Detecting Syntactic Patterns in Speech to Infants: The Utility of Frequent Frames

Toben H. Mintz (tmintz@usc.edu)

Departments of Psychology & Linguistics, Program in Neuroscience, University of Southern California
3620 McClintock Ave., Los Angeles, CA 90089-1061 USA

This study shows the utility of a simple distributional cue from which learners could make crucial syntactic generalizations. Mintz (2003) proposed that frequent frames provide robust distributional cues for bootstrapping into the grammatical categories of a language. Mintz defined a frequent frame as two words co-occurring frequently in a corpus, at a distance of exactly one word. Thus, in She stopped to have ice-cream, she__to is a frame that contains the verb have. If she__to were frequent, Mintz’s procedure would group together all words that occurred in the frame across a corpus. Analyzing six corpora from the CHILDES database (MacWhinney, 2000), Mintz showed that categorizing words in this manner results in extremely accurate categories while requiring limited computational resources. Similar results have been obtained in other languages, such as French (Chemla, Mintz, Bernal, & Christophe, under review), and Mandarin (Xiao, Cai, & Lee, 2006). Furthermore, Mintz (2006) demonstrated that 12-month-old infants can categorize novel words using distributional information consistent with frequent frames. The present study extends these findings by showing that frequent frames can detect important combinatorial regularities of words, in turn allowing for the acquisition of aspects of a language’s syntactic structure.

The first step of the analysis duplicated the procedures in Mintz (2003), analyzing the same six English corpora. This step resulted in a set of frequent frames (henceforth, minimal frames, MF), and the words they categorize. Next, an expanded frame (EF) was generated from each MF. EFs were identical to corresponding MFs, except they were constrained to have exactly two intervening word positions, and they were not subject to any frequency criteria (however, identification of the MFs were subject to frequency criteria, as usual). Thus, the MF she__to from above would be an EF in She could stop to have ice-cream.

The two-word sequences inside EFs were then evaluated in several ways. First, the contents of the EFs were compared to the contents of the corresponding MFs. This revealed that many of the words categorized by MFs occurred within EFs as well. For example, in several corpora the MF you__the contained put, and the EF you__the contained can put and gonna put.

Second, word-order patterns in the content of the EFs were analyzed. In many EFs, these sequences displayed an interesting systematic patterning of a limited set of functors in one position, and a larger range of content words in the other position. E.g., in the you__the EF, the auxiliaries can, and gonna, occurred with many verbs (and many of those verbs occurred in the MF as well); likewise, in the put__back EF, the determiner the occurred in sequence with many nouns. In the majority of cases, across frames and corpora, patterns like these constituted a linguistically significant unit: e.g., determiner-noun sequences form a syntactic constituent, and auxiliaries carry tense and agreement information that is relevant for the main verb. Thus, sequences within EFs could give the learner crucial information about the ordering of function and content words in the language (something that they must learn, as these properties vary cross-linguistically), and in some cases could provide evidence about the organization of phrases. For example, by noticing the small set of frequently recurring words to the left of the nouns (e.g., in the put__back EF) learners could deduce that determiners occur to the left of nouns in noun phrases.

Frequent frames have already been shown to be robustly informative for one type of crucial syntactic information, lexical categories, and infants have been shown to be sensitive to this simple distributional environment. The present study shows that frames could also provide learners with crucial information about aspects of linguistic structure that involve combinatorial regularities of words. These results are important because they show that a simple distributional environment to which infants attend could be a powerful cue to multiple aspects of syntactic structure.

Acknowledgments
This research was supported by the National Institutes of Health (HD40368).

References