Word Segmentation as Word Learning: Integrating Meaning Learning with Distributional Cues to Segmentation

Michael C. Frank, Vikash Mansinghka, Edward Gibson, and Joshua B. Tenenbaum
Massachusetts Institute of Technology

1. Introduction

How do very young children begin to learn the meanings of words? When thinking of word learning, we normally picture a two- or three-year-old engaged in play with a caregiver who ostensively defines words, say by pointing at a toy and naming it with a simple phrase, for example, "look at the *elephant*!" This phrase, combined with a point and a joint focus of eye-gaze, provides the older child with unambiguous evidence that the word *elephant* refers to this toy.

Younger children in this situation may have a more difficult time making this same mapping. They may not fully understand the combination of pointing and eye-gaze that will later become salient for determining reference. (Hollich et al., 2000) They may not already know the frame words, "look at the," and so they may interpret these as potential names for the toy. Finally, they may not even be able to segment the full sentence correctly into its component words (Jusczyk, 1999), leaving them with a large number of potential syllable (or even phone) combinations to map onto the referent (even assuming that the referent is uniquely determined—e.g., there are no other toys around). How are children able to solve this set of interlocking problems in order to learn their first words?

We examine one particularly difficult problem that a young word-learner faces: how to determine which group of sounds (out of a larger sentence) maps onto a novel referent. This problem implicitly includes a larger problem—the problem of word segmentation—but solving word segmentation does not entirely solve it. We refer to this problem as the problem of word-to-referent mapping to denote that we are attempting to find a set of syllables which map to a particular referent that has already been uniquely specified. This is the problem faced by the child who knows it is the concept ELEPHANT that she wants to learn a name for. (This problem can be considered a subset of the problem of word-to-world mapping, where the referent is not uniquely determined and there is uncertainty on both the linguistic and the referential side.)

---

* This research was supported by a Jacob Javits Fellowship for Graduate Study to the first author. The authors wish to thank Anne Fernald, Jenny Saffran, and the members of Tedlab and the Computational Cognitive Science Group for their helpful feedback.

In the current studies, we used an artificial language learning paradigm (e.g., Gomez & Gerken, 2000; Saffran, Newport, & Aslin, 1996; Yu & Smith, in press) to provide an abstract model of the tasks early word learners face. We exposed adult participants to a series of referential events, in which a sentence was used in the presence of a single, salient object. The task of the learner in this paradigm was both to learn the words of the language they heard and to identify mappings between those words and the objects they saw. In addition to our experiments we present preliminary results from simulations in a computational model of this task. The aim of our work was to create a simplified situation in which we could address a number of questions about the problem of word-referent mapping:

- Can learners learn meanings even when there is uncertainty about segmentation?
- Is there any aid to segmentation performance in learning meanings at the same time?
- Does a constant position in the sentence for meaning-bearing words make them easier to learn?

In this paper we present first the results of our experimental study and then the details of our computational model and some preliminary results.

2. Experimental Study

The aim of our study was to expose participants to simple, un-segmented sentences in an artificial language and to test both their segmentation performance and their ability to learn the meanings of words. Accordingly, we designed our task to have three phases: familiarization, segmentation testing, and meaning-learning testing. During the familiarization phase, participants heard sentences in the artificial language for approximately 30 minutes. After half of the sentences, they were tested in a simple change-detection task to make sure they were paying attention to both the pictures and the sentences. We tested four familiarization conditions:

- Random position: words paired with pictures appeared in random positions within each sentence
- End position: words paired with pictures appeared always in sentence-final position within each sentence
- Scrambled meanings: a picture was randomly chosen to go with each sentence (this condition was included to establish that the statistics of the input provided no guidance as to word meanings)
- No meanings: only audio strings were presented (this condition was included to test whether the inclusion of meanings aided segmentation)

In the second part of the experiment, participants were tested on their ability to segment the language in a standard word/part-word forced choice paradigm (Saffran et al., 1996). During the third part, we tested how well participants had learned the meanings of the words in our language by asking them to choose which word best matched a given picture.

2.1. Methods

Participants. Our sample was comprised of 65 adults, both MIT undergraduates and members of the surrounding community.

Materials. Familiarization materials were constructed uniquely for each participant from a randomly generated artificial language. Each language consisted of 18 syllables (*ba, bi, da, du, ti, tu, ka, ki, la, lu, gi, gu, pa, pi, va, vu, zi, zu*) permuted into six words, two of which had two syllables, two of which had three syllables, and two of which had four syllables. In those conditions in which there was a consistent relationship between pictures and words (random position and end position), two words were chosen randomly and one was assigned to each of the two pictures. The two pictures were sepia-tinted bitmaps created using image manipulation software, one resembling a novel plant and the other resembling some sort of artifact.[1]

All speech in the experiment was synthesized using the MBROLA speech synthesizer (Dutoit, Pagel, Pierret, Bataille, & van der Vrecken, 1996) using the us3 diphone database to produce an American male speaking voice. All consonants were 25ms and all vowels were 225 ms in duration and the fundamental frequency of the synthesized speech was ~100 Hz. No breaks were introduced into the speech: the synthesizer created equal co-articulation between every phone in the sentences.

Sentences in our artificial language were constructed by concatenating together three words in the language (chosen without repetitions). In the random position condition, exactly one of these words was one of the two associated with a picture; in the end position condition, this word was always the last word of the sentence. In the scrambled meanings condition, a picture was randomly chosen to match each sentence. In the no meanings condition, there was no picture. There were no other differences between conditions. One hundred and sixty such sentences were generated as the training set.

Segmentation test words were constructed by concatenating all six words in the language and then moving each word boundary one syllable into the next word in the string, in order to produce plausible distractors which appeared in sentences in the language. These distractors were matched in frequency in the test materials with the correct words so that learning was not possible based on the frequencies of words during testing. Meaning-learning distractor materials were created such that one word-picture pair had as a distractor the other word

---

1 Thanks to Lauren Schmidt for the visual stimuli used in the experiment.

that was paired with a picture; the second word-picture pair had as a distractor another word in the language (one that was not paired with any picture).

Procedure. Participants were informed that they were playing a memory game in which they had to learn the words in an alien language by watching alien television and that they would be tested on their memory as they watched. During the familiarization phase, participants were exposed to 160 sentences in the language of the experiment with 1.5 seconds in between each trial. On 50% of all familiarization trials, a picture of an alien entered the screen and participants were instructed to indicate whether a new sentence was the same or different from the one they just heard. The new sentence was either identical or slightly perturbed: either one syllable had been switched or the picture had been switched. Participants received feedback: if correct, they heard a positive sound and their score (indicated on the screen) went up 100 points. If they were incorrect, they heard a beep.

After finishing training, participants were informed that they would be tested on how well they had learned the language. During the segmentation test (which always came first), they were asked to make 30 forced-choice judgments between words in the language and part-word distractors and given no feedback on their answers. During the meaning-learning test, they made 10 forced-choice judgments in which they saw a picture, heard both the correct word and a distractor, and were asked "which word goes with this picture?" Following the meaning-learning test, participants read a short debriefing statement.

Results. We measured performance in each of three dependent variables, memory, segmentation, and meaning learning. We first examined results in the memory task by using an ANOVA to test whether test condition affected percentage correct in the memory task. We found a significant effect of condition ($F[3,64]=4.28$, $p = .008$), reflecting poorer performance in the no meanings condition than the other conditions, perhaps resulting from the lack of memory trials in which the picture rather than the audio was changed. In follow-up pairwise comparisons, there were no significant differences in memory performance between any of the three conditions in which picture-change trials appeared, while each of these conditions differed significantly from the no meanings condition ($t(32)=2.61$, $p = .01$, $t(32)=3.00$, $p = .005$, and $t(31)=2.49$, $p = .02$).

In the segmentation task, we found a main effect of condition on segmentation ($F[3,64]=3.05$, $p = .035$), as shown in Figure 1. Planned comparisons revealed that this effect was due to a significant difference between the end position condition and each of the other conditions ($t(30)=2.07$, $p = .05$, $t(29)=2.07$, $p = .05$, and $t(32)=2.90$, $p = .007$). No other contrasts were found to be statistically significant.

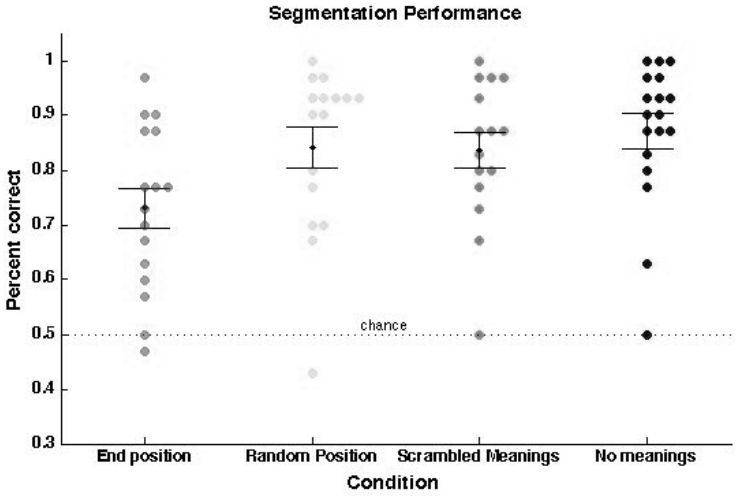

Figure 1. Segmentation performance by condition. Error bars show 95% confidence interval of the mean.

Finally, an ANOVA revealed a main effect of condition on meaning-learning performance as well ($F[2,46]=3.25$, $p = .05$), as shown in figure 2. Planned comparisons revealed a significant difference between the end position and scrambled meanings conditions ($t(29)=2.70$, $p = .01$), but no significant differences between scrambled meanings and random position ($t(29)=1.28$, $p = .21$) or scrambled meanings and end position ($t(30)=1.22$, $p = .23$).

2.2. Discussion

We used a memory task to monitor participants' attention to the familiarization part of our artificial language experiment. Although we found a significant decrease in memory performance in the auditory-only condition, we found no other significant differences in performance between conditions. Results from the memory task thus confirmed that any differences in performance between the three conditions accompanied by visual stimuli were caused by differences in the materials that the participants were presented with, rather than their attention to those materials. We now proceed to evaluate the data with respect to our three questions of interest.

First, were learners able to learn the meanings of words even in the case of uncertainty about segmentation? Results on this question were mixed. Meaning learning performance of participants in the random position condition was not significantly different from performance in the scrambled meanings condition, despite a mean above 60%. This failure to achieve significance stemmed from the extremely high variability between participants; some participants performed

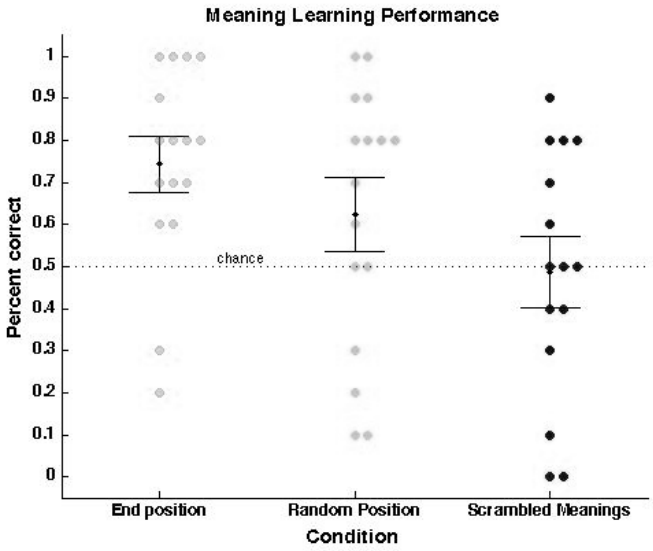

Figure 2. Meaning learning performance by condition. Error bars show 95% confidence interval of the mean.

significantly below chance in this task, while others were at ceiling. It is possible that the repetitiousness of the meaning learning trials may have forced some participants to remain consistent with their first judgments throughout the testing period. Performance in the end position condition did significantly differ from performance in the scrambled meanings condition, however. This result suggests that while it may be possible for learners to segment words from fluent speech and connect them to referents at the same time, this task is facilitated by positional regularities.

We move on to our second question, whether we observed any benefit to segmentation by adding meanings to the task. We did not observe significant differences in segmentation accuracy between the random position condition and either the auditory only or the scrambled meanings condition. However, since word-meaning mappings were not effectively learned in this condition, it remains possible that, had the participants learned the meanings, we would have observed a corresponding increase in segmentation accuracy. However, while participants did learn the correspondences between words and meanings in the end position condition, we in fact observed a significant decrease in segmentation accuracy. Although this evidence militates against a facilitation of segmentation in the presence of meaning words, it is not conclusive. The decrease in segmentation performance is most likely related to the increased salience of the last word of each sentence and could potentially counteract a small increase in segmentation of meaning words.

Finally, we address the question of whether positional regularities facilitate meaning learning. Based on the increase in meaning learning accuracy in the end position condition, we conclude that there is likely a large benefit for placing meaning-bearing words in a constant position in the sentence. One potential cause of this facilitation is the greater salience of patterns found at the ends of sentences (Endress, Scholl, & Mehler, 2005); another could be more general memory effects such as primacy and recency.

3. Computational Modeling

In order to understand better the demands of the task faced by learners in our artificial language experiment, we constructed a computational model to perform the task. While a variety of computational techniques have been used to model statistical word segmentation (e.g., Brent, 1999a; Christiansen, Allen, & Seidenberg, 1998; Perruchet & Vinter, 1998; Swingley, 2005; Yang, 2004) we chose to use the Bayesian framework presented by Brent (1999a) as its use of the generative modeling framework appeared most easily extensible to the more complex task in our experiment. While the details of the implementation of our model are quite different from those of Brent (1999), the overall effect is quite similar (for more details, see Goldwater, Griffiths, & Johnson, 2006) and our work is an extension of Brent's methods. Here we present the structure of our model and some preliminary results from initial simulations.

3.1. Model details

The task of our computational model was to take as input the same sequences of un-segmented sentences which we showed to participants in our experiment and to learn from them a lexicon—a set of words and word-meaning correspondences—from which those sentences could be produced. We defined a generative model of lexicons L, equivalent to a prior probability distribution, using the following procedure:

1. Draw n, the number of words in the lexicon, from a Poisson distribution
2. Draw each element of l, a vector containing the length of each word, from a Poisson distribution
3. Generate $\{w\}$, the set of words in the lexicon by randomly (uniformly) concatenating together syllables observed in the training corpus
4. Generate $\{m\}$, the set of meanings (including a null meaning) corresponding to each of the words in $\{w\}$, by picking the meaning of each word randomly (uniformly) from the observed meanings in the corpus

Based on this procedure, each potential lexicon of the form $(n,l,\{w\},\{m\})$ can be assigned a prior probability $P(L)$ based directly on the values of n and l. Note

that this distribution will favor short lexicons with short words, as in Brent's work (Brent, 1999b).

We next introduce a likelihood *P(s|L)* which gives the probability of a particular sentence *s* given a lexicon *L*:

$$P(s|L) = \sum_{\text{all parses } p \text{ of } s} \left(P(p \mid L) \cdot \prod_{w_i \in p} P(w_i) \right), \text{ where}$$

$$P(p|L) = \begin{cases} 1 \text{ if for all } w_i \in p,\ w_i \in \{w\} \\ 0 \text{ otherwise} \end{cases}$$

$$P(w_i) = \frac{1}{n}$$

In other words, the likelihood of *s* given lexicon *L* is given by the sum over all parses of *s* of the probability of that parse *p* times the product of all of the words in *p*. A parse of *s* is simply some partition of contiguous syllables in *s* into one or more words *w*. The probability of a parse under a particular lexicon is 1 if all the words in the parse are in the lexicon, and zero otherwise. We define the probability of any given word in a lexicon to be 1/*n* where *n* is the number of words in the lexicon; in other words, word frequencies are uniform under our model. While Brent's (1999b) model kept track of individual word frequencies, we take this uniform distribution of frequencies to be a simplifying assumption which will not interfere with segmentation in our artificial stimuli. Under this uniform distribution, the product of the probabilities of the words in the parse is simply 1/*n* to the power of the length of the parse.

Having defined a prior probability distribution *P(L)* and a likelihood function *P(s|L)*, we can now use Bayes' Rule to calculate the posterior probability of some lexicon given a sentence, *P(L|s)*:

$$\begin{aligned} P(L|s) &= \frac{P(s|L)P(L)}{P(s)} \\ &\propto P(s|L)P(L) \end{aligned}$$

This formulation will allow us to use Markov-Chain Monte Carlo (MCMC) simulations to find the posterior probability distribution of lexicons with respect to some corpus of sentences.

3.2. Simulations

We set up a MCMC sampling scheme in order to find the maximum a posteriori (MAP) lexicon—the lexicon which best fits the input corpus given the constraints of the prior. (Gelman, Carlin, Stern, & Rubin, 2004) A MCMC sampler operates by taking a random walk through the space of lexicons while making decisions about the moves in its walk based on the posterior probability

of the destination of its next move. We initialized our model by memorizing the first ten sentences of the input corpus as the ten words in the starting lexicon; each word had as its meaning the picture which accompanied that sentence. This lexicon had a posterior probability under the model which was very low (since it was unable to parse any sentences in the input corpus which were not "words" in the model). Our sampler then generated a proposal—a new lexicon to be compared to the starting lexicon and potentially accepted. This proposed lexicon was created by modifying the current lexicon in one of a number of ways:

- Two words could be merged into one single word
- One word could be split into two separate words
- A word could be deleted outright
- A new word could be added randomly
- The meaning of a word could be swapped with that of another word
- The meaning of a word could be changed randomly
- A word could be truncated

Following the Metropolis-Hastings algorithm, the sampler accepted this proposal with probability *P(new)/P(old)*—always accepting if *P(new) > P(old)*. By continuing to consider proposals for new lexicons in this manner, the sampler "walks" around the space of possible lexicons. When run for a sufficient amount of time, this algorithm guarantees convergence of the distribution of lexicons visited to the posterior distribution of lexicons given the input corpus.

We were initially interested in testing only whether the MAP lexicon under the model was the correct lexicon (including the correct word forms and the correct word-meaning correspondences). Because of the highly peaked probability landscape over the space of lexicons, in practice the sampler tended to converge very quickly to the MAP lexicon; in future work using more complex corpora with several local minima corresponding to multiple plausible lexicons, it may be necessary to use techniques such as simulated annealing (Kirkpatrick, Gelatt, & Vecchi, 1983) to obtain a higher accuracy approximation to the true posterior on lexicons.

We ran a series of simulations on input corpora generated by the same software which generated corpora for the behavioral experiments. Our preliminary results indicated that the sampler reliably converged on the generating lexicon (what we considered to be the target in our behavioral studies) within a relatively small number of proposals (between 500 and 5000). In both the random position and end position condition, both word forms and word-meaning correspondences were correctly recovered in every simulation. As expected, the distributional properties of the input given to experimental participants was sufficiently unambiguous to allow the model to succeed in every case. In the scrambled meanings condition, however, results were more interesting. While the sampler reliably converged on a series of lexicons which

contained only correct word-forms (indicating success on the segmentation), this distribution betrayed high uncertainty about the meanings of each of these word-forms (often including multiple copies of a single lexical form, linked to multiple meanings). Only the fact that the prior significantly favored shorter lexicons kept a fully ambiguous lexicon—in which each word had each meaning—from being learned.

4. Conclusions

In this paper we have introduced a novel formulation of the problem faced by young children who must identify the proper sounds to map to a novel referent, the problem of word-referent mapping. In order to study this problem we created an artificial-language paradigm modeled on the experiments of Saffran, Newport, & Aslin (1996) but including word meaning learning. This paradigm created an environment in which adult participants could be exposed to a randomly generated sample from a simple artificial language which paired un-segmented, synthesized speech with referential pictures. In addition, we created a computational model of this process by extending work on segmentation by Brent (1999) to the task of segmentation and meaning learning that we used with our adult participants.

In our adult experiments, we found that while participants learned the words of the artificial language above chance levels in every condition, the meanings of words were learned reliably above chance only when the meaning-bearing words were placed consistently at the end of sentences. These results suggest that while word meanings may be learnable by adults in such a paradigm given enough exposure, the addition of additional regularities makes the task considerably easier. Although it is tempting to extrapolate this difficulty to the case of children learning words in the absence of certainty about segmentation, there are a large number of abstractions in our experiments which make generalizations from our results highly speculative. These abstractions include but are not limited to: the small vocabulary used in our experiments, the artificial CV structure of our lexical items which kept phonotactic and syllabic structure from offering cues to segmentation (Mattys & Jusczyk, 2001), the lack of prosodic cues such as stress (Johnson & Jusczyk, 2001), and the lack of temporally synchronized cues to reference (Gogate & Bahrick, 1998) such as pointing or eye-gaze. The contribution of each of these factors to word learning is potentially significant, and in future work we hope to address the their contribution to learning by varying them systematically within this paradigm.

In our modeling work, we found that a simple formulation of the word referent problem was solvable by doing Markov-Chain Monte Carlo inference in a Bayesian model. This model was able to succeed in finding the correct words and the word/referent mappings in both the random position and end position conditions. However, when the model was run in the scrambled meanings condition, it discovered the correct word forms while producing ambiguous results for word meanings. These results indicate a qualitative fit between the

performance of our model and the results we observed in our experimental participants.

Our primary aim in future work is to investigate more precise, quantitative connections between our modeling efforts and our experiments. In particular, we would like to investigate whether our model makes the prediction that putting meaning-bearing words in utterance-final position should decrease certainty about segmentation while increasing certainty about the mappings between words and their objects. To this end, we are currently experimenting with a variety of techniques for making empirical predictions in within our simulations. One promising avenue involves asking the model to make the same type of forced-choice judgments as our participants by evaluating the ratio of probabilities between target and distractor under the model. The probability of the correct target should always be higher under the model than the probability of the distractor, but this ratio should change depending on the amount of information present in the input sequences and the plausibility of the distractor under the model. While not modeling human performance directly under the model, this technique may allow us to compare different input sequences to the model with respect the type of test that our human participants completed.

References

Brent, M. R. (1999a). An Efficient, Probabilistically Sound Algorithm for Segmentation and Word Discovery. *Machine Learning, 34*(1), 71-105.

Brent, M. R. (1999b). Speech segmentation and word discovery: A computational perspective. *Trends in Cognitive Science, 3*(8), 294-301.

Christiansen, M. H., Allen, J., & Seidenberg, M. S. (1998). Learning to segment speech using multiple cues: A connectionist model. *Language and Cognitive Processes, 13*(2), 221-268.

Dutoit, T., Pagel, V., Pierret, N., Bataille, F., & van der Vrecken, O. (1996). The MBROLA Project: Towards a Set of High-Quality Speech Synthesizers Free of Use for Non-Commercial Purposes. *Proccedings of the International Conference on Spoken Language Processing, 3*, 1393-1396.

Endress, A. D., Scholl, B. J., & Mehler, J. (2005). The role of salience in the extraction of algebraic rules. *Journal of Experimental Psychology: General, 134*(3), 406-419.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2004). *Bayesian Data Analysis* (2nd ed.). New York: Chapman & Hall/CRC.

Gogate, L. J., & Bahrick, L. E. (1998). Intersensory redundancy facilitates learning of arbitrary relations between vowel sounds and objects in seven-month-old infants. *Journal of Experimental Child Psychology, 69*(2), 133-149.

Goldwater, S., Griffiths, T. L., & Johnson, M. (2006). Contextual Dependencies in Unsupervised Word Segmentation. *Proceedings of the 44th Annual Meeting of the Association for Computational Linguistics*.

Gomez, R. L., & Gerken, L. (2000). Infant artificial language learning and language acquisition. *Trends in Cognitive Sciences, 4*(5), 178–186.

Hollich, G., Hirsh-Pasek, K., Golinkoff, R. M., Brand, R. J., Brown, E., Chung, H. L., et al. (2000). Breaking the Language Barrier: An Emergentist Coalition Model for the Origins of Word Learning. *Monographs of the Society for Research in Child Development, 65*(3), 1-135.

Johnson, E. K., & Jusczyk, P. W. (2001). Word segmentation by 8-month-olds: When speech cues count more than statistics. *Journal of Memory and Language, 44*(4), 548.

Jusczyk, P. W. (1999). How infants begin to extract words from speech. *Trends in Cognitive Sciences, 3*(9), 323-328.

Kirkpatrick, S., Gelatt, C. D., & Vecchi, M. P. (1983). Optimization by Simulated Annealing. *Science, 220*(4598), 671.

Mattys, S. L., & Jusczyk, P. W. (2001). Phonotactic cues for segmentation of fluent speech by infants. *Cognition, 78*(2), 91-121.

Perruchet, P., & Vinter, A. (1998). PARSER: A model for word segmentation. *Journal of Memory and Language, 39*(246-263).

Saffran, J. R., Newport, E. L., & Aslin, R. N. (1996). Word segmentation: The role of distributional cues. *Journal of Memory and Language, 35*(4), 606-621.

Swingley, D. (2005). Statistical clustering and the contents of the infant vocabulary. *Cognitive Psychology, 50*, 86-132.

Yang, C. D. (2004). Universal Grammar, statistics or both? *Trends in Cognitive Sciences, 8*(10).

Yu, C., & Smith, L. (in press). Rapid Word Learning under Uncertainty via Cross-Situational Statistics. *Psychological Science*.