate is obviously the highest, followed by the CBBC recognition rate, while SUBTLEX-US and EWC recognition rates are lower than the other two. The different recognition rates can be partly attributed to

Table 7 Number of words in each grade and word recognition and increase rates

Grade	Number of words	Word recognition rate				Increased rate of word recognition			
		EWC	BNC	SUBTLEX-US	CBBC	EWC	BNC	SUBTLEX-US	CBBC
3	133	10.13%	9.11%	9.40%	10.53%	–	–	–	–
4	194	28.87%	30.80%	28.61%	30.68%	18.74%	21.69%	19.21%	20.15%
5	267	46.27%	54.00%	46.20%	49.32%	17.37%	23.20%	17.59%	18.64%
6	228	50.33%	58.94%	51.34%	53.98%	4.09%	4.94%	5.14%	4.66%

the register and words spelling differences between each corpus and English textbooks, although some other potential factors may affect the results. Table 8 shows the estimated differences of the four corpora in terms of their register and spelling. Specifically, BNC spelling and register are more similar to those of the textbooks, as they each contain formal and informal register and follow British spelling. Moreover, China's administrators and teachers are apt to promote standard English teaching and learning and select more formal language input. Correspondingly, the register of English textbooks tends to be more similar to that of BNC, leading to the relatively high recognition rate. Further scrutiny of the textbook word lists revealed that they all follow British spelling, such as mum, metre, favourite, kilogramme, and labour, instead of their corresponding American spellings (mom, meter, favorite, kilogram, labor). This finding partly provides evidence for the higher recognition rate of BNC and CBBC, since their language input follows British spelling. It also explains the lower recognition rate of SUBTLEX-US and EWC, as they mostly use American spelling. In addition, the relatively high CBBC word recognition rate is partly because the CBBC list contains words from the CBBC channel for primary school children, which tends to be more similar to the language register of child interaction.

Generally, the results of the comparisons reveal that the textbooks' vocabulary cannot fully meet the requirements of online communication or understanding channel and film subtitles. With the difficulties caused by grammar and cultural differences, we can infer that the words learned in primary school can hardly play a proper role in language communication.

2. The amount of English words increases from grade 3 to grade 5, but decreases in grade 6, with an increase rate of only 4.09%, 4.94%, 5.14%, and 4.66%, respectively. Compared with grade 5, the word recognition rate of grade 6 does not improve much, which is not consistent with students' growing cognitive ability and communicative needs at ages 11 and 12. According to the law of pupils' mental and cognitive development, 6th grade students should be at a stage of having relatively strong adaptability and learning ability. Therefore, given certain conditions, we suggest that vocabulary input should be expanded in the 6th grade.

3. To further examine primary school textbooks' word selections, we obtain the intersection of the top 3000 words on the four word frequency lists, i.e., find out the words simultaneously occurring in all of them. Then we filter out the words not occurring in the textbook word lists by comparing the intersection words with the textbook words. A total of 903 words with high frequency do not appear in the textbook vocabulary. Some function words like you, it, what, not, they, from, me, and can, rank in top 10 in the EWC, BNC, SUBTLEX-US, and CBBC intersection word lists. According to the statistics, there are fewer than 400 function words, but they account for more than half of the vocabulary used in daily life (Zhu, 2013). Besides, some content words, like time, life, put, focus, place, night, run, and company, which are commonly used in real life, are not included in the textbook word list. Hence, we suggest that some of those 903 words be included in the textbook vocabulary lists, taking word frequency and contextual diversity into consideration. Doing so can meet the requirements of rapid change in language. The top 20 common words can be seen in Table 9 after ranking the common words according to their total times of occurrence. Through further analysis, we also found that three words in the textbook word list do not exist in any word frequency list. They are turpan, mooncake, and mid-autumn. To further investigate their use in daily life, we use the BCC corpus (http://bcc.blcu.edu.cn) to retrieve these three words. We found that mooncake occurs once, mid-autumn 14, and turpan 0. From this result, we can conclude that these three words are seldom used in English-speaking countries. However, it is not difficult to notice that they have a lot to do with Chinese culture, with apparent Chinese features. Thus, these words should still be selected and learned by Chinese pupils.

Table 8 Comparison of the four corpora and primary school English textbooks' register and spelling

Feature	BNC	CBBC	SUBTLEX-US	EWC	Primary school English textbooks
Register (formal/informal)	Both	Informal	Informal	Both	Both
Spelling (British/American)	British	British	Both	Both	British

Table 9 The intersection words not occurring in textbook word lists ranked by t_w

Word	t_w				Sum
	EWC	BNC	SUBTLEX-US	CBBC	
you	135 807	695 498	2 134 713	372 185	3 338 203
it	208 847	1 090 186	963 712	296 718	2 559 463
with	314 074	675 027	257 465	69 662	1 316 228
what	44 569	249 466	501 965	98 665	894 665
not	76 531	465 486	276 673	62 865	881 555
they	89 536	433 441	209 250	92 087	824 314
from	200 970	434 532	103 992	32 607	772 101
me	18 320	138 151	471 339	63 872	691 682
can	46 019	266 116	267 620	61 863	641 618
out	53 882	201 819	197 131	50 842	503 674
who	114 503	205 432	113 370	26 104	459 409
would	45 135	272 345	90 162	23 772	431 414
as	156 592	101 583	113 068	42 815	414 058
some	96 702	171 174	88 089	25 529	381 494
time	45 939	183 427	99 890	38 621	367 877
here	17 931	70 947	230 788	45 558	365 224
then	17 321	160 652	75 966	22 491	276 430
yeah	732	83 382	152 262	36 944	273 320
into	33 596	163 469	43 074	17 857	257 996
where	57 434	44 496	93 341	18 493	213 764

Meanwhile, we suggest that the word 'twelfth' should be eliminated from primary school English textbooks due to its low frequency in the four corpora we used.

By analyzing the experiment results, we can summarize the findings as follows:

1. By comparison, we find that the coverage rate of the BNC word frequency list is lower than that of the EWC word frequency list, and the Jaccard coefficient of the two lists is not very high. This result indicates that some words on the BNC word frequency list are not as commonly used as they once were. Therefore, as a critical basis of textbook compilation, the word frequency list should be revised at the same time to keep up with the development of social life and dynamic change in language.

2. The word lists contained in the compulsory education primary school English textbooks (grades 3–6) published by the People's Education Publishing House are relatively small, with limited breadth and low EWC, BNC, SUBTLEX-US, and CBBC recognition rates. Primary school students can hardly communicate in English after four years of English study. Therefore, to cultivate their English communication ability, first, some adjustment and updating of vocabulary should be made when designing textbooks, and certain high-frequency words should be added, if possible, while some relatively low-frequency words should be eliminated. Moreover, appropriate extracurricular English reading material and English programs and films should be offered, especially for primary school children. Meanwhile, more attention should be paid to both American and British English spelling when selecting words for teaching and learning. Finally, for each grade's textbook, especially the 6th grade, word selections should cater to students' physical and mental characteristics and their communicative needs to improve their word recognition rate.

5 Conclusions

The aim of this paper is to investigate primary school English recognition rates and provide textbook compilers with some word selection strategies for primary school English textbooks. We employed web crawler technology to build a corpus based on English websites and generated a word frequency list.

Through experiments, we compared EWC and BNC word lists based on primary school English recognition rates and analyzed EWC, BNC, SUBTLEX-US, and CBBC word frequency lists. The overall results indicated two noteworthy findings, which provide a basis for textbook word selections and modifications. First, words on the BNC frequency word list possess the feature of timeliness. Second, the primary school English word recognition rate is relatively low, from the perspectives of both general language and specific language register for primary school children.

Note that different corpora can result in different word frequencies due to the limitations of corpora. The web corpus used in this study was derived chiefly from the BBC, AOL, and Delphi websites, and this study centered around the word lists of textbooks. Therefore, in future work, we will focus on the following three aspects, the perfection of the corpus by enriching the data source and extending the time span for data collection, the extraction and analysis of the formulaic language in English textbooks used by Chinese primary schools, and investigation of the extent to which register and spelling differences cause different word recognition rates.

Acknowledgements

We thank the reviewers for their helpful comments and suggestions. We also appreciate Xiong-fei XU and Zhi-ming CHEN for assistance with the experiments, and Dr. Jin-shan ZENG and Dr. Hong-wei LI for valuable discussion.

References

Beijing Educational Scientific Academy (BESA), 2001. Reflections on the current evaluation system of foreign language teaching and the importance of formative assessment in foreign language teaching, *For. Lang. Teach. Schools (Middle Vers.)*, **24**(6):1-4 (in Chinese).

Brysbaert, M., New, B., 2009. Moving beyond Kučera and Francis: a critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Beh. Res. Meth.*, **41**(4):977-990. http://dx.doi.org/10.3758/BRM.41.4.977

Brysbaert, M., Buchmeier, M., Conrad, M., *et al.*, 2011. The word frequency effect: a review of recent developments and implications for the choice of frequency estimates in German. *Exp. Psychol.*, **58**(5):412-424. http://dx.doi.org/10.1027/1618-3169/a000123

Cunningsworth, A., 1995. Choosing Your Coursebook. Heinemann Publishers, Oxford.

Fu, Y.C., 2013. A Vocabulary Study in Textbooks for Primary School and Junior High School Students. MS Thesis, Nanjing Normal University, China (in Chinese).

Halliday, M.A.K., Hasan, R., 1976. Cohesion in English. Longman, London, UK.

Jescheniak, J.D., Levelt, W.J.M., 1994. Word frequency effects in speech production: retrieval of syntactic information and of phonological form. *J. Exp. Psychol. Learn. Mem. Cogn.*, **20**:824-843.

Kennedy, G., 1998. An Introduction to Corpus Linguistics. Longman.

Kilgarriff, A., 1995. BNC database and word frequency lists. http://www.kilgarriff.co.uk/bnc-readme.html

Kilgarriff, A., 1997. Putting frequencies in the dictionary. *Int. J. Lexicogr.*, **10**(2):135-155. http://dx.doi.org/10.1093/ijl/10.2.135

Lang, J.G., Li, J., 2009. On English frequent words and frequency annotations of four English learners' dictionaries. *For. Lang. Teach. Res.*, **42**(1):61-66 (in Chinese).

Leech, G., 2001. The role of frequency in ELT: new corpus evidence brings a re-appraisal. *For. Lang. Teach. Res.*, **33**(5):328-339.

Lei, L., Liu, D.L., 2016. A new medical academic word list: a corpus-based study with enhanced methodology. *J. Engl Acad. Purp.*, **22**:42-53. http://dx.doi.org/10.1016/j.jeap.2016.01.008

Lenneberg, E.H., 1967. Biological Foundations of Language. Wiley, New York.

Liang, M.C., Li, W.Z., Xu, J.J., 2010. Using Corpora: a Practical Coursebook. Foreign Language Teaching and Research Press, Beijing.

Liu, L., 2014. An Evaluation of 2012 PEP Primary English. MS Thesis, Ludong University, China (in Chinese).

Liu, X.C., Sun, Y.J., 2013. Research on correlation of English vocabulary class information processing and comprehensive English ability. *Inform. Sci.*, **34**(7):64-67 (in Chinese).

Monsell, S., Doyle, M.C., Haggard, P.N., 1989. Effects of frequency on visual word recognition tasks: where are they? *J. Exp. Psychol. Gen.*, **118**(1):43-71. http://dx.doi.org/10.1037/0096-3445.118.1.43

Nation, I.S.P., 1990. Teaching and Learning Vocabulary. Heinle ELT, London, UK.

Nation, I.S.P, 2001. Learning Vocabulary in Another Language. Cambridge University Press, London, UK.

Nation, P., Waring, R., 1997. Vocabulary size, text coverage and word lists. *In*: Schmitt, N., McCarthy, M. (Eds.), Vocabulary: Description, Acquisition and Pedagogy. Cambridge University Press, London, UK.

Rietveld, T., van Hout, R., Ernestus, M., 2004. Pitfalls in corpus research. *Comput. Human.*, **38**(4):343-362. http://dx.doi.org/10.1007/s10579-004-1919-1

Schmitt, N., 2010. Researching Vocabulary: a Vocabulary Research Manual. Palgrave MacMillan.

Shi, S.T., 2015. A Research on the Vocabulary Setting of Primary School Textbooks and the Students' Communicative Competence. MS Thesis, Shanghai Normal Uni-

versity, China (in Chinese).
Sinclair, J., 1996. Preliminary Recommendations on Corpus Typology. EAGLES Document TCWG-CTYP/P. http://www.ilc.cnr.it/EAGLES96/corpustyp/corpustyp.html
Spolsky, B., 1998. Sociolinguistics. Oxford University Press.
Sun, W.K., 2005. On the compilation principles and methods of English teaching vocabulary syllabus of basic education—with comments on vocabulary syllabus of basic education. *Curricul. Teach. Mat. Meth.*, **25**(3):61-65 (in Chinese).
Svartvik, J., 1996. Corpora are becoming mainstream. *In*: Thomas, J., Short, M. (Eds.), Using Corpora for Language Research. Longman, London, p.3-13.
Thornbury, S., 2006. How to Teach Vocabulary. Pearson Education, India.
Trudgill, P., 2000. Sociolinguistics: an Introduction to Language and Society. Penguin, UK.
van Heuven, W.J.B., Mandera, P., Keuleers, E., *et al*., 2014. SUBTLEX-UK: a new and improved word frequency database for British English. *Q. J. Exp. Psychol.*, **67**(6): 1176-1190. http://dx.doi.org/10.1080/17470218.2013.850521
Verschueren, J., 1999. Understanding Pragmatics. Oxford University Press.
Wang, Z.Q., Wu, X., 2008. The construction of corpus of English textbooks in China and its application in primary English textbook writing. *Curricul. Teach. Mat. Meth.*, **6**:53-57 (in Chinese).
Wardhaugh, R., 1972. Introduction to Linguistics. McGraw-Hill, New York.
White, R., 1998. The ELT Curriculum: Design, Innovation and Mangement. Wiley-Blackwell.
Wilkins, D.A., 1972. Linguistics in Language Teaching. PhD Thesis, Edward Amold, London.
Xie, J.C., He, A.P., 2008. A study on the appendix vocabulary of middle school English textbooks. *For. Lang. Teach. Schools (Middle Vers.)*, **31**(9):1-5 (in Chinese).
Zhang, W., Ma, G.H., 2007. Analysis on the vocabulary of *Go for It*. *For. Lang. Teach. Schools (Middle Vers.)*, **30**(1): 9-13 (in Chinese).
Zhao, X.B., 2007. A Study on Recognition and Extraction Method of Contemporary Chinese Basic Vocabulary Based on Dynamic Circuit Corpus. PhD Thesis, Beijing Language and Culture University, China (in Chinese).
Zhu, X.M., 2013. A Study of Second Language Function Words Acquisition Based on Attention Theory. MS Thesis, Sichuan International Studies University, China (in Chinese).

Appendix: Statistical results of the first 20 words in the EWC word frequency list

Table A1 Statistical results of the first 20 words in the EWC word frequency list

Rank	Word	t_w	d_w	tf_w	df_w
1	the	1 350 902	63 055	0.0451	0.9875
2	and	922 603	62 677	0.0308	0.9816
3	to	846 416	62 710	0.0282	0.9821
4	be	784 486	62 743	0.0262	0.9826
5	a	773 315	63 435	0.0258	0.9935
6	of	612 463	62 191	0.0204	0.9740
7	in	574 539	62 241	0.0192	0.9748
8	for	374 803	58 869	0.0125	0.9219
9	that	329 155	59 151	0.0110	0.9264
10	on	322 042	59 892	0.0107	0.9380
11	with	314 074	57 112	0.0104	0.8944
12	have	271 996	49 239	0.0090	0.7711
13	say	254 593	45 199	0.0085	0.7078
14	it	208 847	55 472	0.0069	0.8687
15	from	200 970	53 202	0.0067	0.8332
16	this	167 817	49 117	0.0056	0.7692
17	he	159 797	30 535	0.0053	0.4782
18	as	156 592	51 778	0.0052	0.8109
19	cover	140 637	20 051	0.0047	0.3140
20	people	136 874	37 498	0.0045	0.5872