Speech Segmentation and Word Discovery: PARSER

Taylor Cassidy¹ and Danny Nassre¹

¹City University of New York
contact: nassre@gmail.com

This study develops a version of PARSER¹, PARSER2 – a psychologically feasible model of the earliest stages of speech segmentation and word discovery (SSWD). Unlike several models of SSWD, PARSER2 makes minimal assumptions about the computational capacities of language learners². PARSER2 is a first-stage process that can feed into other processes. In this case, it is used to estimate the word boundary probability distribution (WBPD). Because the probability of a word boundary given two adjacent segments is usually either very high or very low, the WBPD is extremely bimodal, and it can be thought of as a representation of phonotactic knowledge. It can be used to segment a corpus of English child-directed speech with 76% precision and 69% recall³.

PARSER2 models two processes: the chunking of continuous, unfamiliar input into discrete units, and the retention in memory of more frequently chunked units. PARSER2 (initially randomly) chunks phonemically transcribed speech, creating a lexicon which rewards or penalizes candidate words based on their frequency, with higher weighted candidates more likely to be actual words. These candidates are used to aid subsequent segmentation, and lower-weighted ones are deleted from the lexicon if they are not encountered with sufficient frequency.

The original PARSER was run on syllabified artificial language data. Some psycholinguistic justification for the use of syllabic input exists: even 4-day-old infants can discriminate utterances with different numbers of syllables, but not different numbers of phonemes⁴; and, among older children, analysis of utterances into phonemes is more difficult than analysis into syllables⁵. Nevertheless, the perceptual salience of syllables does not mean their boundaries are a priori clear to the child; a model using syllabified input that is consistent with word boundaries begs the question of how that consistency is achieved.

To avoid presupposing what has to be learned, PARSER2 uses input that can be described as “partially”-syllabic. The chunking process outlined above is constrained so that all chunks must contain at least one vowel, reflecting the perceptual salience of syllables. The model illustrates a process whereby children form words (or syllables) around vowels, and refine their understanding of syllable and word boundaries as they proceed. Regular phonemic input can also be used, however, with minimal degradation of precision.

PARSER2 was run on CHILDES⁸ transcribed speech with variations in the forgetting rate, which is the primary determinant of its performance. A higher forgetting rate leads to a more accurate lexicon, at the expense of completeness, and the reverse is true for a lower forgetting rate. Additionally, words that attained a sufficient weight (subject to parameterization) were automatically segmented for the rest of the process, enabling better segmentation of adjacent words.

The more accurate PARSER2 lexicons did not produce the best segmentation results, because they were too small to be representative. The best segmentation performance came from a lexicon with 904 word candidates, 64% of which were actual words. The precision increases as lower weighted candidates are excluded, with 77% of the top 200 candidates being actual words. Most of the growth in the lexicon occurs in the first few thousand utterances (Figure 1).

This lexicon was used to segment a second batch of utterances, and a word boundary probability distribution was calculated from the segmented utterances, which was then used to segment a third batch of utterances (a cut was made between two segments $p_1$ and $p_2$ when the probability of $p_1p_2$
occurring within a word was less than .5), achieving 55% precision and 54% recall. The numbers might not be impressive at first glance, but the amount of utterances PARSER was trained on—200,000—probably only represents a few hundred hours of parental interaction, according to our extrapolation from the duration-marked transcripts in the Brown corpus in CHILDES. Additionally, the algorithm for segmenting the second batch of utterances with words from the PARSER2 lexicon was rather naive (the longest matching word candidates were segmented from the beginning of each utterance or utterance remainder); a more sophisticated word recognition algorithm could make greater use of PARSER2’s output. PARSERS2’s performance, and the simplicity of the model itself, suggest that it may be a plausible mechanism for the earliest stages of SSWD and the acquisition of phonotactic knowledge.

References:


Figure 1: Development of PARSER Lexicon