Qiong Hu, Ming Yue, 2017. 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*. *Frontiers of Information Technology & Electronic Engineering*, **18**(7):867-881. http://dx.doi.org/10.1631/FITEE.1700418

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*

Key words: Zipf's law, corpora, textbook, vocabulary list

Corresponding author: Ming Yue
E-mail:yueming@zju.edu.cn
ORCID: http://orcid.org/0000-0001-6457-8176

Motivation

- Xiao *et al.* (2017), by using four corpora, concluded that the primary school English word recognition rate is relatively low, and they suggest an addition of 903 words to the textbook vocabulary list while deleting low frequency words such as *twelfth*.
- As applied linguists in language acquisition, we would like to comment on their study from a Zipfian perspective.

Major results

The coverage rate of a fixed number of top words in larger corpora (e.g., BNC) tends to be lower than that in smaller corpora (e.g., EWC) with limited registers (Fig. 1).

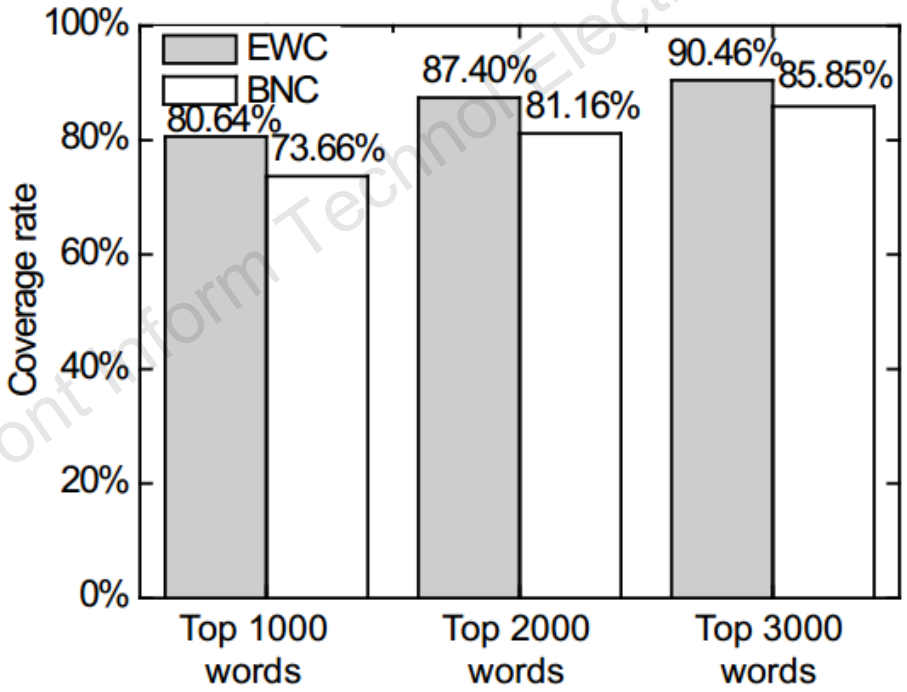

Fig. 1 Coverage rate of the top 3000 words in EWC and BNC (data from Xiao *et al.* (2017))

Major results

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

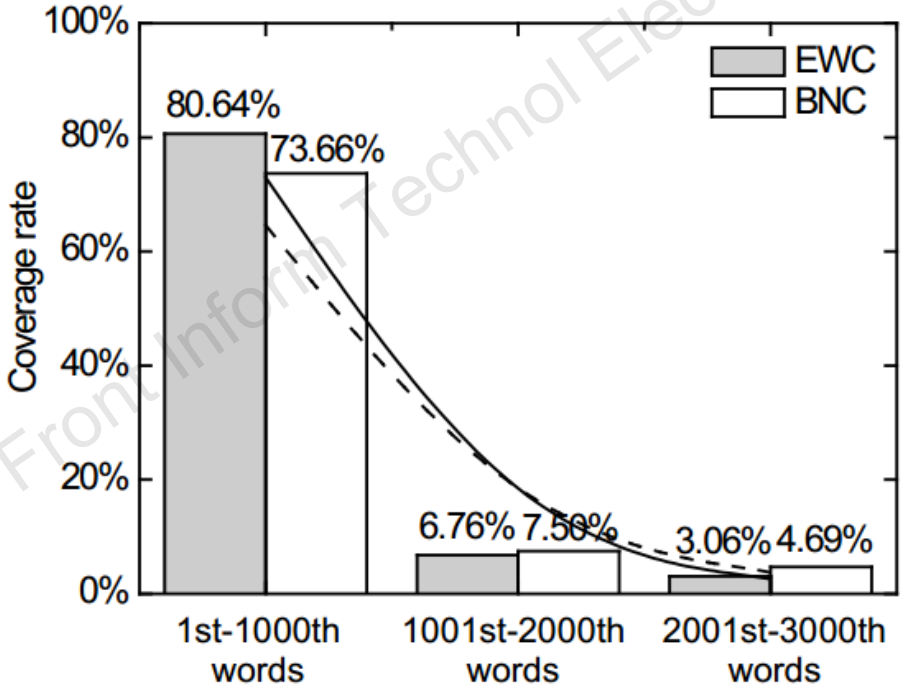

Fig. 2 Coverage increments of the top 1000/2000/3000 words in EWC and BNC

Major results

Zipf's law predicts that, under ideal conditions, the coverage rate increments of the 6th-grade vocabulary are 4.06% in EWC and 4.02% in BNC. The actual increments 4.09% and 4.94% reported by Xiao *et al*. are as good as, or even slightly better than, these two estimated numbers (Fig. 3).

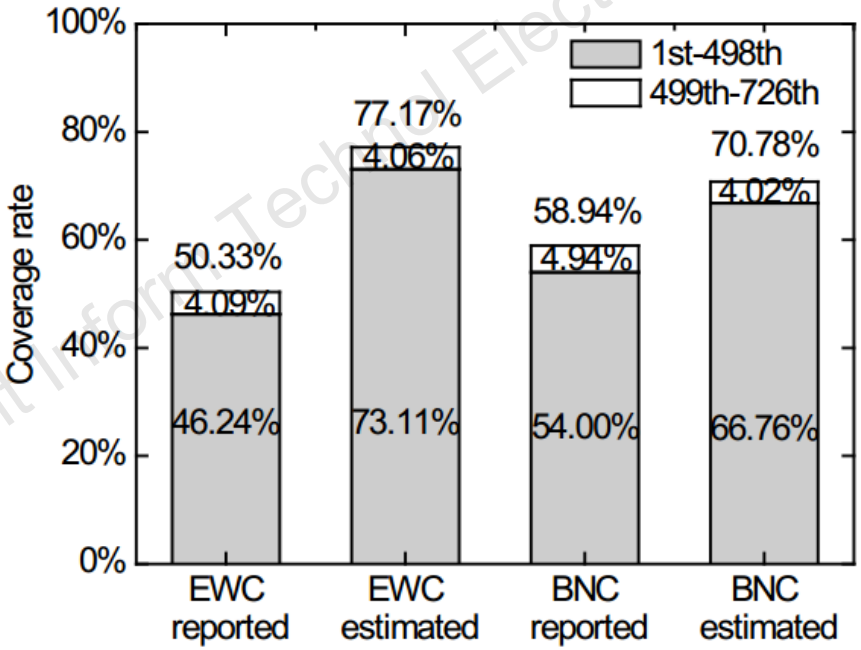

Fig. 3 Reported and estimated coverage rates of 726 words (listed textbook words in the reported data and top words in the estimated data) in EWC and BNC

Major results

- The absence in the vocabulary list does not mean the absence in the textbook.
- By adding only the top 20 words in the 903 words suggested by Xiao *et al.*, the coverage rate will significantly increase by 5.93%-14.80% in the four reference corpora.

Conclusions

- Sampling issues are important when constructing reference corpora.
- Zipf's law can provide evidential support for interpreting word frequencies and vocabulary list analysis. As far as word coverage rate is concerned, the current textbook wordlists provide reasonable increments.
- Practical constraints such as pupils' workloads and cognitive features should be considered if vocabulary size in textbooks is to be expanded.
- Cultural words, such as *twelfth*, should not be deleted.
- Joint attention should be given by scholars of various backgrounds for textbook compilation.