**What can bound roots teach us about morphemes and morphology?**

Matthew T. Carlson & Amy C. Crosson

Penn State University

Whatever our theoretical approach to the nature and structure of words, they possess an undeniable phenomenological reality. Language users have strong shared intuitions about how words divide into smaller units, and about how those units contribute to the meaning of the whole word. This phenomenological intuition is reflected in the basic definition of a morpheme, as often cited in introductory linguistics materials. Namely, the morpheme is the smallest meaningful unit of language. Research in both linguistics and psycholinguistics paints a more subtle picture, but substantial focus remains on the idea of morphological *representations*, expressed in terms of morphological units. A different, connectionist approach (Baayen et al., 2019) has sought to model the mental lexicon without explicit representations of morphemes, or even of words, by learning associations between sequences of phonological units and complex semantic (e.g. vector) representations. We might understand this approach as focusing on *processing*, i.e. accessing semantic information based on form, or vice versa, to the exclusion of any explicit lexical representations, and without compositionality of meaning. This approach leads us to question: What does it mean to be a morphological or a lexical unit?

We seek to shed light on this question by examining words whose morphological structure is arguably marginal, specifically, English words with bound roots. Bound roots are interesting because whatever clear morphological role they once played in the language has been obscured over time, such that they are not always easily segmented from neighboring material (e.g. *circ-us*, *ver-dict*, *min-strel*), and their meanings are often vague (cf. *de-fect, per-fect*), multiple (*cert-ain, con-cert*), or obscure (e.g. *eti-ology*). Thus, the presence (or absence, in the case of folk etymologies) of a historical root does not reliably indicate whether the root plays a synchronic role in speakers’ mental lexicons. Nevertheless, given their importance in academic vocabulary, the apparent or historical morphological complexity of these words has been explored as a resource in pedagogy for helping children and adolescents learn them more effectively (Crosson & McKeown, 2016), making these words of both theoretical and practical interest.

We consider the form and lexical semantics of these words separately through computational modeling, and we present some initial comparisons of our results with large-scale lexical processing data (Balota et al., 2007). Concerning form, we used two unsupervised parsers to assess the identifiability of a set of bound roots, and to derive measures of morphological family size. One parser (Smit et al., 2014) isolated the roots much more frequently than the other (Xu et al., 2020). Concerning semantics, we explored the clustering of words in a semantic space (Baroni et al., 2014), using Principal Components Analysis to derive measures of semantic coherence and polysemy as well as the relationship of specific words to meanings attributable to the bound root. While formal parsability and semantic measures were not necessarily closely related, both form-based and semantic measures predicted behavioral data. Interestingly, we found evidence of a tradeoff in that root polysemy effects were stronger for words in which the parser did not isolate the root.

How does this confluence of formal and semantic information relate to the experience of recognizing, producing, learning, or creating words? We argue that these initial explorations point away from a representational view of morphemes, and towards a a conceptualization of morphemes as intersections of form and meaning that makes little sense apart from human language processing. We hypothesize that morphological structure, family relationships, and so on do not define the structure of the mental lexicon in long-term memory, nor are they static properties of long-term memory representations of words. Our interpretation thus has something in common with the end-to-end models based on discriminative learning (such as in Baayen et al., 2019). Nevertheless, we seek insight into the interplay of form and meaning in such models, conceptualizing morphology as affording intersecting pathways for navigating between form and meaning during production, comprehension, learning, and creating words.

References

Baayen, R. H., Chuang, Y.-Y., Shafaei-Bajestan, E., & Blevins, J. P. (2019). The Discriminative Lexicon: A Unified Computational Model for the Lexicon and Lexical Processing in Comprehension and Production Grounded Not in (De)Composition but in Linear Discriminative Learning. *Complexity*, *2019*, 1–39. https://doi.org/10.1155/2019/4895891

Balota, D. A., Yap, M. J., Cortese, M. J., Hutchison, K. A., Kessler, B., Loftis, B., Neely, J. H., Nelson, D. L., Simpson, G. B., & Treiman, R. (2007). The English lexicon project. *Behavior Research Methods*, *39*(3), 445.

Baroni, M., Dinu, G., & Kruszewski, G. (2014). Don’t count, predict! A systematic comparison of context-counting vs. Context-predicting semantic vectors. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 238–247. https://doi.org/10.3115/v1/P14-1023

Crosson, A. C., & McKeown, M. G. (2016). Middle School Learners’ Use of Latin Roots to Infer the Meaning of Unfamiliar Words. *Cognition and Instruction*, *34*(2), 148–171. https://doi.org/10.1080/07370008.2016.1145121

Smit, P., Virpioja, S., Grönroos, S.-A., & Kurimo, M. (2014). Morfessor 2.0: Toolkit for statistical morphological segmentation. *Proceedings of the Demonstrations at the 14th Conference of the European Chapter of the Association for Computational Linguistics*, 21–24. https://doi.org/10.3115/v1/E14-2006

Xu, H., Kodner, J., Marcus, M., & Yang, C. (2020). Modeling Morphological Typology for Unsupervised Learning of Language Morphology. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 6672–6681. https://doi.org/10.18653/v1/2020.acl-main.596