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

Correspondence:

Zipfian interpretation of textbook vocabulary lists: comments on Xiao *et al.*'s *Corpus-based research on English word recognition rates in primary school and word selection strategy*[*]

Qiong HU, Ming YUE[†‡]
(*Department of Linguistics, Zhejiang University, China*)
[†]E-mail: yueming@zju.edu.cn

http://dx.doi.org/10.1631/FITEE.1700418

Now that the latest technology can process huge amounts of data that was previously unimaginable, scientists can challenge established beliefs and practices in many information-related fields. Xiao *et al*. (2017) made such a challenge, focusing on the effectiveness of English textbooks popularly used in Chinese primary schools. They first assumed that pupils' word recognition rate would equal the coverage rate of their textbook vocabulary lists. Then they used four reference corpora to calculate the latter, one of which was self-developed with an automatic web crawler. Finally, they concluded that textbook vocabulary lists were limited in timeliness, that the word recognition increment of the 6th graders was relatively small, and that word selection in textbook compilation should be adjusted. We would like to comment on their study from a Zipfian perspective as applied linguists in language acquisition.

While we agree that traditional word frequency lists should be regularly reevaluated with the help of the latest information technology, we hold the following five points: (1) Sampling issues are important when constructing reference corpora; (2) Zipf's law can provide evidential support for interpreting word frequencies and vocabulary list analysis; (3) Various practical constraints should be considered if vocabulary size in textbooks should be expanded; (4) The word *twelfth* should be kept in the list; (5) Joint attention should be given by scholars of various backgrounds for textbook compilation.

1 Frequency list and textbook compilation

Not all words are created equal in a language. One measurement of the usefulness of a word (type) is its frequency rank in a reference corpus; i.e., how often the word is used in representative samples of the language (Nation and Waring, 1997). If learners acquire new words in the order of the frequency rank, their recognition rates would increase most efficiently, as high frequency brings high coverage increment in the language.

Frequency lists have long been applied in pedagogy (Feng, 1998). Like dictionaries, they also need regular updates, since language never stops changing, with novel words coming and going as the most obvious phenomenon. One recent example is the word *Brexit*, which came into frequent use in 2016 after the British had voted to leave the EU.

2 Reference corpora: sampling issues

Researchers have traditionally used large corpora to evaluate the effectiveness of textbook vocabulary (Chujo, 2004). Xiao *et al*. (2017)'s attempt was interesting in that they developed their EWC corpus (30M tokens) with selected website texts published in 2016. The corpus was taken as a useful resource to observe the latest status of high frequency words in English. In fact, given the availability of big data

‡ Corresponding author
* Project supported by the Fundamental Research Funds for the Central Universities, China (No. 172220173)
ORCID: Ming YUE, http://orcid.org/0000-0001-6457-8176

technology, the size of EWC could have been much larger.

Corpus size and register variety are important for retrieving a representative word frequency list (Nation and Waring, 1997). Most existing corpora, however, are composed mainly of written texts, due mainly to the prohibitive costs of speech transcription when they were created. BNC (100M tokens), for instance, contains oral materials of only about 10%. Consequently, certain words common in oral communication could be missing at the top of the list. Xiao *et al*. (2017) effectively made up for this by introducing SUBTLEX-US (42M tokens) (Nation and Waring, 1997) and CBBC (i.e., Children's BBC, a sub-corpus of SUBTLEX-UK, 12M tokens) (van Heuven *et al*., 2014) to the study. Originally designed for psycholinguistic research, these subtitle corpora have been recognized by linguists as indicative of daily oral speech. In Xiao *et al.* (2017), they were taken as representatives of the British and American varieties of English, particularly children's speech, though not sufficiently fairly balanced.

Overall, Xiao *et al*. (2017) rightly paid attention to temporal, regional, social, and pragmatic variables when selecting/composing their reference corpora.

3 Coverage increments: Zipf's law

Larger and more diversified corpora usually have more different words. The coverage rate of a fixed number of top words in these corpora tends to be lower than that in smaller corpora with limited registers (Nation and Waring, 1997). Xiao *et al*. (2017) gave immediate evidence in line with this tendency. The coverage rates of the top 1000/2000/3000 words in EWC (30M tokens) are all higher than those in BNC (100 M tokens) (Fig. 1).

Zipf (1936) proposed that when text size reached a certain level, the frequency of a word would be inversely proportional to its rank in the frequency list. Zipf's observation has been further generalized to a power law relation evident in many languages as the following: Frequency×Rank$^{\gamma}$=Constant.

Many researchers have investigated the value of exponent γ based on different corpora, and found that it depends on factors such as language, author, and domain. The value obtained for general English is around 1, just as Zipf (1936) reported. Thus, it would be good for us to consider the exponent as 1 to estimate the coverage rates of top words in the following investigation.

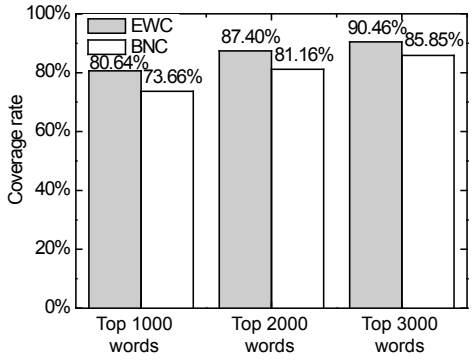

Fig. 1 Coverage rate of the top 3000 words in EWC and BNC (data from Xiao *et al.* (2017))

According to Zipf's law, the word frequency drops as rank increases, and the increment of coverage rates drops accordingly. This is well illustrated by the top 3000 words in EWC and BNC (Fig. 2).

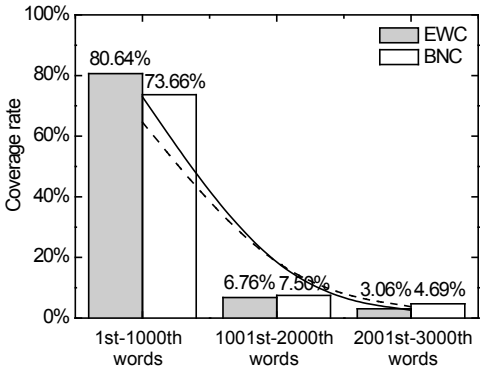

Fig. 2 Coverage increments of the top 1000/2000/3000 words in EWC and BNC

The predictive power of Zipf's law can be further attested to by data from BNC. Based on the statistics of its top 1000 words, the estimated coverage rate of its topmost 726 words (which is the number of unrepeated types/lemmas in the primary textbooks of People's Education Press reported in Xiao *et al*. (2017)) is 70.78%. The accurate coverage rate of the topmost 726 words, according to the BNC word frequency list, is 70.28% (Kilgarriff, 1995). This minor difference of 0.5% demonstrates the good predictive power of Zipf's law.

As to the 228 new words (/types/lemmas) for the 6th graders, Xiao *et al*. (2017) reported that these

words (/types/lemmas) contribute a coverage increment of only 4.09%–5.14% in the four corpora. To them, such an increase was not satisfactory, as compared with coverage increments of 18.74%–21.69% by 194 words and 17.37%–23.20% by 267 words for the 4th and 5th graders, respectively.

Our estimation based on Zipf's law, however, offers a different interpretation. Suppose that all these 228 words (/types/lemmas) are non-repetitively added (i.e., well selected and all new to the 6th graders) and ideally ranked from 499th to 726th, the expected increments in the coverage rate are 4.06% in EWC and 4.02% in BNC at best, with adjusted increments even lower. Viewed in this light, the actual increased rates in the 6th grade (4.09% in EWC and 4.94% in BNC) are reasonable, or even slightly better than expected (Fig. 3).

Zipf's law predicts that the coverage increment will slow down as words of lower frequencies are added. Since Xiao *et al*. (2017) reported close coverage contributions by the 4th and 5th grade vocabulary in all the four corpora, we can infer that new words are not taught in the exact order of their frequency ranks; i.e., some words of higher frequencies are taught later than some others of lower frequencies to the pupils.

Notably in Fig. 3, the topmost 726 words cover 77.17% and 70.78% (in estimation) of the running texts in EWC and BNC, respectively. However, according to Xiao *et al*. (2017), the 726 words in textbook vocabulary account only for 50.33% in EWC and 58.94% in BNC, with a discrepancy of 26.84% and 11.84%, respectively. This confirms that some top words are missing in the textbook wordlists.

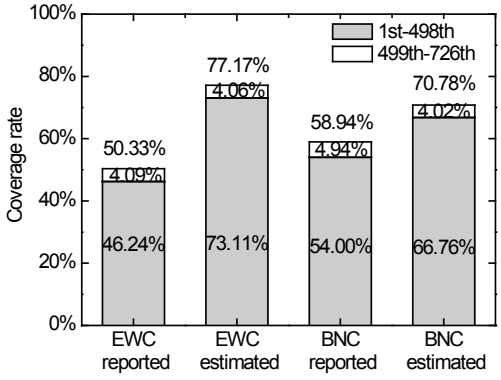

Fig. 3 Reported and estimated coverage rates of 726 words (listed textbook words in the reported data and top words in the estimated data) in EWC and BNC

4 Vocabulary size in textbooks: why and how should it be expanded?

Xiao *et al*. (2017), by comparing the top 3000 words in the four reference corpora, suggested the addition of 903 words. The basis of their reasoning is from Nation and Waring (1997), who claimed that a vocabulary size of 2000 to 3000 words provides a solid basis for language use. It would be ideal if children could master this number of words at their best ages for acquiring language.

If all 903 words were added, however, the compulsory vocabulary size for pupils would be more than doubled, clearly involving heavier workloads. The 6th graders, in particular, have upcoming graduation exams, and because of this, textbooks usually arrange fewer new materials for them. Such practical constraints need to be considered.

Among the 903 words in the intersection of the four frequency lists, are *you*, *it*, *with*, *what*, *not*, *they*, *from*, *me*, *can*, *out*, *who*, *would*, *as*, *some*, *time*, *here*, *then*, *yeah*, *into*, and *where* in the top 20 (Xiao *et al.*, 2017). Almost all are common words that English learners encounter.

If these 20 words are added, textbook vocabulary will significantly increase its coverage rate in general English. Using the statistics given in Xiao *et al*. (2017), the coverage rate will rise from 50.33% to 56.26% in EWC and from 58.94% to 65.29% in BNC (with increases of 5.93% and 6.35% respectively). Coverage for spoken materials will increase even more significantly: from 51.34% to 66.14% in SUBTLEX-US and from 53.98% to 66.67% in CBBS (with increases of 14.80% and 12.69% respectively).

Note that the absence of these 903 words from the textbook wordlists does not necessarily mean their absence in the textbooks. Words such as *be*/*was*, *or*, *year*, and *China* appear rather frequently (with 57/36, 48, 33, and 27 occurrences respectively) in the textbook series under investigation. Unlisted words such as *these* also exist in another popular series of primary textbooks, New Standard English (Li and He, 2010).

Therefore, it seems that pupils are presupposed to know these words in advance. Or rather, students are not required to master all the possible meanings and functions of these words because of their extreme flexibility. Although lack of occurrence in the vocabulary list does not equate to pupils' inability to

recognize, we still agree with Xiao *et al*. (2017) on the inclusion of these words for both their high frequencies and crucial functions.

English is taught in China from Grade 3 in the nine-year obligatory education together with Chinese and other courses, and continues to be taught in high schools and universities. To what extent the English textbook wordlists should be expanded and how the compulsory vocabulary should be reasonably allocated into different periods, remains to be seriously studied. A practical step could be to compare the current primary textbook wordlists with the top 1000 words intersection (instead of the top 3000 words), and then to select a smaller number from this scope.

5 *Twelfth*: to be or not to be?

Xiao *et al*. (2017)'s suggestion that the word *twelfth* be excluded sounds rather bold to us. After all, word recognition is not the only thing that matters in the development of a person's language proficiency. Other crucial issues, such as formulaic expressions, knowledge structures, and cultural awareness, should all be part of enlightening our pupils. The *Twelve Days of Christmas* is a well-known expression as well as a Christmas song, and the *Twelfth Night* has its significance there being the title of one of Shakespeare's best-known plays. As a connection between cardinal and ordinal numbers, the word *twelfth* belongs to an essential word class of human knowledge and English culture. It therefore must be systematically acquired with no random deletion.

6 Coda

Language is a self-adaptive complex system. Its changing vocabulary requires that traditional word frequency lists be regularly reevaluated with the help of the latest information technology. Sampling issues need to be taken care of when constructing reference corpora. Also, distribution patterns such as those suggested by Zipf's law need to be considered when interpreting data.

Various practical constraints should also be considered if theoretical models are to be applied in education. To effectively get students prepared for lively communication, language textbooks should not only be concerned with vocabulary coverage rates, but also make texts interesting to read and easy to follow. The coordination between language skills, emotional attitudes, and cultural awareness, and the coordination between the mother tongue, a second language, and other courses of study should be balanced when designing syllabuses (Ministry of Education of China, 2011). It requires the joint attention of scholars of various backgrounds to improve the effectiveness of textbooks with the aid of latest technological advances. In this sense, Xiao *et al*. (2017)'s attempt is noteworthy.

References

Chujo, K., 2004. Measuring vocabulary levels of English textbooks and tests using a BNC lemmatised high frequency word list. *Lang. Comput.*, **51**(1):231-249. https://doi.org/10.1163/9789004333758_013

Feng, Z., 1998. Linguistics is the bridge between mathematics and humanities. *Mod. Sci.*, (2):21-23 (in Chinese).

Kilgarriff, A., 1995. BNC Database and Word Frequency Lists. http://www.kilgarriff.co.uk/bnc-readme.html

Li, Q., He, S., 2010. Comparison between vocabulary in primary school textbooks and English Curriculum Standards: a case study on New Standard and PEP. *Teach. Res. Prim. Mid. Schools*, (11):8-10 (in Chinese).

Ministry of Education of China, 2011. English Curriculum Standards for Compulsory Education. Beijing Normal University Press, Beijing (in Chinese).

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, Cambridge, UK, p.6-19.

van Heuven, W.J., 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. https://doi.org/10.1080/17470218.2013.850521

Xiao, W., Wang, M., Zhen, W., *et al*., 2017. Corpus-based research on English word recognition rates in primary school and word selection strategy. *Front. Inform. Technol. Electron. Eng.*, **18**(3):362-372. https://doi.org/10.1631/FITEE.1601118

Zipf, G.K., 1936. The Psychobiology of Language: an Introduction to Dynamic Philology. Routledge, Abingdon, UK.