AACL 2022 Abstract

Sooyeon Cho & Hakyung Sung

Deadline: February 7, 2022

Thematic Stream: Linguistic analyses of corpora
Submission Category: Full Paper
Word Count: 300 words

Title

Investigating lexical diversity in L2 Korean writing: Focusing on multiple indices and Korean tokenizers

Abstract

Indices of lexical diversity (LD), or the variety of words, are commonly used in L2 writing assessments. The simplest LD index is the type-token ratio (TTR; Johnson, 1944), which is calculated as the number of types divided by the number of tokens. However, due to its strong correlation to text length, there has been considerable effort in recent years to develop indices that are less dependent on text length (e.g., McCarthy & Jarvis, 2007). The developed indices have been studied with L2 English writing but not often extended to L2s other than English.

The present study aims to address this gap by assessing twelve established LD indices in L2 Korean learners' written corpus and evaluating their correlations with the learners' L2 proficiency levels. The data comprises a sample of 4,208 argumentative essays extracted from the National Institution of Korean Language, and the number of tokens ranged from 80 to 480. The corpus was preprocessed, subdivided via parallel sampling (Hess, Sefton, & Landry, 1986) and tokenized with five different tokenizers from the KoNLPy Python package (Park & Cho, 2014). Researchers analyzed twelve LD indices by adapting algorithms developed for English writing assessment (Kyle et al., 2021) to function appropriately with Korean, where a word can be divided into many morphemes.

The results indicated that MATTR and MTLD were both strongly correlated with learner proficiency score while also demonstrating minimal text length effects. Number of word types (abundance) was the index that was most strongly correlated with proficiency scores, and TTR was the most weakly correlated, but both were susceptible to the text length. The figures for each index differed slightly based on how different tokenizers analyzed morphemes. Implications and limitations related to the ways of tokenization and how they affect the calculation of LD indices in Korean texts will be discussed.

Word Count: 300 (300 max)