"Less is More" in Bayesian word segmentation: When cognitively plausible learners outperform the ideal

Lawrence Phillips (lawphill@uci.edu)
Department of Cognitive Sciences, 2235
SBSG Irvine, CA 92697 USA

Lisa Pearl (lpearl@uci.edu)
Department of Cognitive Sciences, 2314
SBSG Irvine, CA 92697 USA

Abstract

Purely statistical models have accounted for infants' early ability to segment words out of fluent speech, with Bayesian models performing best (Goldwater et al. 2009). Yet these models often incorporate unlikely assumptions, such as infants having unlimited processing and memory resources and knowing the full inventory of phonemes in their native language. Following Pearl, et al. (2011), we explore the impact of these assumptions on Bayesian learners by utilizing syllables as the basic unit of representation. We find a significant "Less is More" effect (Pearl et al 2011; Newport 1990) where memory and processing constraints appear to help, rather than hinder, performance. Further, this effect is more robust than earlier results and we suggest this is due a relaxing of the assumption of phonemic knowledge, demonstrating the importance of basic assumptions such as unit of representation. We argue that more cognitively plausible assumptions improve our understanding of language acquisition.

Keywords: language acquisition; Bayesian modeling; cognitively plausible learning; less is more; statistical learning; word segmentation

Introduction

Knowledge of words plays a crucial role in language acquisition but requires a child to identify words out of fluent speech. Children seem to accomplish this word segmentation very early (~7.5 months (Jusczyk & Aslin 1995; Echols et al. 1997; Jusczyk et al., 1993a)), and therefore many strategies have been proposed for this early success. One popular explanation for initial language learning relies purely on distributional information, rather than language-specific biases. This idea is bolstered by findings that infants keep track of the statistical regularities in speech (Saffran et al. 1996), and because languages vary greatly in their cues to word boundaries which would weaken the use of language specific knowledge. One very successful, purely distributional, learning approach uses Bayesian inference (Goldwater, Griffith & Johnson 2009 (***GGJ***), Pearl, Goldwater & Steyvers 2011 (***PGS***)). However, these Bayesian models incorporate modeling assumptions that are unlikely to be true. Both have assumed that the basic unit of representation available to the infant is the phoneme. We will argue from experimental evidence that syllables (or syllable-like representations) are a more natural representation for infants at this stage of acquisition. In addition, GGJ conducted an ideal learner analysis, which assumes unlimited processing and memory resources for the learner. PGS investigated the impact of this assumption, finding a limited "Less is More" effect (Newport 1990) where cognitive resource limitations help, rather than hinder, some Bayesian learners. We examine the effect of the phoneme assumption in addition to these cognitive resource assumptions. We find not only that syllable-based Bayesian learners can do well at word segmentation but also a much more robust "Less is More" effect in our constrained Bayesian learners. This suggests that the unit of representation for models of language acquisition plays a crucial role. Here, using more cognitively plausible assumptions showcases a surprising learning effect, the "Less is More" effect that has been hypothesized to explain language acquisition success in children.

The syllables as the representational unit

The first evidence that infants possess categorical representations of syllabic units appears at 3 months: Eimas (1999) finds that infants have categorical representations of syllables whereas infants at this age have no categorical representation of phonemes. Since word segmentation first occurs around 7.5 months (Jusczyk & Aslin 1995), it is likely that infants have robust access to syllables at this age. In contrast, knowledge of phonemes does not occur until approximately 10 months (Werker & Tees 1984) making it unlikely the learner has adult knowledge of their native language phonemes during the initial stages of word segmentation. Although it is possible that word segmentation and phoneme learning bootstrap from one another, we consider a more conservative approach which assumes infants only have access to syllabic information.

While the success of previous statistical word segmentation models is heartening, how dependent is their success on the assumption of the phoneme as a representational unit? With this question in mind, we modify existing phoneme-based statistical models of word segmentation that use Bayesian inference (GGJ, PGS) to operate over syllables. All of our modified Bayesian learners treat syllables as atomic units in the same way phonemes are thought of as atomic units. This mimics the performance of infants who are able to discriminate between syllables such as /ba/, /bu/, and /lu/, but who are unable to

recognize the phonemic similarity between /ba/ and /bu/ which does not exist between /ba/ and /lu/ (Jusczyk & Derrah 1987).

Utilizing syllables alleviates the learning problem somewhat because it reduces the number of potential boundary positions (e.g., a baby has three syllables but five phonemes). However, a potential sparse data problem then surfaces: A model operating over English phonemes must track statistics over approximately 40 units; a model operating over English syllables must track statistics over approximately 4000 units, while using less data than a phoneme-based model since there are fewer syllable tokens than phoneme tokens. This increases the statistical difficulty of the task tremendously. Additionally, because syllables are treated as atomic, almost all phonotactic information about English is lost in the model. Although previous work (e.g. Gambell & Yang 2006) shows that heuristic syllable-based models can perform quite well, it is unclear a priori whether a distributional learner with phonemes or syllables will produce better results for Bayesian word segmentation, due to the tradeoffs just mentioned.

In changing our unit of representation, we attempt to create a more psychologically faithful model of word segmentation. To foreshadow our results, we show that successful Bayesian word segmentation does not depend on the phoneme assumption. Moreover, by utilizing a more cognitively plausible unit of representation, we find a much more robust "Less is More" effect. The success of our models demonstrates the effectiveness of this purely statistical approach. Replicating and extending results from PGS concerning the surprising utility of processing constraints for Bayesian word segmentation. This suggests that the task of word segmentation may be structured to be more easily learned with strong memory limitations, such as those that infants have. Moreover, Bayesian models may be on the right track with respect to the kind of strategies infants are using during early word segmentation, since the learners demonstrate this "Less is More" behavior, and infants are thought to as well. In addition, the fact that this pattern of results was only hinted at by the phoneme-based models of PGS means that the unit of representation for models of language acquisition has a strong, non-trivial effect on the results found.

Methods

Corpus

We test our syllable-based models using English child-directed speech from the Pearl-Brent corpus (CHILDES: MacWhinney, 2000). This modification of the Brent corpus contains 100 hours of child-directed speech from 16 mother-child pairs. We restrict ourselves, however, to child-directed utterances before 9 months of age, leaving 28,391 utterances (3.4 words per utterance, 10.4 phonemes per utterance, 4.2 syllables per utterance, on average).

While there are many ways to syllabify a corpus automatically, we opted for a two-stage approach. First, where we have human judgments of syllabification we used them; second, when not, we automatically syllabify our corpus in a language-independent way. We take human judgments of syllabification from the MRC Psycholinguistic Database (Wilson 1988), but not all words in the Pearl-Brent corpus have syllabifications in the MRC dictionary. To solve this problem we used the Maximum-Onset Principle to syllabify all remaining words. This principle states that the onset of any syllable should be as large as possible while still remaining a valid word-initial cluster. We use this principle out of convenience for the kind of syllabification that infants might possess. Given a lack of experimental evidence as to the exact nature of infant syllabification, this representation is likely only an approximation. Approximately 25% of lexical items were syllabified automatically. Only 3.6% of human judgments on our items differ from automatic syllabification. Each unique syllable is then treated as a single, indivisible unit losing all sub-syllabic phonetic (and phonotactic) information.

Models

Bayesian models are well suited to questions of language acquisition because they explicitly distinguish between the learner's pre-existing beliefs (prior) and how the learner evaluates incoming data (likelihood), using Bayes' theorem:

$$P(h|d) \propto P(d|h)P(h)$$

The Bayesian learners we use are those of GGJ as well as the constrained learners of PGS. All learners are based on the same underlying hierarchical Bayesian models developed by GGJ. The first of these models assumes independence between words (a *unigram* assumption) while the second assumes words depend only on the word before them (a *bigram* assumption). To encode these assumptions into the model, GGJ use a *Dirichlet Process* (Ferguson, 1973), which supposes that the observed sequence of words $w_1 \ldots w_n$ is generated sequentially using a probabilistic generative process. In the unigram case, the identity of the ith word is chosen according to:

$$P(w_i = w|w_1 \ldots w_{i-1}) = \frac{n_{i-1}(w)+\alpha P_0(w)}{i-1+\alpha} \quad (1)$$

where $n_{i-1}(w)$ is the number of times w appears in the previous $i-1$ words, α is a free parameter of the model, and P_0 is a *base distribution* specifying the probability that a novel word will consist of the phonemes $x_1 \ldots x_m$:

$$P(w = x_1 \ldots x_m) = \prod_{j=1}^{m} P(x_j) \quad (2)$$

In the bigram case, a *hierarchical Dirichlet Process* (Teh et al. 2006) is used. This model additionally tracks the frequencies of two-word sequences and is defined as in:

$$P(w_i = w|w_{i-1} = w', w_1 \ldots w_{i-2}) = \frac{n_{i-1}(w',w)+\beta P_1(w)}{n_{i-1}(w')+\beta} \quad (3)$$

$$P_1(w_i = w) = \frac{b_{i-1}(w)+\gamma P_0(w)}{b_{i-1}+\gamma} \quad (4)$$

where $n_{i-1}(w',w)$ is the number of times the bigram (w',w) has occurred in the first $i-1$ words, $b_{i-1}(w)$ is the number of times w has occurred as the second word of a bigram, b_{i-1} is the total number of bigrams, and β and γ are free model parameters.

In both the unigram and bigram case, this generative model implicitly incorporates preferences for smaller lexicons by preferring words that appear frequently (due to (1) and (3)) as well as shorter words in the lexicon (due to (2) and (4)). The ideal learner based on this model is fit using Gibbs sampling (Geman & Geman 1984), run over the entire corpus, sampling every potential word boundary 20,000 times. GGJ found that their bigram ideal learner performed better than their unigram ideal learner, so we begin by examining this distinction in our syllable-based Bayesian learners. In addition, we will consider the constrained learners that PGS investigated—incorporating processing and memory constraints.

The Dynamic Programming Maximization (DPM) learner incorporates a basic processing limitation: linguistic processing occurs online rather than in batch after a period of data collection. Thus, the DPM learner processes one utterance at a time, rather than processing the entire corpus at once. This learner uses the Viterbi algorithm to converge on the optimal word segmentation for the current utterance, conditioned on the utterances seen so far. In all other aspects, the DPM learner is essentially identical to the Ideal model: it has perfect memory for previous utterances and unlimited processing resources.

The Dynamic Programming Sampling (DPS) learner is similar to the DPM learner in processing utterances incrementally, but is additionally motivated by the idea that infants, and human beings in general, are not ideally rational. This could mean that infants do not *always* select the best segmentation. Instead, infants select segmentations probabilistically. So, they will often choose the best segmentation but occasionally choose less likely alternatives, based on the likelihood of the various segmentation alternatives. To implement this, the DPS learner uses the Forward algorithm to compute the likelihood of all possible segmentations and then chooses a segmentation based on the calculated distribution.

The Decayed Markov Chain Monte Carlo (DMCMC) learner also processes data incrementally, but uses a DMCMC algorithm (Marthi et al. 2002) to implement a memory constraint. This learner is similar to the original GGJ ideal learner in that it uses Gibbs sampling. However, the DMCMC learner does not sample all boundaries; instead, it samples some number s of previous boundaries using the decayed function b_a^{-d} to select the boundary to sample, where b_a is the number of potential boundary locations between b and the end of the current utterance a and d is the decay rate. Thus, the further b is from the end of the current utterance, the less likely it is to be sampled. Additionally, larger values of d indicate a stricter memory constraint. All our results here use a set, non-optimized value for d of 1.5, which was chosen to implement a heavy memory constraint. Having sampled a set of boundaries, the DMCMC learner can then update its beliefs about those boundaries and subsequently update its lexicon.[1] Because of the decay function, the DMCMC's sampling is biased towards boundaries in recently seen utterances and thus the DMCMC learner implements a recency effect.

In addition to comparing our syllable-based learners against the original phoneme-based learners, we also compare our learners against other syllable-based learners. The first baseline is the Transitional Probability (TP) model based on Gambell & Yang (2006), which calculates TPs over syllables and places boundaries at all local minima. Our second baseline is a "Syllable=Word" learner which simply assumes that all syllables are words (a strategy that can be very useful in languages containing many monosyllabic words, like English).

Results

We measure our results in terms of precision, recall and F-score, where precision is defined as (5) and recall is defined as (6):

$$Precision = \frac{\#\ correct}{\#\ guessed} \quad (5)$$

$$Recall = \frac{\#\ correct}{\#\ actual} \quad (6)$$

F-score is the harmonic mean of the two:

$$F - score = \frac{2*Precision*Recall}{Precision+Recall} \quad (7)$$

Precision and recall are considered jointly, through the harmonic mean, because it is possible for learners to succeed on one measure while failing on the other. For instance, a learner that posits only a single boundary scores 100% on precision if that boundary is correct. In comparison, the same learner will have just over 0% recall. Similarly, a learner could posit boundaries at every position, producing a 100% recall with very low precision because many of the boundaries were false. As the F-score balances these two measures, a high F-score indicates the learner is succeeding at both precision and recall. We can make these measurements over individual word tokens, word boundaries, and lexical items.

In order to prevent overfitting, we train each learner on 90% of the corpus and then test the learner on the remaining 10%. This train-test validation was done five times for each learner. Given the probabilistic nature of our learners, all

[1] All DMCMC learners sample s=20,000 boundaries per utterance. According to PGS, this works out to approximately 89% less processing than the original ideal learner in GGJ, which samples every boundary 20,000 times.

results presented here are averaged over the five iterations to ensure the validity of each learner's performance.

Table 1 shows the F-score for word tokens over all of the syllable-based learners. First, we observe that, in all cases, the Bayesian bigram learners outperform their unigram equivalents. In the unigram case, all constrained learners (DPM, DPS, DMCMC) significantly outperform the ideal learner; in contrast, in the bigram case this is true for the DMCMC learners only. This indicates that constrained learning helps generally if statistics cannot be tracked across words. However, if bigram statistics can be tracked, a memory constraint is only beneficial for the DMCMC strategy. Additionally, all learners outperform the TP baseline learner and all bigram learners outperform the Syl = Word baseline.

	Unigram	Bigram
Ideal	53.12	77.06
DPM	**58.76**	75.08
DPS	**63.68**	77.77
DMCMC	**55.12**	**86.26**
TP	43.98	
Syl=Word	72.41	

Table 1. Word token F-scores across all syllable-based models. Constrained Bayesian learners that significantly outperform their ideal counterpart ($p<.05$) are in bold.

	Syl-U	Phon-U	Syl-B	Phon-B
Ideal	53.1	**54.8**	**77.1**	71.5
DPM	58.8	**65.9**	**75.1**	69.4
DPS	**63.7**	58.5	**77.8**	39.8
DMCMC	55.1	**67.8**	**86.3**	73.0

Table 2. Token F-scores for syllable-based (Syl) vs. phoneme-based (Phon) models, comparing Unigram (U) and bigram (B) learners. Learners in bold outperform their baseline counterparts.

Clearly our syllable-based learners perform well, but are syllables a better unit of representation than phonemes for this task? Table 2 compares our syllable-based learners with the original phoneme-based models of PGS. We see that in the unigram case, phoneme-based learners outperform their syllable-based counterparts, except in the case of the DPS learner. In the bigram case, however, all syllable-based models outperform their phoneme-based equivalents. This suggests that the bigram assumption is crucial to a syllable-based learner. We speculate that this is due to an additional source of information that the bigram learner has access to. In particular, because the unigram learner assumes that words are independent of one another, the TPs between syllables are the only source of boundary information. Because there are roughly 4000 syllables, there will often be cases where a problem of sparse data arises. In contrast, the bigram learner has access to the boundary information inherent in word bigrams, in addition to TPs. These word bigrams may help supplement the sparseness of the TP data.

A desired behavior for all learners is undersegmentation since children are known to undersegment the input they receive (Peters 1983). All of our Bayesian learners exhibit this behavior. This can be seen by comparing the values of boundary precision and recall. High boundary precision (indicating the boundaries are often correct) but low recall (indicating not enough boundaries are put in) indicates general undersegmentation, whereas high boundary recall (indicating a lot of boundaries are put in) but low boundary precision (indicating the boundaries are not often correct) indicates oversegmentation. Although this trend of undersegmentation exists for both unigram and bigram learners, we present data only on our bigram learners since the results are qualitatively similar. Table 3 shows the boundary precision and recall for all Bayesian bigram and comparison learners. The Syl=Word baseline learner tends to oversegment, so although it performs much better than the TP learner and the Bayesian unigram learners, its error pattern does not match what we expect from infants. In contrast, all of our Bayesian learners are producing more undersegmentations than oversegmentations. Table 4 presents sample segmentation errors from the Ideal and DMCMC bigram learners.

	Boundary Precision	Boundary Recall
Ideal	**96.50**	80.45
DPM	**96.49**	76.21
DPS	**95.78**	79.72
DMCMC	**94.11**	91.57
TP	**90.00**	53.14
Syl = Word	76.26	**100**

Table 3. Boundary precision and recall for all bigram Bayesian and comparison learners.

Bigram Ideal	Bigram DMCMC
putit away	put it away
Iloveyou	I love you
Let'ssee what that *feltlike*	Let's see what that *feltlike*
If *youdon't* like it	*Ifyou* don't like it

Table 4. Example output from Bigram Ideal and DMCMC learners. Undersegmentation is marked in italics.

To explain the difference between our ideal and constrained learner results, we can examine the token and lexicon item recall scores, as shown in Table 5. We observe that both the DMCMC learners identify fewer word types than their ideal learner counterparts, as shown by their comparatively low lexicon recall scores. The token recall score for the DMCMC learners, however, is higher than their ideal learner counterparts. Since this requires the DMCMC learners to identify more word tokens from a smaller stock of lexical items, it can be inferred that these

DMCMC learners are identifying more frequently occurring words than the ideal learners.

	Token Recall	Lexicon Recall
Uni-Ideal	44.96	73.44
Uni-DMCMC	48.09	68.9
Bi-Ideal	72.47	79.69
Bi-DMCMC	85.43	76.84

Table 5. Token and lexicon recall for the Ideal and DMCMC learners. Lower lexicon recall with higher token recall implies that the DMCMC learners identify more frequently occurring words.

Discussion

Our results support two broad findings. First, we find that memory-constrained learners outperform their "ideal" equivalents, which we take as support for the "Less is More" hypothesis (Newport 1990). In particular, limited cognitive resources, rather than hurting learner, seem to help word segmentation. Second, because this effect was obscured in the phoneme-based learners of PGS, we argue that the unit of representation posited by a model of language acquisition has a crucial impact on the results found. In particular, making more cognitively plausible assumptions may yield answers to the puzzling behaviors we observe—namely, that children, who are more cognitively limited than adults, nonetheless are far more successful at language acquisition.

What exactly is causing the "Less is More" effect here? Perhaps it is due to the properties of online vs. batch unsupervised probabilistic learning algorithms. Liang & Klein (2009) show that for unsupervised models using Expectation-Maximization, online models not only converge more quickly than batch models, but, also in cases as varied as word segmentation, part-of-speech induction and document classification, can actually outperform their batch equivalents. However, this explanation fails to account for our results in two ways: (a) the most direct online equivalent of our batch model (the DPM learner) actually performs worse than the Ideal model, and (b) this does not explain the performance boost caused by sub-optimal segmentation (the DPS learner).

Perhaps the answer lies in the kinds of words these models identify. We find, as in table 5, that our ideal bigram learner segments 72.5% of the words in the input, building a lexicon that contains 80% of the actual word-types it encounters. Yet we find that a learner with memory constraints (the DMCMC learner) can successfully segment 85% of the words in the input, although this makes up only 76.8% of the word-types encountered. This suggests that while an ideal learner identifies more lexical items, the memory-constrained learner identifies more *frequent* lexical items. Not only is this true in both the unigram and bigram syllable-based learners, but it is also true of the equivalent phoneme-based learners of PGS. The robustness of this phenomenon suggests that, irrespective of the representational unit, memory-constrained learners are biased towards identifying more commonly occurring units, a potentially useful bias in language acquisition.

In effect, this strategy in word segmentation may help in learning the *important* things. Although this has been hypothesized by the literature on "Less is More" in artificial language learning (Kersten & Earles 2001; Cochran et al. 1999), we are unaware of experimental support for why constrained processing helps in real language acquisition. The fact that we can help to explain, from a computational perspective, why "Less is More" is beneficial highlights a very major contribution computational modeling can make to developmental research more generally.

For our claim regarding the impact of the unit of representation, we can compare the syllable-based learner results with those of phoneme-based learners. Table 2 highlights a number of crucial distinctions. First, and most basically, syllable-based learners perform well, and in the bigram case better than phoneme-based learners. This suggests that the tradeoff between number of potential boundaries and number of potential transitional probabilities works out in favor of the syllable-based learner. This underscores the utility of a Bayesian inference strategy for the initial stages of word segmentation – without access to phonotactics, stress, acoustic cues, or innate linguistic knowledge, a learner can be very successful at segmenting words from fluent speech.

Still, there is a major difference in the performance of the sub-optimal (DPS) learner – the syllable-based DPS learner has comparable performance to the Ideal learner while its phoneme-based equivalent suffers greatly. We speculate that this is due to the number of potential segmentations the phoneme-based learner considers, compared to the syllable-based learner since the DPS learner chooses a segmentation probabilistically, the phoneme-based learner may be more easily led astray in the initial stages of segmentation, and never recover. In addition, we also notice a strengthening of the "Less is More" effect in the syllable-based learner, compared to its phoneme-based counterpart (Ideal vs. DMCMC). By making more realistic assumptions about the learner's unit of representation, we also create a learner that exhibits the kind of behavior that infants show. This highlights one benefit of pursuing more cognitively plausible computational models, as opposed to models that are more idealized.

In that vein, there are a number of areas where we could improve the existing syllable-based Bayesian learners. First, some segmental cue information is likely available to infants such as phonotactics or articulatory cues. Similarly, suprasegmental cues such as primary stress are known to affect infant word segmentation (Jusczyk et al. 1999) and there is evidence that stressed and unstressed syllables are represented separately in infants (Pelucchi, Hay, & Saffran 2009). Finally, the exact form which infants use to represent syllables is unclear. While it is our view that syllabification must be learned by infants, we make no attempt here to

explain by what means this occurs. When one looks cross-linguistically, languages treat syllabification in very different ways. In addition, languages vary significantly on the number of syllable types they have – languages such as English number their unique syllables in the thousands, while some languages, like Japanese, have very few unique syllables. To ensure that our pattern of results is truly representative of word segmentation *generally* and not just in English, syllable-based word segmentation models must be tested across multiple languages.

In conclusion, this study highlights the benefits of using empirical research from psychology to inform decisions on how to model language acquisition: not only can we identify the strategies that are likely to be used by children, but we may also discover potential explanations for existing, sometimes puzzling, observations about child language acquisition, as with the "Less is More" hypothesis.

Acknowledgments

We would like to thank Caroline Wagenaar and James White for their help on syllabifying the Pearl-Brent corpus. In addition, we are very grateful to Robert Daland, Constantine Lignos, and the audiences at the Psycho-Computational Models of Human Language Acquisition 2012 and the Linguistic Society of America meeting in 2012 for their helpful comments.

References

Brent, M.R. & Siskind, J.M. 2001. The role of exposure to isolated words in early vocabulary development. *Cognition, 81*, 31-44.

Cochran, B., McDonald, J. & Parault, S. 1999. Too smart for their own good: The disadvantage of superior processing capacity for adult language learners. *Journal of Memory and Language*, 41, 30-58.

Echols, C.H., Crowhurst, M.J. & Childers, J.B. 1997. The perception of rhythmic units in speech by infants and adults. *Journal of Memory and Language*, 36, 202-225.

Eimas, P.D. 1999. Segmental and syllabic representations in the perception of speech by young infants. *Journal of the Acoustical Society of America*, 105(3), 1901-1911.

Ferguson, T. 1973. A Bayesian analysis of some nonparametric problems. *Annals of Statistics, I*, 209-230.

Frank, M. C., Goodman, N. D., & Tenenbaum, J. 2009. Using speakers' referential intentions to model early cross situational word learning. *Psychological Science, 20*, 579 585.

Gambell, T. & Yang, C. 2006. Word Segmentation: Quick but not dirty. Manuscript. New Haven: Yale University

Geman S. & Geman D. 1984. Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6, 721-741.

Goldwater, S., Griffiths, T. & Johnson, M. 2009. A Bayesian framework for word segmentation: Exploring the effects of context. *Cognition 112*(1), 21-54.

Jusczyk, P.W. & Derrah, C. 1987. Representation of speech sounds by young infants. *Developmental Psychology*, 23(5), 648-654.

Jusczyk, P.W., Cutler, A. & Redanz, N.J. 1993a. Infants' preference for the predominant stress patterns of English words. *Child Development*, 64(3), 675-687.

Jusczyk, P.W., Friederici, A.D., Wessels, J.M.I., Svenkerud, V.Y. & Jusczyk, A.M. 1993b. Infants' sensitivity to the sound patterns of native language words. *Journal of Memory and Language*, 32, 402-420.

Jusczyk, P.W., Luce, P.A. & Charles-Luce, J. 1994. Infants' sensitivity to phonotactic patterns in the native language. *Journal of Memory and Language*, 33, 630-645.

Jusczyk, P.W., Houston, D.M. & Newsome, M. 1999. The beginnings of word segmentation in English learning infants. *Cognitive Psychology*, 39, 159-207.

Kersten, A.W. & Earles, J.L. 2001. Less really is more for adults learning a miniature artificial language. *Journal of Memory and Language*, 44, 250-273.

Liang, P. & Klein, D. 2009. Online EM for unsupervised models. *Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the ACL*, 611-619.

MacWhinney, B. 2000. *The CHILDES project: Tools for analyzing talk*. Mahwah, NJ: Lawrence Erlbaum Associates.

Marthi, B., Pasula, H., Russell, S. & Peres, Y., et al. 2002. Decayed MCMC filtering. In *Proceedings of 18th UAI* 319-326.

Newport, E. 1990. Maturational constraints on language learning. *Cognitive Science, 14*, 11-28.

Pearl, L., Goldwater, S., & Steyvers, M. 2011. Online Learning Mechanisms for Bayesian Models of Word Segmentation, *Research on Language and Computation*, special issue on computational models of language acquisition. DOI 10.1007/s11168-011-9074-5.

Pelucchi, B., Hay, J., & Saffran, J. 2009. Learning in reverse: Eight-month-old infants track backward transitional probabilities. *Cognition, 113*, 244-247.

Peters, A. 1983. *The Units of Language Acquisition, Monographs in Applied Psycholinguistics*, New York: Cambridge University Press.

Saffran, J.R., Aslin, R.N. & Newport, E.L. 1996. Statistical learning by 8-Month-Old Infants. *Science, 274*, 1926 1928.

Teh, Y., Jordan, M., Beal, M., & Blei, D. 2006. Heirarchical Dirichlet processes. *Journal of the American Statistical Association, 101*(476), 1566-1581.

Wilson, M.D. 1988. The MRC Psycholinguistic Database: Machine Readable Dictionary, Version 2, *Behavioral Research Methods, Instruments and Computers*, 20 6-11.

Werker, J.F. & Tees, R.C. 1984. Cross-language speech perception: Evidence for perceptual reorganization during the first year of life. *Infant Behavior & Development, 7*, 49-63.

Xu, F. & Tenenbaum, J.B. 2007. Word learning as Bayesian inference. *Psychological Review, 114*(2), 245-272.