How Ideal Are We? Incorporating Human Limitations into Bayesian Models of Word Segmentation

Lisa Pearl[1], Sharon Goldwater[2], and Mark Steyvers[1]
[1]University of California, Irvine and [2]University of Edinburgh

1. Introduction

Word segmentation is one of the first problems infants must solve during language acquisition, where words must be identified in fluent speech. A number of weak cues to word boundaries are present in fluent speech, and there is evidence that infants are able to use many of these, including phonotactics (Mattys et al., 1999), allophonic variation (Jusczyk et al., 1999b), metrical (stress) patterns (Morgan et al., 1995; Jusczyk et al., 1999c), effects of coarticulation (Johnson and Jusczyk, 2001), and statistical regularities among sequences of syllables (Saffran et al., 1996). However, with the exception of the last cue, all these cues are language-dependent, in the sense that the infant must know what some of the words of the language are in order to make use of the cue. For example, in order to know what the common stress pattern is for words of her native language, an infant has to know some words already. Since the point of word segmentation is to identify words in the first place, this seems to present a chicken-and-egg problem. Statistical learning has generated a lot of interest because it may be a way out of this problem, by providing an initial language-independent way to identify some words. Since infants appear to use statistical cues earlier than other kinds of cues (Thiessen & Saffran, 2003), statistical learning strategies could indeed be providing an initial bootstrapping for word segmentation.

Statistical learning is often associated with transitional probability (Saffran et al., 1996), which has been shown to perform poorly on realistic child-directed speech (calculated over syllables: Gambell & Yang (2006); calculated over phonemes: Brent (1999)). However, a promising alternative approach is Bayesian learning. Researchers have recently shown that Bayesian model predictions are consistent with human behavior in various cognitive domains, including language acquisition (e.g., Griffiths & Tenenbaum, 2005; Xu & Tenenbaum, 2007). Goldwater, Griffiths, and Johnson (2007) (henceforth GGJ) found that Bayesian learners performed very well on the word segmentation problem when given realistic child-directed speech samples, especially when compared to transitional probability learners.

One critique of GGJ's model is that it is an "ideal learner" or "rational"

* We would like to thank Tom Griffiths, Michael Frank, and audiences at the Computational Modeling of Language Learning seminar at UC Irvine, the Psychocomputational Models of Human Language Acquisition workshop in 2009, and the Boston University Conference on Language Development in 2009. In addition, this work was supported by NSF grant BCS-0843896 to LP.

model (Oaksford & Chater, 1998). Rational models seek to explain why humans behave as they do, given the task and data they encounter. Specifically, if we view language acquisition as an induction problem, rational models seek the optimal solution for that induction problem, given certain assumptions about the representation of knowledge in the human mind and available information from the learner's environment. However, such models typically avoid a question addressed by more traditional psychological models: namely, *how* the observed behavior is generated given human limitations on memory and processing. That is, rational models do not address how humans could identify the optimal solution given cognitive limitations; rather, rational models identify what the optimal solution is given the data, and may use computational procedures that humans cannot use.

In this paper, we investigate how to incorporate human limitations into the Bayesian model of GGJ. In particular, we create several constrained learning algorithms that take limitations on memory and processing into account. Notably, our constrained learners still use the same generative probabilistic Bayesian model that an ideal Bayesian learner would. It is only the learning process that differs.

After testing our constrained learners on child-directed speech, we find that a constrained learner may not necessarily benefit from the biases and assumptions that are helpful for an ideal learner. Specifically, some assumptions about the data representation may only be useful when the learner has sufficient processing resources to exploit this information. This suggests that the transition from an ideal model of human cognition for a particular acquisition problem to a constrained model is not straight-forward.

In addition, one notable discovery is that our constrained learners are able to utilize the statistical information in the data quite well, sometimes achieving performance at or even above that of the ideal learner model. Also, most of our constrained learners significantly out-perform other cognitively motivated statistical learning strategies that use transitional probability.

2. Bayesian Word Segmentation

The starting point of our research is the work of GGJ, who develop two models of word segmentation within a Bayesian framework. The Bayesian learner seeks to identify some internalized representation (e.g., involving a lexicon of words) that provides a good explanation for how the observed data (e.g., the utterances) were generated. A good explanation should both account for the observed data and conform to the learner's prior expectations about the internalized representation. For GGJ, the learner is presented with some data *d*, which is an unsegmented corpus of phonemes[1]. The learner seeks a hypothesis

[1] Note that the units over which this model operates are phonemes, rather than phonetic features (Christiansen et al., 1998) or syllables (Swingley, 2005). This is not

h, given *d*, that is a sequence of words that matches the observed data and also has a high prior probability. This idea can be stated formally using Bayes' rule:

(1) $P(h \mid d) \propto P(d \mid h)P(h)$

A hypothesis matches the data if concatenating words from that hypothesis can create the data. Since a hypothesis is only a sequence of words, if the hypothesis sequence matches the observed sequence of phonemes, the likelihood ($P(d|h)$) is 1; if the hypothesis sequence does not match the observed sequence, the likelihood is 0. For example, hypotheses consistent with the observation sequence *lookatthedoggie* (we use orthographic rather than phonemic transcriptions here for clarity) include *lookatthedoggie, look at the doggie, lo oka t th edo ggie,* and *l o o k a t t h e d o g g i e*. Inconsistent hypotheses, for which $P(d|h) = 0$, include *i like pizza, a b c,* and *lookatthat*.

Since the likelihood is either 0 or 1, all of the work in the model is done by the prior distribution over hypotheses. A hypothesis has high prior probability if it accords with biases or assumptions the learner has about the internalized representation (encoded in $P(h)$). For GGJ, the prior of *h* encodes the intuitions that words should be relatively short, and the lexicon should be relatively small. In addition, each of the two models encodes a different expectation about word behavior: in the *unigram* model, the learner assumes that words are statistically independent (i.e., the context preceding the word is not predictive); in the *bigram* model, words are assumed to be predictive units. The optimal hypothesis is the one with the highest probability, given the data ($P(h|d)$). Importantly, only the counts of hypothesized lexicon items are required to calculate $P(h)$, and thus to calculate the highest probability segmentation (see Appendix for details). This means that this probabilistic model can be used by *any* learner able to track the frequency of lexicon items, not just an ideal learner.

3. Ideal and Constrained Bayesian Inference

3.1. Ideal learners

To evaluate the performance of both the unigram and bigram Bayesian models for an ideal learner, GGJ used Gibbs sampling, a stochastic search procedure often used for ideal learner inference problems. The Gibbs sampling algorithm, when used for word segmentation, iterates over the entire corpus of utterances multiple times, identifying segmentations with high probability for each utterance by deciding for each word boundary whether to insert or remove that boundary. The algorithm decides each utterance's segmentation based on evidence from the entire corpus (i.e., every other utterance in the corpus) – and so a learner using Gibbs sampling must hold all utterances ever heard in memory to make this estimation. GGJ found a good approximation for

---

uncontroversial, as it makes the model insensitive to feature-based similarity between sounds and abstracts away from many details of phonetic and acoustic variation.

segmentation by allowing the ideal learner to sample each boundary in the corpus 20000 times. They discovered that an ideal bigram learner, which believes words are predictive, achieves far more successful segmentation than an ideal unigram learner, which assumes words are not predictive. Moreover, a unigram ideal learner will severely under-segment the corpus, identifying common collocations as single words (e.g., *you want* segmented as *youwant*). This is most likely because the only way a unigram learner can capture strongly predictive word sequences is to assume those words are actually a single word. These ideal learning results tells us about the expected behavior in learners if humans were capable of these memory and processing feats – that is, what in principle are the useful biases for humans to use, given the available data.

3.2. Constrained Learners

If we were to embody GGJ's ideal learner as a human learner, we would have a human capable of remembering all the utterances she was exposed to in enough detail to sample each word boundary and able to do a significant amount of processing (recall that each boundary in the corpus is sampled 20000 times). Here we investigate three algorithms that make more cognitively plausible assumptions about memory and processing, asking how such limitations might affect the learner's ability to find good segmentation solutions.

To simulate limited resources, all our constrained algorithms process one utterance at a time rather than processing the entire corpus simultaneously. We, like GGJ and other previous work, assume that utterance boundaries are available in the input to the learner since they are marked by pauses. Recall that under GGJ's Bayesian model, the only information necessary to compute the probability of any particular segmentation of an utterance is the number of times each word (or bigram, in the case of the bigram model) has occurred in the model's current estimation of the segmentation. Thus, in each of our constrained learners, the counts of lexicon items are updated after processing each utterance.

We first tried to find the most direct translation of the ideal learner to a constrained learner whose only limitation is that utterances must be processed one at a time. One idea for this is an algorithm that uses a standard method in computer science known as dynamic programming to efficiently compute the probability of every possible segmentation of the current utterance, given the current lexicon. It then chooses the segmentation with the highest probability, adds the words from that segmentation to the lexicon, and moves to the next utterance. We refer to this algorithm as Dynamic Programming Maximization (DPM), because it chooses the maximum probability segmentation for each utterance. Details of this algorithm are described by Brent (1999), who uses the same algorithm on a different probabilistic model (which means that the probabilities of the possible segmentations would be different for his learner than the probabilities calculated here for our learners).

We then created a variant, called Dynamic Programming Sampling (DPS), that also uses dynamic programming to efficiently compute segmentation

probabilities, but does not necessarily choose the most probable segmentation. Instead, it bases its likelihood of choosing a segmentation on how likely each segmentation is. So, if a segmentation has probability 0.20, the learner will choose it with probability 0.20 as the correct segmentation of the utterance, even if other segmentations are more probable.

We also examined a learning algorithm that encodes the idea of human memory decay and so focuses processing resources more on recent data than on data heard further in the past (a recency effect). We implemented this using a Decayed Markov Chain Monte Carlo (DMCMC) algorithm (Marthi et al., 2002), which processes an utterance by probabilistically sampling s word boundaries from all the utterances encountered so far. When sampling a potential word boundary, the learner decides whether to insert or remove that boundary, based on the knowledge (as encoded in the current lexicon counts) accumulated so far by the learner. The probability that a particular potential boundary b is sampled is given by the exponentially decaying function b_a^{-d}, where b_a is how many potential boundaries away b is from the end of the current utterance, 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, with the exact probability of sampling based on the decay rate d.

After each boundary is sampled, the learner updates the lexicon. We examined a number of different decay rates, ranging from 2.0 down to 0.125. To give a sense of what these really mean for the DMCMC learner, Table 1 shows the likelihood of sampling a boundary within the current utterance assuming the learner could sample a boundary from any utterances that occurred within the last 30 minutes. Calculations are based on samples from the alice2.cha file from the Bernstein corpus (Bernstein-Ratner, 1984) in the CHILDES database (MacWhinney, 2000), where the average utterance is 3.5 seconds long. As we can see, the lower decay rates cause the learner to look further back in time, and thus require the learner to have a stronger memory.

Table 1. Likelihood of sampling a boundary from the current utterance, based on decay rate.

Decay rate	2	1.5	1	0.75	0.5	0.25	0.125
Likelihood	0.942	0.772	0.323	0.125	0.036	0.009	0.004

4. Bayesian Model Results

We tested the GGJ Ideal learner and our three constrained learners on five randomly generated training sets (~8800 utterances each) and separate test sets (~900 utterances each), where each training and test set were non-overlapping subsets of the data set used by GGJ and each training and test set together formed the complete data set. This data set was the Bernstein corpus (Bernstein-Ratner, 1984) from the CHILDES database (MacWhinney 2000), which contained 9790 child-directed speech utterances (33399 tokens, 1321 types, average utterance length = 3.4 words, average word length = 2.9 phonemes) that

had been phonemically transcribed using characters equivalent to IPA characters that were easy to type (Brent, 1999). See Table 2 below for some examples and GGJ for the full mapping of IPA symbols to characters.

Table 2. Sample phonemic encodings of the Bernstein corpus.

English orthography	Phonemic transcription
You want to see the book	yu want tu si D6 bUk
look there's a boy with his hat and a doggie	lUk D*z 6 b7 wIT hIz h&t &nd 6 dOgi
you want to look at this	yu want tu lUk &t DIs

We assessed the performance of these different learners, based on precision (sometimes called accuracy) and recall (sometimes called completeness) over word tokens, word boundaries, and lexicon items. Precision is defined as the number correct divided by the number found (i.e, "how many words did I correctly identify?" compared to "how many words did I identify total?"). Recall is defined as the number correct divided by the number found in the correct segmentation (i.e., "how many words did I correctly identify?" compared to "how many words should I have identified if I had actually segmented this perfectly?"). To demonstrate how these measures differ when assessed across word tokens, word boundaries, and lexicon items, consider the evaluation of the utterances "*look at the doggie*" and "*look at the kitty*". Suppose the algorithm decided the best segmentation was "*lookat the doggie*" and "*lookat thekitty*" (we again use orthographic forms for ease of clarity). For word tokens, precision is 2/5 (*the* and *doggie* are correct, but *lookat, lookat,* and *thekitty* are not), while recall is 2/8 (*the* and *doggie* are correct, but 8 separate words should have been found). For word boundaries, utterance-initial and utterance-final boundaries are excluded because they are provided in the input; so, precision is 3/3 (all the boundaries identified are true boundaries), while recall is 3/6 (there are three boundaries missing). For lexicon items, precision is 2/4 (*the* and *doggie* are correct, but *lookat* and *thekitty* are not), while recall is 2/5 (*the* and *doggie* are correct, but *look*, *at*, and *kitty* should also have been found).

Table 3 reports the scores for each learner, including F-scores (bolded for easy comparison across models) that combine precision and recall (F-score = (2*precision*recall)/(precision+recall)). All DMCMC learners have s = 20000 (20000 samples per utterance), as we found this gave the best segmentation performance. While this may still seem like a lot of processing, this learner nonetheless takes 89% fewer samples than the ideal learner in GGJ, which is a substantial savings in processing resources. In addition, the DMCMC unigram learners fared best with d = 1.0, while the DMCMC bigram learners fared best with d = 0.25.

Table 3. Average performance of different learners on the test sets.
Precision, recall, and F-scores over word tokens (TP, TR, TF), word boundaries (BP, BR, BF), and lexicon items (LP, LR, LF) resulting from the chosen word segmentation are shown. Standard deviations are italicized.

	TP	TR	TF		BP	BR	BF		LP	LR	LF
Unigram Models (Words are not predictive)											
GGJ – Ideal	63.2 *0.99*	48.4 *0.80*	**54.8** *0.85*		92.8 *0.67*	62.1 *0.42*	**74.4** *0.42*		54.0 *1.89*	73.6 *1.89*	**62.3** *1.30*
DPM	63.7 *2.82*	68.4 *2.68*	**65.9** *2.73*		77.2 *1.86*	85.3 *1.67*	**81.0** *1.64*		61.9 *2.17*	56.9 *2.07*	**59.3** *2.09*
DPS	55.0 *4.82*	62.6 *3.99*	**58.5** *4.45*		70.4 *3.73*	84.2 *1.79*	**76.7** *2.85*		54.8 *1.64*	49.2 *3.14*	**51.8** *2.20*
DMCMC	71.2 *1.57*	64.7 *2.31*	**67.8** *1.97*		88.8 *0.89*	77.2 *2.17*	**82.6** *1.53*		61.0 *1.18*	69.6 *0.43*	**65.0** *0.67*
Bigram Models (Words are predictive)											
GGJ – Ideal	74.5 *1.41*	68.8 *1.53*	**71.5** *1.46*		90.1 *0.75*	80.4 *1.01*	**85.0** *0.82*		65.0 *1.19*	73.5 *1.71*	**69.1** *1.15*
DPM	67.5 *1.13*	71.3 *0.74*	**69.4** *0.90*		80.4 *0.96*	86.8 *0.63*	**83.5** *0.57*		66.0 *1.00*	63.2 *1.46*	**64.6** *1.05*
DPS	34.2 *2.16*	47.6 *2.16*	**39.8** *2.13*		54.9 *1.40*	85.3 *2.07*	**66.8** *1.00*		39.0 *2.02*	34.4 *2.42*	**36.5** *2.19*
DMCMC	72.0 *1.24*	74.0 *1.76*	**73.0** *1.43*		84.1 *0.98*	87.4 *1.47*	**85.7** *0.94*		61.1 *1.41*	64.2 *1.35*	**62.6** *1.17*

A few observations: First, while there is considerable variation in the performance of our constrained learners, nearly all of them out-perform a transitional probability learner on realistic test data (Gambell & Yang (2006) report a precision of 41.6 and recall of 23.3, for an F-score of 29.9 over word tokens; precision and recall scores over word tokens from Brent (1999) appear to be in the lower 40s while precision over lexicon items appears to be around 15). Second, when we examine the impact of the unigram and bigram assumptions on word token performance, we find that the bigram learners do not always benefit from assuming words are predictive of other words. While the Ideal, DPM and DMCMC learners do (bigram F > unigram F, Ideal: $p < .001$, DPM: $p = .046$, DMCMC: $p = .002$), the DPS learner is harmed by this bias (unigram F > bigram F: $p < .001$). This is also true for the lexicon F scores: while the Ideal and DPM learners are helped (bigram F > unigram F, Ideal: $p < .001$, DPM: $p = .002$), the DPS and DMCMC learners are harmed (unigram F > bigram F, DPS: $p < .001$, DMCMC: $p = .006$).

Third, when comparing our ideal learner to our constrained learners, we find – somewhat unexpectedly – that some of our constrained learners are performing equivalently or *better* than their ideal counterparts. For example, when we look at word token F-scores for our bigram learners, the DMCMC learner seems to be performing equivalently to the Ideal learner (DMCMC $\neq$ Ideal: $p = 0.144$). Among the unigram learners, our DPM and DMCMC learners are equally out-performing the Ideal learner (DPM > Ideal: $p < .001$, DMCMC >

Ideal: $p < .001$, DPM ≠ DMCMC: $p = 0.153$) and the DPS is performing equivalently to the Ideal learner (Ideal ≠ DPS: $p = 0.136$). Turning to the lexicon F-scores, the results look a bit more expected for the bigram learners: the Ideal learner is out-performing the constrained learners (Ideal > DPM: $p < .001$, Ideal > DPS: $p < .001$, Ideal > DMCMC: $p < .001$). However, among the unigram learners we again find something unexpected: the DMCMC learner is out-performing the Ideal learner (DMCMC > Ideal: $p = .006$). The Ideal learner is still out-performing the other two constrained learners, however (Ideal > DPM: $p = .031$, Ideal > DPS: $p < .001$).

Fourth, GGJ found that both their ideal learners tended to under-segment (putting multiple words together into one word), though the unigram learner did so more than the bigram learner. One way to gauge whether under-segmentation is occurring is to look at the boundary precision and recall scores. When boundary precision is higher than boundary recall, under-segmentation is occurring; when the reverse is true, the model is over-segmenting (i.e., splitting single words into more than one word, e.g. *the dogs* segmented as *the dog s*). In Table 3, we can see that the Ideal learners are still under-segmenting, with the bigram model doing so less than the unigram model. Looking at our constrained learners, we can see that the unigram DMCMC learner is also under-segmenting. However, every other constrained model is over-segmenting, with the DPS learners being the most blatant over-segmenters; the bigram DMCMC learner appears to be over-segmenting the least.

5. Discussion

Using simulated learners, we have discovered several interesting things. First, our constrained learners were able to extract statistical information from the available data well enough to out-perform other cognitively motivated statistical learners that tracked transitional probability. This underscores how statistical learning can be considerably more successful than is sometimes thought when only transitional probability learners are considered. In addition, our results suggest that even with limitations on memory and processing, a learning strategy that focuses explicitly on identifying words in the input (as all our learners here do) may work better than one that focuses on identifying where word boundaries are (as transitional probability learners do). This ties into the purpose of using a statistical strategy for word segmentation in the first place: To identify a seed pool of words reliable enough for language-dependent strategies to become useful. Specifically, if a child is trying to find units in fluent speech, then it seems intuitive that the child would use a strategy that seeks to identify these units explicitly, rather than one where the units are simply a by-product of the strategy.

Second, we discovered that a bias that was helpful for the ideal learner – to assume words are predictive units (the bigram assumption) – is not always helpful for constrained learners. This suggests that solutions we find for ideal learners may not necessarily transfer to learners that have constraints on their

Lisa Pearl 12/18/09 5:38 PM
Comment: I wanted to tie this into why children are using statistical learning in the first place. It may need a little rewording to make it flow, though.

memory and processing the way that humans do. In this case, we speculate that the reason some of our constrained learners do not benefit from the bigram assumption has to do with the algorithm's ability to search the hypothesis space; when tracking bigrams instead of just individual words, the learner's hypothesis space is far larger. It may be that some constrained learners do not have sufficient processing resources to find the optimal solution (and perhaps to recover from mistakes made early on). However, not all constrained learners suffer from this. There were constrained learners that benefited from the bigram assumption, which suggests less processing power may be required than previously thought to find good word segmentations. In particular, if we examine the DMCMC learner, we can decrease the number of samples per utterance to simulate a decrease in processing power. Table 4 shows the F-scores by word tokens for both the unigram and bigram DMCMC learner with varying samples per utterance. Though performance does degrade when processing power is more limited, these learners still out-perform the best transition probability learner (which had scores in the 40s for word tokens), even when sampling only 0.06% as much as the ideal learner. Moreover, the bigram assumption continues to be helpful, even with very little processing power available for the DMCMC learner.

Table 4. Performance on test set 1 for DMCMC learners with varying samples per utterance. Learners were tested with the decay rate that yielded the best performance at 20000 samples per utterance (unigram = 1, bigram = 0.25). F-scores over word tokens are shown, as well as the processing comparison to the ideal learner (as measured by number of samples taken).

# of samples/utterance	20000	5000	1000	500	100
% Ideal learner samples	*11.3*	*2.84*	*0.57*	*0.28*	*0.06*
Unigram	69.3	65.5	63.4	60.0	51.1
Bigram	74.9	68.3	64.6	61.2	60.9

Turning to the broader comparison of the ideal learner to our constrained learners, we discovered – somewhat surprisingly - that some of our constrained unigram learners out-performed the ideal learner. This may seem counterintuitive, as one might naturally assume that less processing power would lead to worse performance. (Though see Newport (1990) for the "Less is More" hypothesis for language acquisition, which suggests that less processing power may in fact be beneficial for language acquisition.) While we currently have no specific explanation for this behavior for our case study, we do plan to test the robustness of the phenomena by examining our learners' performance on corpora of speech to younger children (e.g., the Brent corpus in the CHILDES database (MacWhinney, 2000) contains English speech to children younger than nine months) and speech to children in languages besides English (e.g., the JacksonThal corpus in the CHILDES database (MacWhinney, 2000) contains Spanish speech to children under a year old).

We also discovered that the tendency to under-segment the corpus depends on how constraints are implemented in our learners, as well as whether the learners assume words are predictive or not. According to Peters (1983), children tend to make errors that indicate under-segmentation rather than over-segmentation, so perhaps learners that under-segment are a better match for children's behavior. Here, the models that under-segmented were both of the ideal learners as well as the unigram DMCMC learner.

6. Conclusions & Future Work

Simple intuitions about human cognition, such as humans having limited memory and processing abilities, can be implemented in numerous ways using cognitively-motivated learning algorithms. Our learners incorporated the ideas that human processing is incremental and human memory shows a recency effect, and we found that the implementation of these ideas non-trivially determined which learning assumptions were helpful and which were not. Still, there are obviously other ways of implementing constrained learning algorithms. We view these investigations as a first step towards understanding how to incorporate human limitations into rational models of human cognition. It is also useful to ask if the effects discovered here are robust, and persist across different corpora and different languages. Moreover, we can take further inspiration from what is known about the representations infants attend to, and allow our algorithms to operate over syllables (Jusczyk et al., 1999a) and track stressed and unstressed syllables separately (Curtin, Mintz, & Christiansen, 2005).

In the larger picture, this study speaks to the problem of translating ideal learner solutions for an acquisition problem to constrained learner approximations of those solutions. This process is not necessarily straight-forward – as we have seen here, learning biases and assumptions that were helpful for the unconstrained learner were not always helpful to the constrained learners. By integrating what we know about human statistical learning abilities with what we know about human limitations, we can hopefully come to understand how infants solve the language acquisition problems that they do when they do.

Appendix. Definition of the Bayesian Model

The probabilistic model of GGJ, used by all modeled learners in this paper, is defined by the equations below. We can imagine that the sequence of words $w_1 \ldots w_n$ in h is generated sequentially using a probabilistic generative process. In the unigram model, the identity of the ith word is chosen according to

$$\text{(A1)}\quad P(w_i = w \mid w_1 \ldots w_{i-1}) = \frac{n_{i-1}(w) + \alpha P_0(w)}{i - 1 + \alpha}$$

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

$$\text{(A2)}\quad P_0(w = x_1 \ldots x_m) = \prod_{j=1}^{m} P(x_j)$$

The equation in (A1) enforces the preference for a small lexicon by giving a higher probability to hypotheses where a small number of words occur frequently compared to those with larger lexicons, where each word occurs less often. The probability of a word is completely determined by the equation in (A2). Since this is the product of the phonemes in the word, words with fewer phonemes (i.e., shorter words) will be preferred. This model is known as a *Dirichlet Process* (Ferguson, 1973).

The bigram model, defined in (A3) and (A4), is conceptually similar to the unigram model except that it tracks not only the frequencies of individual words, but also the frequencies of pairs of words (i.e., bigrams). Just as the unigram model prefers hypotheses where a small number of words appear with high frequency, the bigram model prefers hypotheses where a small number of bigrams appear with high frequency (in addition to the assumptions of the unigram model).

$$\text{(A3)}\quad P(w_i = w \mid w_{i-1} = w', w_1 \ldots w_{i-2}) = \frac{n_{i-1}(w', w) + \beta P_1(w)}{n_{i-1}(w') + \beta}$$

$$\text{(A4)}\quad P_1(w_i = w) = \frac{b_{i-1}(w) + \gamma P_0(w)}{b_{i-1} + \gamma}$$

Here, $n_{i-1}(w', w)$ 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 model parameters. The preference for hypotheses with relatively few distinct bigrams is enforced in the equation in (A3), by making a bigram's probability approximately proportional to the number of times it has occurred before. When a new bigram is created, its probability is determined by the equation in (A4), which assigns higher probability to new bigrams that use words that already occur in many other bigrams (i.e., the model assumes that a few words create bigrams very promiscuously, while most do not). This model is known as a *hierarchical Dirichlet Process* (Teh at al, 2006).

References

Bernstein-Ratner, N. (1984). Patterns of vowel Modification in motherese. *Journal of Child Language. 11,* 557-578.

Brent, M. (1999). An efficient, probabilistically sound algorithm for segmentation and word discovery. Machine Learning, 34:71–105.

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

Curtin, S., Mintz, T., & Christansen, M. (2005). Stress changes the representational landscape: evidence from word segmentation in infants. *Cognition*, *96*, 233-262.

Ferguson, T. (1973). A Bayesian analysis of some nonparametric problems. *Annals of Statistics, 1*, 209-230.

Gambell, T. & Yang, C. (2006). Word Segmentation: Quick but not Dirty. Manuscript. Yale University.

Griffiths, T. L., & Tenenbaum, J. B. (2005). Structure and strength in causal induction. *Cognitive Psychology, 51,* 354-384.

Goldwater, S., Griffiths, T., and Johnson, M. (2007). Distributional Cues to Word Boundaries: Context is Important, In Caunt-Nulton, H., Kulatilake, S., and Woo, I. (eds), *BUCLD 31: Proceedings of the 31st annual Boston University Conference on Language Development,* Somerville, MA*:* Cascadilla Press, 239-250.

Johnson, E. & Jusczyk, P. (2001). Word segmentation by 8-month-olds: When speech cues count more than statistics. *Journal of Memory and Language, 44*, 548-567.

Jusczyk, P., Goodman, M., and Baumann, A. (1999a). Nine-month-olds' attention to sound similarities in syllables, *Journal of Memory & Language, 40*, 62–82.

Jusczyk, P., Hohne, E., and Baumann, A. (1999b) Infants' sensitivity to allophonic cues for word segmentation. *Perception and Psychophysics, 61*, 1465-1476.

Juszcyk, P., Houston, D., and Newsome, M. (1999c). The beginnings of word segmentation in English-learning infants. *Cognitive Psychology, 39,* 159-207.

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.

Mattys, S., Jusczyk, P., Luce, P., and Morgan, J. (1999). Phonotactic and prosodic effects on word segmentation in infants. *Cognitive Psychology, 38*, 465-494.

Morgan, J., Bonamo, K., and Travis, L. (1995). Negative evidence on negative evidence. *Developmental Psychology, 31*, 180-197.

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

Oaksford, M. & Chater, N. (1998). *Rational models of cognition*. Oxford, England: Oxford University Press.

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

Saffran, J., Aslin, R., & Newport, E. (1996). Statistical learning by 8-month-olds. *Science*, *274*, 1926-928.

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

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

Thiessen, E., & Saffran, J. R. (2003). When cues collide: Use of stress and statistical cues to word boundaries by 7- to 9-month-old infants. *Developmental Psychology, 39*, 706–716.

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