ZWgen: A Python-based Chinese Lexical Stimuli Generator

Abstract

Recent research has developed various Chinese lexicon databases focusing on different aspects. For example, Cai and Brysbaert (2010) provided SUBTLEX-CH — a Chinese word and character corpus sourced from film subtitles. Sze, Liow and Yap (2014) and Tse et al. (2017) released two versions of the Chinese Lexicon Project — a database of lexical decision performance on Chinese compound words. An integrated database that incorporates all of these data is lacking at this moment. Furthermore, the standard of selecting Chinese stimuli is crucial in the lexical decision task. However, previous research did not standardize the stimuli generation criteria due to the lack of resources. Forster (2000) proposed that the ultimate solution is automatic stimuli selection by the computer. Such platforms as the English Lexicon Project (Balota et al., 2007) now provides such automatic selection service, but for English and several alphabetic languages only. Combining features of multiple databases, this study provides an attempt to automatically generate Chinese lexical stimuli. ZWgen — the python-based generator consists of a variety of lexical features that focus on orthographic, phonological, morphological and syntactic aspect, including but not limited to character stroke number, orthographic neighbourhood size, homophone density and frequency of homophones. By manipulating these lexical features, ZWgen will automatically return lexical items that meet the selection criteria. Furthermore, ZWgen also generates pseudo-word stimuli by randomly combining two Chinese character together and cross-check the combined word with a Chinese dictionary to ensure that the combined word does not make sense. Currently, the generator is written in Python and executed on a Jupyter Notebook. The application will be available soon for public users.

References

Balota, D. A., Yap, M. J., Hutchison, K. A., Cortese, M. J., Kessler, B., Loftis, B., ... & Treiman, R. (2007). The English lexicon project. *Behavior Research Methods*, *39*(3), 445-459.

Cai, Q., & Brysbaert, M. (2010). SUBTLEX-CH: Chinese word and character frequencies based on film subtitles. *PloS ONE*, *5*(6), e10729.

Forster, K. I. (2000). The potential for experimenter bias effects in word recognition experiments. *Memory & Cognition*, *28*(7), 1109-1115.

Sze, W. P., Liow, S. J. R., & Yap, M. J. (2014). The Chinese Lexicon Project: A repository of lexical decision behavioral responses for 2,500 Chinese characters. *Behavior Research Methods*, *46*(1), 263-273.

Tse, C. S., Yap, M. J., Chan, Y. L., Sze, W. P., Shaoul, C., & Lin, D. (2017). The Chinese Lexicon Project: A megastudy of lexical decision performance for 25,000+ traditional Chinese two-character compound words. *Behavior Research Methods*, *49*(4), 1503-1519.