Wen-yan Xiao, Ming-wen Wang, Zhen Weng, Li-lin Zhang, Jia-li Zuo, 2017. Corpus-base research on English word recognition rates in primary school and word selection strategy. *Frontiers of Information Technology and Electronic Engineering*, **18**(3):362-372.
http://dx.doi.org/10.1631/FITEE1601118

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

Key words: Corpus; Primary English; Recognition rate; Word frequency; Coverage rate

Contact: Wen-yan Xiao
E-mail: wyxiao@jxnu.edu.cn
ORCID: http://orcid.org/0000-0001-6253-2414

Introduction

- Acquiring vocabulary is important when studying English, as it assists in listening, speaking, reading, and writing.
- The development of social life may bring about new words and relevant changes in vocabulary system. Thus, vocabulary in primary English textbooks should be updated accordingly.
- To investigate primary school English word recognition, we develop an English webpage corpus (EWC) and compare the word lists in primary school English textbooks with the word frequency lists of several corpora.

Corpora statistics

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

Table 2 EWC statistics

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

Table 3 SUBTLEX-US and CBBC statistics

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

Relevant indexes

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	$J(A,B)=\frac{\|A\cap B\|}{\|A\cup B\|}$	Used to obtain the degree of similarity between two word lists

Comparison results (1)

- This result not only indicates that there are some similarities between the EWC and BNC word frequency list, but also reveals the differences between them.

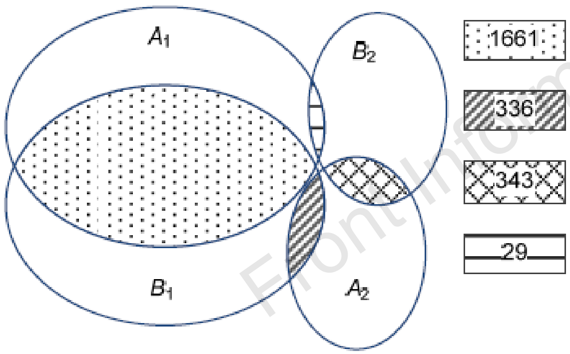

Fig. 1 Comparison of EWC and BNC word frequency lists

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

Comparison results (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.

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%

Conclusions

- Words on the BNC frequency word list possess the feature of timeliness.
- The primary school English word recognition rate is relatively low, not only from the perspective of general language, but also the specific language register for primary school children.