Complex words as shortest paths in the network of lexical knowledge

Sergei Monakhov[1], Karsten Schmidtke-Bode[2] & Holger Diessel [3]
[1, 2, 3] Friedrich-Schiller University of Jena, holger.diessel@uni-jena.de

Keywords: grammar network, complex words, morphological constructions

Models of word recognition diverge on the question of how to represent complex words. Under the morpheme-based approach, each morpheme is represented as a separate unit (Pinker, 1999) while under the word-based approach, morphemes are represented in lexical networks (Bybee, 1995; Baayen et al., 2015). The word-based approach is consistent with construction morphology (Booij, 2010) and recent research on the grammar network (Diessel, 2019). However, while the network view of constructions has become popular in recent years, there is little computational and experimental research on this topic.

In the current study, we used a network model and an experiment to investigate the Latinate component of English morphology. Specifically, we assumed that complex words can be conceptualised as paths in a weighted directed network of morphemes. The edges of the network have different weights that are determined by usage frequency. New words are created easily if they follow well-trodden paths.

The network construction process ran as follows. We created a graph G = (V, E), such that $V = V_{Lb} \cup V_{dm}$, where V_{Lb} is a set of vertices that represent the 100 most frequent Latinate bases and V_{dm} is a set of vertices that represent all derivational morphemes encountered in 12,950 unique English lexemes with these bases. For any pair of morphemes v_i and v_j, the edge $v_i \rightarrow v_j$ was added to the set of edges E only if this sequence of morphemes was attested in at least one lexeme in the data. The total number of such lexemes was assigned to the edge as its weight.

If this network is an accurate representation of lexical knowledge, one expects to find (1) that the attested words will represent the most heavily weighted ('shortest') among all possible paths connecting their initial and final morphemes, and (2) that, in an experimental setting, unattested 'possible' (Aronoff, 1976) words derived along the shortest paths will get higher acceptability ratings and require less decision time.

To test hypothesis (1), we constructed for each word in the data all possible paths connecting its initial and final morphemes and going through its base. After arranging the paths in the descending order of the sums of weights, we obtained the ranks of the attested words. We found that in 98% of all cases these words are among top 15 paths. Importantly, the ranks and frequencies are perfectly negatively correlated ($r = -0.99$, $p < 0.001$), suggesting a power-law distribution.

To test hypothesis (2), we randomly selected several possible English words from the network. The participants were asked to rate the acceptability of a given word on a scale of 1 ('unlikely') to 5 ('likely'). For each possible word, we obtained its median rating and task completion time, the final measure was calculated as a ratio of the two. The correlation coefficient ρ (rating/time, path weight) was found to be 0.81, $p < 0.001$.

Generalizing across these findings, we argue that complex words are best analyzed in a network model of morphological constructions that is shaped by language use.

References

Aronoff, M. 1976. *Word Formation in Generative Grammar*. Cambridge, MA: MIT Press.

Baayen, H., Shaoul, C., Willits, J., & Ramscar, M. 2015. Comprehension without segmentation: A proof of concept with naive discrimination learning. *Language, Cognition, and Neuroscience*, 31(1): 106–128.

Booij, G. 2010. Construction morphology. *Language and Linguistics Compass*, 4/7: 543–555.

Bybee, J. 1995. Regular morphology and the lexicon. *Language and Cognitive Processes*, 10: 425–455.

Diessel, H. 2019. *The Grammar Network. How Linguistic Structure is Shaped by Language Use*. Cambridge: CUP.

Pinker, S. 1999. *Words and Rules: The Ingredients of Language*. New York: Harper Perennial.