Identifying lexical bundles in Chinese

Methodological issues and an exploratory data analysis

Chan-Chia Hsu and Shu-Kai Hsieh
National Taipei University of Business / National Taiwan University

Recurrent word sequences, referred to as "lexical bundles", may be structurally incomplete, but they serve important communicative functions. Despite the essential roles of lexical bundles in discourse, many methodological issues have been raised in the process of identifying lexical bundles, which is generally frequency-based. The present study identifies three-word and four-word bundles in Chinese conversation and news, and efforts are made to respond to methodological challenges encountered in previous studies. We employ a more sensitive dispersion measure, DP, and an internal association measure, G, which help filter out high-frequency word sequences with no identifiable function and reduce the workload of further manual interventions. An exploratory data analysis is then conducted to compare the distributional patterns of lexical bundles in Chinese conversation and news. In Chinese, both the type number and the density of lexical bundles are higher in conversation than in news. This appears to be a strong cross-linguistic tendency that reflects the real-time pressure speakers face in spontaneous speech. The exploratory data analysis also shows that the elements in Chinese bundles are closely associated with each other. This suggests that lexical bundles are useful phrasal units in Chinese discourse, and thus invites further investigations of how lexical bundles are used in Chinese.

Keywords: lexical bundle, multi-word unit, frequency, dispersion measure DP, word association measure G