Use of Lexical Statistics for Compound Word Recognition and Segmentation in Turkish

Ozkan Kilic
Graduate Student

Abstract: Compound words are cross-linguistic morphological phenomena that occur in all languages. Compound words are widely accepted to be stored in the lexicon but their constituents need to be accessed during both language learning and production processes. In this study, the use of corpora was investigated for how to differentiate single-stem words from single-word compounds and then how to segment compound words when no phonological information is available. Stems and morphs discovered in manual segmentations of the METU-Sabancı Turkish Treebank and the CHILDES were employed in the compound word recognition task and the results were compared. The METU Turkish Corpus (with about 2 million words) and a web-corpus (with about 490 million of Turkish words) were utilized in the segmentation task. The results emphasize that the lexicon can be morpheme-based; and lexical frequencies are effective heuristics in compound word recognition and segmentation.