Processing Non-Concatenative Morphology – A Developmental Computational Model

Tamar Johnson, Inbal Arnon

Acquiring verbal systems with non-concatenative morphology is a challenge for infants, adult learners and computational models. Computational models designed to learn concatenative morphologies performed poorly when applied to Semitic languages, in which verbs are formed by integrating the root into one of a number of templates (Bergmanis & Goldwater, 2017). One model that targets root-templatic morphology learning is Fullwood and O'Donnell's (2013) Bayesian model. While this model exhibits high performance in matching words with their correct templates, it has several crucial limitations as a model for language development. First, the type of input it receives (filtered verbal stems from the Quranic Arabic Corpus) does not represent the input infants receive. Second, the model performs multiple iterations over the entire input to match each word with its template. Furthermore, Fullwood and O'Donnell's model works within the generative approach and has a priori notions of roots and templates as the components structuring Semitic verbs.

In this study, I developed a cognitively feasible computational model for processing root-templatic verbs and extracting their roots. Identifying the root characters and linking them to the core semantics of the verb is claimed to be an important stage in acquiring the verbal paradigm in Hebrew, a Semitic language (e.g., Berman, 1985), and in other Semitic languages (Ravid, 2003). The model presented here is a developmental model within the usage-based framework. Hence, a priori notions regarding the structure of verbs (i.e., built out of roots and templates) are not given to the model but instead are acquired through exposure.

The model functions in a fully incremental fashion and makes use of three types of information assumed to be available to the learner: the sentential context the verb appeared in (specifically the bigram followed by the verb in the input), the phonological representation of the verb and the meaning of the verb. The model therefore assumes that the learner can link the linguistic input with the conceptual representation of the event she takes part in or observes (Alishahi & Stevenson, 2008). The sentential context of the verb enables the model to predict which verb follows each bigram in the input, as the model involves predictive and error-driven learning, both claimed to be features of human language learning (e.g., Christiansen & Chater, 2016). In the case of a false prediction the predicted and the correct verb are compared, both phonologically and semantically. This produces two strings of characters extracted as the difference between the two verbs. For example, if the trigram *ma ʔat ʕoṣā* (what are you doing) is highly frequent in the input, then when the bigram *ma ʔat* appears again, the model will predict that the verb *ʕoṣā* will follow. If a different verb appears instead, say, *Šotā* (drinking), the two verbs are compared and the two sets of characters, [*Š, t*] and [*ʕ, ṣ*], are extracted, each linked with the meaning of the relevant verb. The root linked with a specific meaning is the one that was extracted for a given verb meaning with the highest frequency in the input. The output of this model, then, is a 'dictionary' matching each root with the verb meaning.

The model was applied to Hebrew child-directed speech utterances from the CHILDES corpora, phonetically transcribed and morphologically annotated (with verbs translated to English,

Proceedings of the Society for Computation in Linguistics (SCiL) 2019, pages 389-390.
New York City, New York, January 3-6, 2019

which was used for the semantic representations). The Hebrew corpus (BSF-Long by Armon-Lotem & Berman, 2003) contains 25,716 child-directed utterances, 4.3 words long on average, directed to Hebrew-speaking children between the ages of 16 and 42 months.

To evaluate the model's performance, the accuracy of the root strings extracted by the model was computed. The mean accuracy was evaluated against the accuracy of a baseline model, which does not use predictive learning, but processes verbs independently from the input without using their sentential context. Mean accuracy of the model presented here was found to be 0.503 Which was higher than the mean accuracy of 500 runs of the baseline model (mean accuracy = 0.412) and also higher than the highest accuracy achieved by the baseline model (0.426).

This study outlines a novel developmental model for extracting roots from verbs in non-concatenative morphologies and demonstrates that using the sentential context in which verbs appear aids the task of extracting roots for Hebrew verbs. The model is predicted to perform with good accuracy when applied to concatenative languages in extracting verbs' stems, and future work will test that.

Alishahi, A. & Stevenson, S. (2008) A computational model of early argument Structure acquisition. *Cogn. Sci.* 32, 789–834.

Armon-Lotem, S. & Berman, R. A. 2003. The emergence of grammar: Early verbs and beyond. *Journal of Child Language* 30: 845-878.

Bergmanis, T., & Goldwater, S. (2017). From Segmentation to Analyses: A Probabilistic Model for Unsupervised Morphology Induction. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers* ,1, 337-346.

Berman, R. A., & Slobin, D. (1985). The acquisition of Hebrew. *The cross-linguistic study of language acquisition*, *1*, 255-371.

Christiansen, M. H., & Chater, N. (2016). The Now-or-Never bottleneck: A fundamental constraint on language. *Behavioral and Brain Sciences*, *39*.

Fullwood, M. A., & O'donnell, T. J. (2013). Learning non-concatenative morphology. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics* (pp. 21-27).

Ravid, D. (2003). A developmental perspective on root perception in Hebrew and Palestinian Arabic. In Y. Shimron (ed.), Language Processing and Acquisition in Languages of Semitic, Root-Based Morphology, 293-319. Amsterdam: Benjamins.