A Sample-size-invariant Estimation of Lexical Diversity

Shohei Hidaka
Indiana University

Abstract: In a long history of studies on language development, lexical diversity has been used as one measure of vocabulary growth. A typical analysis, type-token ratio, or a ratio of sample size of words to the number of unique types of words, has been used as a measure of lexical diversity. As the type-token ratio is not valid for vocabulary sets with different token sample sizes, an improved measure has recently been proposed. In the present study, however, we show that the diversity estimated by past proposed measures is not robust enough for the corpus of different sample sizes. Accordingly, we propose a new formal model of lexical diversity, which distinguishes the latent number of types and property of frequency distribution robustly for different token sizes. We then compared this proposed model to other measures of lexical diversity. In addition, we applied the method to actual vocabulary development data.