Using Latent Semantic Analysis to Estimate Semantic Transparency and Study How Two-character Chinese Words are Accessed under Different Transparency Conditions

Hsueh-Cheng Wang

Department of Computer Science, University of Massachusetts, Boston, USA

Li-Chuan Hsu

Graduate Institute of Neural and Cognitive Sciences, Taiwan

Yi-Min Tien

Department of Psychology, Chung Shan Medical University, Taiwan

Marc Pomplun

Department of Computer Science, University of Massachusetts, Boston, USA

Abstract

In this study, we propose a computational method for measuring semantic transparency of Chinese words and English compound words and use it to study how language users access a two-character Chinese word's meaning. The method is based on Latent Semantic Analysis (LSA), which is a theory and method for extracting and representing the contextual-usage meaning of words by statistical computations applied to a large corpus of text (Landauer & Dumais, 1997). In the present study, the LSA-based method is found to successfully estimate transparent and opaque constituents and is then used to select Chinese words with different transparency conditions for an eye-movement experiment of Chinese reading. In this experiment, two-character words with opaque-transparent, transparent-opaque, high-frequent opaque-opaque, and low-frequent opaque-opaque combinations of characters were paired with matched transparent-transparent words, and each target word was embedded in an individual sentence. Eye-movement measures including first fixation duration and gaze duration were examined at both the whole-word and the individual-character level. The results are in general consistent with previous studies, but we also found evidence that (1) fully and partially opaque words were accessed differently, (2) fully opaque words were accessed faster than transparent words, and (3) word frequency influenced how fully opaque words were processed, and conclude that the overall results support the dual-route model (Baayen, Dijkstra, & Schreuder, 1997).