What can distributional semantics tell us about (mor)phonotactics?

The goal of the paper is to model the interaction between phonology and morphology in Polish using Distributional Semantics. Emphasis is placed on word-initial consonant clusters. Previous research has suggested that long and universally marked sequences of consonants are primarily triggered by the intervention of morphology (morphonotactics, Dressler & Dziubalska-Kołaczyk 2006). Interestingly, the parsability of clusters emerging due to morphological operations is not always obvious in Polish, resulting in several degrees of morphotactic transparency of the prefix and the stem.

In the present analysis, we ask the reverse questions: Can the morphological status of consonant clusters, their their graded transparency and markedness be inferred from semantic vectors without consulting the morphological structure of words? Can the phonological grammar of a language be modelled without consulting the phonemic representation of syllables? We answer these questions on the example of Polish, a language characterized by non-trivial morphology and an impressive inventory of morphologically-motivated consonant clusters (e.g. Orzechowska 2019). Instead of determining morpheme boundaries within a cluster, which has been the main goal of the previous studies, we use statistical and computational techniques that operate on word embeddings (e.g. t-SNE, (Maaten & Hinton 2008; Linear Discriminative Learning, Baayen et al. 2019) obtained with the FastText algorithm (Bojanowski et al. 2017).

The study demonstrates that - apart from encoding rich semantic and syntactic information - semantic vectors can infer sub-lexical linguistic units such as phoneme strings, morphemes, along with their structure and function. That is, phonotactic complexity and morphotactic transparency can be modelled without accessing the words' internal structure. We shows a link, to say the least, between word meaning and aspects of phonotactic complexity. In particular, FaxtText vectors are very accurate in predicting the number and type of consonants in a string as well as the sonority slope, or a sequence of such slopes in a cluster. The findings contribute to the state of the art by demonstrating that semantics may serve as a cue for phonotactic complexity. That is, words that are semantically related start with similar consonant clusters in terms of their length and constituent consonants. A consonant cluster and a morphological boundary within it map a well-defined set of words (for similar observations in Polish see Dziubalska-Kołaczyk et al. 2011).

Keywords:
(mor)phonotactics
morphotactic transparency
distributional semantics
Polish