Dissecting the interplay between corpus properties, algorithm, and word segmentation performance

Jun Ho Chai
Sunway University, Subang Jaya, Malaysia

Seongmin Mun
Ajou university, Suwon, Korea, Republic of

Eon-Suk Ko
Chosun University, Gwangju, Korea, Republic of

Abstract

This study investigates how corpus-level properties in Korean child- and adult-directed speech shape word segmentation across four algorithms: Transitional Probability, Diphone-Based Segmentation, PUDDLE, and Adaptor Grammar. Utterance length consistently impacts segmentation, with shorter utterances improving performance, particularly for PUDDLE, DiBS, and AG. Word length affects transitional probability algorithms, while hapax legomena introduce challenges for forward TP and AG. Interjections negatively influence AG, but not the others, and larger corpus size benefits PUDDLE. Register effects are limited, with forward TP and PUDDLE performing better on child-directed speech. These patterns highlight algorithm-specific sensitivities, with utterance length emerging as the most consistent factor. Our findings underscore the importance of considering both input properties and algorithm design when studying word segmentation in Korean. Future work should explore cross-linguistic comparisons, larger balanced corpora, and the role of multimodal cues in segmentation.

In D. Barner, N.R. Bramley, A. Ruggeri and C.M. Walker (Eds.), *Proceedings of the 47th Annual Conference of the Cognitive Science Society*