Comparing predictions of lexical norm data obtained using word associations and word collocation

Hendrik Vankrunkelsven
University of Leuven, Leuven, Belgium

Steven Verheyen
University of Leuven, Leuven, Belgium

Simon De Deyne
University of Adelaide, Adelaide, SA, Australia

Gert Storms
University of Leuven, Leuven, Belgium

Abstract: We compared the quality of prediction of word variables based on a Dutch word association and text corpus. We derived estimates for: valence, arousal, dominance, concreteness and age of acquisition (AoA) for 2831 words. Based on the similarity between words we: (1) used projections on a dimension identified as the variable in question in a multidimensional representation, (2) used the k-nearest neighbors values, weighted according to their proximity. Estimates prevailed when based on word associations. Differences between the predictions of the two methods were small. Based on the word association corpus it yielded correlations of .92, .85, and .85, for valence, arousal, and dominance, respectively. Its corresponding correlations based on the text corpus were .80, .74, and .67. For concreteness and AoA, both the association and the text corpus yielded correlations of .88 and .73, respectively. This suggests word associations are better at capturing human ratings of affective word variables.