Developing an Integrated and Comprehensive Traditional Chinese Corpus Based on Multi-Character Words for Studying relations between words and lexicons

Chung-Ching Wang
National Cheng Kung University, Tainan, Taiwan

Sau-chin Chen
Department of Human Development, Tzu Chi University, Hualien, Taiwan

Yueh Lin Tsai
National Cheng Kung University, Tainan, Taiwan

Yong-Ru Hsiao
National Cheng Kung University, Tainan, Taiwan

Jon-Fan Hu
National Cheng Kung Univeristy, Tainan, Taiwan, ROC

Abstract: Most of Chinese corpus were created for single-character words with indexes, such as frequency, stroke number, and phonetic information, for the purposes of basic research. However, multi-character Chinese words are recognized of referring alterations of meaning and more useful for investigating reading processes and comprehension. Therefore, for studying complete relations between words and lexicons of Chinese, a corpus requires statistics based on more than single-character words with valid and reliable indexes. In this study, we illustrate a corpus of Traditional Chinese providing five word indexes, including word sound, word position, word form, semantics, and competence of forming multi-character words by integrating current credible corpus. The integration approach of the present study is beneficial not only for minimizing inconsistencies of word entities between corpus, but also for calculating quantitative properties of character-to-character relationship. The utilization of the present corpus will significantly impact the studies of Chinese words and reading comprehension.