Acquiring formulaic language: A cross-linguistic model of children's comprehension and production

Stewart M. McCauley & Morten H. Christiansen
Cornell University

Recent work in psycholinguistics has underscored the importance of formulaic expressions in everyday language use, in line with predictions emerging from cognitive linguistics. Adults' sensitivity to the properties of multiword sequences in comprehension and production (e.g., Arnon & Snider, JML 2010) suggests a greater role for formulaic processing than has previously been assumed. Such results are mirrored in developmental studies (e.g., Bannard & Matthews, Psych. Science 2008), suggesting that children's item-based linguistic units—and their active use during processing—do not diminish, but persist throughout development and into adulthood. If this is indeed the case, it holds that we can better understand the role of formulaic sequences in adult language by studying the processes whereby children discover and use such units.

To this end, we describe a computational model of acquisition which instantiates the view that the discovery and on-line use of formulaic sequences forms the backbone for children's language processing, such that the importance of multiword units grows rather than lessens over time. Our model simulates both comprehension and production, learns incrementally using simple statistics, offers broad, cross-linguistic coverage, and accommodates a range of developmental findings. Importantly, the model features no distinction between language learning and language use; it learns by merely attempting to comprehend and produce language.

The model learns from corpora of child-directed speech, acquiring item-based knowledge incrementally, through on-line learning of simple statistics in the form of backward transitional probabilities (which 8-month-olds can track; Pelucchi et al., Cognition 2009). The model gradually builds up an inventory of chunks consisting of one or more words, uniting comprehension and production within a single framework. The model groups words together as they are encountered, incrementally building an item-based “shallow parse” as each incoming utterance unfolds (reflecting evidence for shallow comprehension based on local information; e.g., Ferreira & Patson, Lang. Ling. Compass 2007). When the model encounters a multiword utterance produced by the target child of a corpus, it incrementally produces its own utterance using chunks and statistics learned up to that point. The model's comprehension abilities are scored against a state-of-the-art shallow parser, and its production abilities are scored against the target child's original utterances (the model's utterances must match the child's).

The model exhibits strong performance with over 200 single-child corpora (from the CHILDES database) representing a typologically diverse array of 29 languages, achieving high shallow parsing accuracy and correctly producing the majority of the child utterances encountered. In each case, the model outperforms various Markov models (p's<0.0001). The model also provides close quantitative fits to key psycholinguistic findings regarding children’s distributional and item-based learning (e.g., phrase frequency effects: Bannard & Matthews, Psych. Review 2008; morphological processing: Arnon & Clark, Lang. Learn. Development 2011; artificial grammar learning: Saffran, JML 2002; relative clause production: Diessel & Tomasello, Language 2005).

Using these results, we argue that a great deal of children's early linguistic behavior can be accounted for by incremental, on-line learning of formulaic sequences using simple distributional cues. Moreover, we argue that a complete understanding of the role played by formulaic sequences in adult language can only be arrived at by adopting a developmental perspective.