Incorporating Cognitive Realism into Models of Early Language Acquisition

Lawrence Phillips

March 13, 2013

1 Introduction

Language acquisition presents a very interesting problem from a human learning standpoint. Few human-created systems are quite as complex, nor as easily and rapidly acquired as one's native language. Over the course of a child's first few years of life, quite a number of linguistic problems must be met and conquered, including (i) learning a sound inventory (*phoneme identification*), (ii) learning word boundaries (*word segmentation*), (iii) word-meaning mapping (*word learning*), and (iv) learning how words combine to produce utterances (*syntax learning*). Before a child can tie his or her own shoes, all of these tasks are achieved with somewhat startling constraints on the learning process. First, children rarely receive explicit instruction for these tasks. Second, they rarely receive direct negative evidence, which would indicate which hypotheses about language are incorrect. Third, they learn so rapidly without the luxury of a fully-developed brain.

These facts pose unique challenges to cognitive scientists attempting to make sense of language acquisition. What prior knowledge about language, if any, do infants bring to the language learning problem? For example, we know that even very young infants are sensitive to probabilistic information (Saffran et al. 1996, Xu & Garcia 2008, Xu & Denison 2009), but to what extent can this information inform language learning? Furthermore, any model of language learning needs to explain why language acquisition proceeds so easily for children while acquiring a language as an adult (*second language acquisition*) proves to be a difficult experience for most adults.

Computational modeling is a tool we can use to shed light on these issues. Computational models require us to make specific commitments about the nature of language acquisition before we can implement the models, and then a given model can serve as an explicit instantiation of a particular learning strategy. If a particular model then fails, this tells us that children must be learning in some manner which is different from the strategy implemented in that model. Although a successful model does not necessitate that children's learning occurs in a similar fashion, it does serve as an existence proof that language learning could proceed that way. Moreover, because language learning is so difficult, a model completely without merit is unlikely to pass (Hoff 2008).

Bayesian inference models have been increasing in popularity within cognitive science and especially within language acquisition research (Xu & Tenenbaum 2007, Griffiths & Kalish 2007, Perfors et al. 2011, Goldwater et al. 2011). One advantage of these models is they allow us to explicitly divide particular aspects of the learning process into specific parts of the model. For instance, any Bayesian model is split between its prior, which encodes the learner's beliefs over a set of possible hypotheses about the language, and the likelihood, which encodes the child's perception of the match between a given hypothesis and the observable data. Further, the particular way in which the inference, which combines the prior and likelihood, occurs can be made more realistic by incorporating cognitive constraints that reflect the limitations human learners have (Phillips & Pearl 2012).

At the same time, a potential downside of computational modeling is that explanatory power of a model is often directly related to how realistic the model is. Therefore, it is very important to both accurately frame the learning problem and also implement learning in a cognitively plausible way. While there are many ways to interpret "cognitive plausibility", I suggest that the following are reasonable aims for a cognitively plausible model: (i) learning should occur incrementally, (ii) the model should incorporate knowledge that learners of the appropriate age have access to, and (iii) the model should incorporate processing constraints which human learners of the appropriate age are likely to possess. What all of these things have in common is that they aim to make a computational model implement more a "realistic" learning process.

2 Realistic learning

Many problems in cognitive science, including language acquisition, can be tackled from multiple directions simultaneously. The problem of realistic computational modeling can be seen as separate problems at two of Marr's levels of explanation (Marr 1983). At the computational level, it is important to describe the problem in a way that realistically represents the problem children are actually solving. Only by making the model realistic can we draw inferences as to which strategies children could possibly pursue. At the algorithmic level, if we take our model to represent in some fashion the process occurring in human learners, then it is clearly important that this learning algorithm be realistic. The question is not whether learning can (or cannot) occur, but whether this learning can occur given the constraints which human learners have.

2.1 Computational-level approaches to language acquisition

Most computational models of language learning exist on the computational level. They make no direct claims that humans learn in such a manner, but rather argue only that a strategy incorporating particular representations or particular learning assumptions can in principle solve the learning problem. Thus, these representations or assumptions may be something incorporated by human learners. These kinds of studies have covered many aspects of language learning, including phonology (Feldman et al. 2009, Dillon et al. 2011), word segmentation (Johnson & Goldwater 2009, Goldwater et al. 2009), word learning (Xu & Tenenbaum 2007, Frank et al. 2009) and syntax (Perfors et al. 2011). This type of work has made many contributions to the field of language acquisition, proposing novel solutions to problems for which no mathematically coherent method of learning previously existed. They all possess, however, a number of implicit assumptions which call into question the claims they make about the success of the implemented learning strategies.

In order to create a mathematical model of language learning, a number of assumptions must be built into the model. One type of assumption which is often taken for granted has to do not with the mathematics of the model, but with the domain in which it operates. For instance, the assumed units of representation play a crucial role in defining the space in which learning occurs. Models of phone identification almost always assume that learning occurs separately for different types of sounds (de Boer & Kuhl 2003, McMurray et al. 2009, Feldman et al. 2009). This has led to the fact that learning models exist for vowels and for stop consonants, but are nonexistent for all other sounds. This problem arises out of the fact that learning has to operate over some continuous acoustic unit. Phoneticians have long proposed relevant acoustic units such as formant frequencies and voice-onset time to describe the acoustic differences between contrasting phones. These units, however, tend to be relevant or defined only for a particular set of sounds. This creates a situation where learning models that utilize phonetically relevant measures are unable to be applied beyond a particular subset of a language's phonetic inventory. This makes it less obvious that the learning strategy proposed can generalize beyond the particular subset of sounds the strategy is implemented for. Also, as these have been computational-level models, it is unclear whether or not sound identification proceeds in such a manner.

In the case of word segmentation, the ability to access a particular unit of representation plays a similarly crucial role in defining the learning problem. Conversational speech proceeds fluidly without punctuation or clear acoustic markers of word boundaries. Even where potential markers do exist, they are highly language-dependent and therefore must be learned for the particular language being segmented. Therefore, it has been proposed that the earliest stages of word segmentation proceed via statistical learning (Thiessen & Saffran 2003). Experimental evidence shows that infants are able to track transitional probabilities (TPs; Saffran et al. 1996) which lends credence to this hypothesis. Interestingly, word segmentation models have typically used phonemes as the unit of representation (Brent 1999, Blanchard & Heinz 2008, Goldwater et al. 2009) - despite the fact that word segmentation begins as early as 6 months (Bortfeld et al. 2005; 7 months Jusczyk et al. 1993a, Jusczyk & Aslin 1995, Echols et al. 1997) while phone identification often does not fully occur until 10-12 months (Werker & Tees 1984). By defining the word segmentation problem in this way, it is unclear that previous models could be extended to the actual task which infants face, which likely first occurs without reference to phonemes.

Another way in which computational-level analyses can be made more realistic is through the incorporation of multiple problems into a single learning model (Johnson & Goldwater 2009, Blanchard et al. 2009, Feldman et al. 2009). This has become a relatively popular approach in language acquisition because infants often learn about multiple levels of language simultaneously. For example, as mentioned above, word segmentation is likely occurring at the same time as phoneme identification. While solving two problems simultaneously might seem like a harder task than solving the problems individually, recent computational-level modeling research has suggested that this may actually be easier than solving the problems separately. These synergies between the statistical information relevant for each problem can then be leveraged for solving the other.

2.2 Algorithmic-level approaches to language acquisition

Ideally, however, our models are not only computational analyses of the utility of particular representations or assumptions in language learning. Instead, models can also show the plausibility of a particular learning strategy, given the cognitive limitations of human learners. Simply put, it is only a first step to show that the learning problem can be solved with statistical learning (for example); it must then be shown that the learning problem can be solved with statistical learning as implemented by a human learner. Increasingly, statistical learning models have been attempting to bridge this gap and show that they are plausible models of learning, given what we know about the cognitive limitations of human learners (Wang & Mintz 2007, Pearl et al. 2011, Lignos 2011).

Just as with computational level approaches, there are many ways in which one might investigate an algorithmic-level analysis. Whereas the computational level deals primarily with the specifics of the model, its prior, likelihood function, and input, the algorithmic level deals with the step-by-step process of how learning occurs for that system. For example, a typical method for learning within Bayesian models is called *batch learning*. This requires that all data be evaluated simultaneously. While this shows that a problem can be solved, it does not show that the problem could plausibly be solved by human learners, as it is unlikely humans are implementing these resource-intensive, iterative ways of selecting the hypothesis with the highest posterior.

A very basic change to the learning algorithm which can be made is to create a model which learns incrementally, as the data come in. It seems clear that infants learn in this way, paying attention to information as it appears and integrating it into their existing hypotheses about language, rather than waiting some pre-appointed amount of time before making decisions. This *online learning* approach therefore seems much closer to the learning process occurring in human learners, and has seen increased use in past years (Wang & Mintz 2007, Pearl et al. 2011, Phillips & Pearl 2012).

Besides the time course of learning, human learners also have constraints on the cognitive resources that they bring to the language acquisition task. While it is clear that children (and especially infants) have fewer cognitive resources than adults, one might reasonably wonder whether these differences impact statistical learning and language acquisition more generally. Longitudinal studies investigating children's abilities make it clear that differences on cognitive measures correlate with current language ability and predict future language growth (Rose et al. 2009). A striking difference between adults and children relates to how they deal with inconsistent input. Adults tend to probability match the inconsistencies they encounter, while children create generalizations about what structures to use (Hudson Kam & Newport 2005, Hudson Kam & Chang 2009). Experimental evidence suggests that this behavior arises from children's difficulties with memory retrieval (Hudson Kam & Chang 2009). When adults are tested in conditions that tax their memory retrieval abilities, they perform quite similarly to children, relying on productive forms rather than probability matching the input.

In order to accurately model language acquisition from the perspective of a young human learner, we would ideally like to incorporate the appropriate cognitive abilities which these learners possess. Of course, this can difficult to determine. The brain, even at a young age, is capable of making complex inferences through processes which currently are unclear. That said, it is possible for models to incorporate constraints that we have reason to believe are more cognitively plausible such as (i) online processing, (ii) sub-optimal decision making (Börschinger & Johnson 2011, Pearl et al. 2011), (iii) a preference for frequent information (Mintz 2003, Wang & Mintz 2007), and (iv) memory constraints (Pearl et al. 2011).

I will now discuss investigations that are underway for two language learning problems children solve when they are very young: word segmentation and phone identification. In each case, computational modeling is used to show the importance of utilizing statistical information to infer linguistic structure. In the case of word segmentation, this work involves translating a computational-level approach to the algorithmic-level. Doing so uncovers surprising behavior that fits a perspective about language acquisition called the "Less is More" hypothesis. For phone identification, this involves reframing the problem at a computational-level in order to unify all phone learning under a single model, the infinite hidden Markov model.

3 Investigation 1: Word segmentation

3.1 Introduction

3.1.1 "Less is More"

"Less is More" (LiM) (Newport 1990) is a somewhat counterintuitive hypothesis that posits that cognitive limitations may help, rather than hinder, children acquiring their native language. It is typically used when explaining why children are so good at learning language while adults often struggle to achieve native-level competency in a non-native language. The original LiM proposal suggests that a trade-off exists between rote memorization and abstract generalization. An example of this can be seen in the form of English past tense verbs, which either take nonproductive memorized forms (e.g., *go~went*, *run~ran*), or instead follow a linguistic rule such as "add *-ed*" (i.e., *laugh~laughed*, *talk~talked*). The intuition is that if children had unlimited cognitive resources, they could memorize everything without the need for generalization; however, since cognitive resources are limited, they are forced to make (ultimately helpful) generalizations. The reflections of this process are seen in children's productions of English past tense forms, which form a U-shaped performance trajectory (Brown 1973, Kuczaj 1977):

(i) good initial performance: Children initially memorize past tense forms, both regular and irregular, and produce them correctly (e.g., production: *go~went, laugh~laughed*).

(ii) poor intermediate performance: Children generalize the "add *-ed*" rule due to their cognitive limitations and increasing verb vocabulary, and end up over-generalizing its use to irregular verbs (e.g., production: *go~goed, laugh~laughed*).

(iii) good final performance: Children eventually learn the balance between generalization and memorization, and only generalize the "add *–ed*" rule to regular verbs (e.g., production: *go~went, laugh~laughed*).

According to this instantiation of the LiM hypothesis, the failure of adult language acquisition stems from having too much memory – adult learners memorize too much, and fail to generalize the way that children do. Experimental evidence supports the idea that children generalize in different ways than adults – for example, children generalize more frequently and easily than adults both in morphological (Hudson Kam & Newport 2005) and syntactic acquisition (Hudson Kam & Chang 2009). Nevertheless, these studies do not pinpoint memory as the unique factor producing these generalization differences. However, computational work by Elman (1993) has suggested that memory constraints can be helpful, as language learning in artificial neural networks often benefits from starting with explicit memory constraints that are gradually relaxed over time. Additionally, Bayesian modeling work by Perfors (2011) has shown that (over)regularization can result from a combination of memory limitations and a bias towards generalization, though it is unlikely to occur from memory limitations alone.

3.1.2 "Less is More" in word segmentation

Though LiM is most often thought of as an explanation for morphological or syntactic acquisition, it is in principle a more general hypothesis about the source of children's exceptional language acquisition abilities. Here I examine potential LiM effects in the process of word segmentation, where children must learn to identify word forms in fluent speech, which is a foundation for much linguistic knowledge. This process is thought to begin around 6 months (Bortfeld et al. 2005) and is certainly in place by around 7.5 months (Jusczyk et al. 1993a, Jusczyk & Aslin 1995, Echols et al. 1997), when infants presumably would have much more limited cognitive resources than adults. One current idea for the learning strategies active at this stage is that infants are leveraging purely distributional information, i.e., statistical cues. This is due in part to the lack of universal cues to word segmentation. In particular, many helpful cues are language-specific, such as phonotactics (Mattys et al. 1999), allophonic variation (Jusczyk et al. 1999a), metrical stress patterns (Morgan et al. 1995, Jusczyk et al. 1999b) and coarticulation effects (Johnson & Jusczyk 2001). Using these cues thus requires infants to already know some words in the language in order to identify the language-specific instantiation of the cue (e.g., stress-initial for the metrical stress pattern cue). Therefore, they are unlikely to be helpful in the initial stage of word segmentation, when infants do not know many words. In contrast, statistical cues can be used initially, before any words are known. The idea that infants are sensitive to statistical cues is bolstered by findings that infants track statistical regularities in speech (Saffran et al. 1996) and in other domains (Xu & Garcia 2008, Xu & Denison 2009).

One very successful purely distributional learning strategy for word segmentation involves Bayesian inference (Goldwater, Griffith & Johnson 2009 (henceforth ***GGJ***), Pearl, Goldwater & Steyvers 2011 (henceforth ***PGS***)). To investigate LiM, I compare ideal Bayesian word segmentation models that have performed quite well (GGJ) to more cognitively plausible models that (i) do not have unlimited processing resources and (ii) incorporate memory restrictions (PGS). Although word segmentation lacks the instance- versus rule-learning trade-off which characterizes traditional LiM phenomena, the study by PGS nonetheless showed a limited LiM effect. The existence of a LiM effect in word segmentation implies that the effect is broader than previously characterized and the exact nature of the effect in word segmentation may have implications for how to understand the causes of LiM more generally. We investigate this effect further here and suggest that limited cognitive resources help push language learners away from naïve assumptions about language. For word segmentation, this leads to more cognitively limited learners discovering the useful units of language, i.e., the word forms that will be assigned meaning in the developing lexicon.

3.1.3 Psychological plausibility in cognitive modeling

I am also interested in investigating learning models that are more faithful to what is currently known about infant learning. While incorporating limitations on processing resources and memory, as PGS did, is quite important in terms of psychological plausibility, I further investigate assumptions about the basic unit of representation for word segmentation. Previous Bayesian modeling studies assumed the basic representational unit for word segmentation was the phoneme (GGJ, PGS). However, experimental evidence from Werker & Tees (1984) suggests that infants are unlikely to recognize and use phonemes until at least 10 months. At the initial stages of word segmentation (i.e., 6 months: Bortfeld et al. 2005), syllables may be a more plausible representational unit, given evidence of categorical perception among syllables (Eimas 1999) as well as infant abilities to use statistical cues defined over syllables (Saffran et al. 1996). By combining a more realistic unit of representation with more psychologically faithful learning models, I find a significant, robust LiM effect in which cognitively constrained learners out-perform their idealized counterparts.

This somewhat surprising result is in line with a broader view of LiM, where limited memory is believed to help learning. Notably however, it is not consistent with the traditional view of LiM coming from morphology that believes the underlying cause of LiM is due to a balance between memorization and generalization. In word segmentation, it is unclear how large a role memorization and generalization could play (though there is still a trade-off between the number and length of word forms in the developing lexicon, as discussed below in section 3.3.1). Given this, I propose a secondary explanation for LiM phenomena: Learners with a naïve model of language are biased by their cognitive limitations into discovering more regular and frequent structures that aid later learning - such as word forms, morphemes, or syntactic units - by updating a very naïve model of the language system. Under this interpretation, the LiM effect arises from a learning bias towards analyzing particularly frequent structures which give the learner a better starting point from which to generalize the language's inherent structure.

3.2 Designing psychologically faithful Bayesian learning models of word segmentation

The goal of cognitive modeling is not just to create an algorithm which solves a task, but rather to create a learning model that helps us understand how humans solve the same task. We can therefore break our modeled learners into two groups: ideal learners and constrained learners. The goal of an ideal learner (such as those in GGJ) is to investigate the utility of a particular learning strategy. For example, GGJ showed that an ideal Bayesian learner with a naïve language model could succeed at word segmentation without relying on any language-specific cues. This style of investigation corresponds to a computational-level analysis (Marr 1982): It defines the problem (often as a specific computation to be done) and asks if this problem is solvable with a learning strategy that incorporates specific learning biases or assumptions, irrespective of the actual algorithm that carries out the computation. The goal of a constrained learner (such as those in PGS) aligns more closely with the algorithmic-level: Is the learning strategy in question useable by humans? To do this, the learner must employ a psychologically faithful and plausible learning algorithm when using that learning strategy to solve the task.

Here I investigate whether Bayesian inference is a good potential learning strategy for the initial stages of word segmentation, focusing on the algorithmic level. Bayesian inference is a strategy that is useful in many domains of language learning, both at the ideal learner level (Foraker et al. 2009, Feldman et al. 2009, Frank, Goodman, & Tenenbaum 2009, Perfors et al. 2011, Dunbar et al. forthcoming) and at the constrained learner level (Regier & Gahl 2004, Xu & Tenenbaum 2007, Pearl & Lidz 2009, Pearl & Mis 2011, Pearl & Mis 2012, Gagliardi et al 2012). Although constrained learners are most useful for investigating the algorithmic level, I implement GGJ's ideal learner as well for comparison. By implementing an ideal learner as well as several constrained learners that use the same underlying learning strategy, I am able to investigate how specific cognitive constraints affect learning with that strategy.

3.2.1 Implementing constraints on cognitive resources

Since I am interested in investigating the effects of learning models that are more faithful to what we know about infant learning, it is important to consider learning algorithms that place constraints on cognitive resources. To this end, I draw on three basic facts about human learning when constructing the constrained Bayesian learners. First, humans – both adults and infants – are likely to analyze information as it comes to them, rather than waiting a predetermined amount of time before analyzing what they have encountered. This leads us to create *online* learners, who process information from data as the data are encountered (as opposed to *batch* learners, who wait a predetermined time before analyzing the data). Second, humans are not completely optimal learners (Tversky & Kahneman 1974, Cascells et al. 1978). This leads to creating learners who may make sub-optimal decisions given the information available (similar to the learners in Borschinger & Johnson 2011). Third, humans have memory limitations, some of which concentrate resources on recent events, often creating a recency effect (Ebbinghaus 1902). This leads to creating a learner that replicates this memory effect (described in more detail in section 3.3.2).

3.2.2 Using syllables as a representational unit

There has been considerable debate regarding the basic unit of representation both within infant learning and in adult speech perception and production. While knowledge of phonemes is generally assumed to be a part of adult-level linguistic knowledge (though see Liberman et al. (1967) and Massaro (1974) for arguments against the phoneme as a basic unit of adult speech perception), I pursue the idea that the syllable (or a syllable-like unit) may be the basic representation for infant speech perception.

The first evidence that infants possess categorical representations of syllabic units appears at 3 months (Eimas 1999), where infants have categorical representations of word-initial syllables (e.g., /be/ in *baby*, and /ba/ in *bottle*). Notably, infants at this age have no categorical representation of phonemes (e.g., they would not recognize the that two syllables /be/ and /ba/ are similar by both beginning with the phoneme /b/). Since word segmentation first occurs around 6 months (Bortfeld et al. 2005), 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) with infants showing evidence of vowel discrimination generally before consonants (Polka & Werker, 1994). This scenario makes it unlikely the learner has full, adult knowledge of the native language phonemes during the initial stages of word segmentation. Although it is possible that word segmentation and phoneme learning bootstrap from one another, I consider the situation where infants only have access to syllabic information.

Further evidence for the syllable as a basic unit of representation comes from acoustic properties of the data. While vowels, which are usually the center of syllables, may fit the acoustic properties necessary for perceptual units – namely, relatively invariant sound patterns in different contexts (although see Raphael, 1972) - many consonants do not (Delattre et al. 1955). In addition, stop consonants in CV syllables (e.g., /da/ or /di/) are unable to be identified by adults before the following vowel has also been identified (Massaro 1974, 1975), which would not be expected if phonemes were the basic unit of perception, given the linear nature of auditory perception. Thus, it may be that syllables are a potential basic perceptual unit from which phonemes are then recovered. For example, backward recognition masking experiments suggest that recognition first occurs over complete CV or VC units and individual phonemes are only recovered afterwards (Massaro 1974, 1975).

While the success of previous Bayesian word segmentation models is heartening for the Bayesian inference learning strategy, how dependent is their success – and the limited LiM effect found - on the assumption of the phoneme as a representational unit? With this question in mind, I modify existing phoneme-based Bayesian models of word segmentation (GGJ, PGS) to operate over syllables. All the modified Bayesian learners treat syllables 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/ that does not exist between /ba/ and /lu/ (Jusczyk & Derrah 1987).

Utilizing syllables as the basic unit of representation has both benefits and drawbacks for word segmentation. On the one hand, it alleviates the learning problem somewhat because it reduces the number of potential word boundary positions. For example, *pretty baby* (/pɹɪɾi bebi/) has four syllables (pɹɪ, ɾi, be, bi) but nine phonemes (p, ɹ, ɪ, ɾ, i, b, e, b, i), yielding three potential word boundary positions for a syllable-based learner but eight potential word boundary positions for a phoneme-based learner. Thus, the syllable-based learner's job is considerably easier, since there are only three decisions to make (i.e., yes or no for a boundary in each position), as compared to eight for the phoneme-based learner.

On the other hand, a potential sparse data problem surfaces for a syllable-based learner. A model operating over English phonemes must track statistics over approximately 40 units; a model operating over a corpus of English syllables must track statistics over approximately 2500 units, while using less data per unit than a phoneme-based model since there are fewer syllable tokens than phoneme tokens in any given corpus. This increases the statistical difficulty of the task tremendously. Additionally, because syllables are treated as atomic units, all phonotactic information about English is lost in the model as there is no representation of phonemes (or phoneme sequences).

Previous work on syllable-based word segmentation strategies (e.g. Yang 2004, Gambell & Yang 2006, Lignos 2011) has demonstrated that heuristic syllable-based learning strategies can perform quite well when segmenting English child-directed speech data. Still, due to the trade-offs discussed, it is unclear a priori whether a syllable-based or phoneme-based Bayesian learner will demonstrate better word segmentation performance.

3.3 Bayesian word segmentation

3.3.1 The Bayesian learning model

Bayesian models are well suited to questions of language acquisition because they explicitly distinguish between the learner's pre-existing beliefs (the *prior*) and how the learner evaluates incoming data (the *likelihood*). This information is combined using Bayes' theorem (1) to generate the updated beliefs of the learner (the *posterior*). Bayesian models take advantage of the distinction between likelihood and prior in order to make a trade-off between model fit to the data and knowledge generalizability (Perfors et al. 2011).

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

The underlying Bayesian models for all of our learners are taken from GGJ. These Bayesian models infer a lexicon of word forms from which the observable data are drawn. These models incorporate a prior over hypotheses which favor "simpler" hypotheses, where simpler translates to two distinct biases: prefer (i) a smaller lexicon and (ii) shorter words in that lexicon. These models are also generative, meaning that they predict how the words and utterances of the observable data are generated. This requires that the learner have some idea of how sentences are generated. Given the limited knowledge of language structure which infants may possess at this age, GGJ posit two simple generative models.

The first model assumes independence between words (a *unigram* assumption) – the learner effectively believes word forms are randomly generated with no relation to each other. To encode this assumption in 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 *i*th word is chosen according to (2):

$$(2) P(w_i = w | w_1 \ldots w_{i-1}) = \frac{n_{i-1}(w) + \alpha P_0(w)}{i - 1 + \alpha}$$

where $n_{i-1}(w)$ is the number of times *w* appears in the previous $i - 1$ words, α is a free parameter of the model which encodes how likely the learner is to encounter a novel word, and P_0 is a *base distribution* (3) specifying the probability that a novel word will consist of particular units (e.g., phonemes or syllables) $x_1 \ldots x_m$. P_0 can be interpreted as a parsimony bias, giving the model a preference for shorter words, since the more units that comprise a word, the smaller the probability of that word is - thus, shorter words are favored. α can be interpreted as controlling the bias for the number of unique lexical items in the corpus, since α controls the probability of creating a new word in the lexicon. For example, when α is small, the learner is less likely to hypothesize new words to explain the observable corpus data, and so prefers fewer unique items in the lexicon.

$$(3) P_0(w = x_1 \ldots x_m) = \prod_{j=1}^{m} P(x_j)$$

The second model makes a slightly more sophisticated assumption about the relationship between words. A learner using this model believes a word is related to the previous word – i.e., a word is generated based on the identity of the word that immediately precedes it. GGJ call this a *bigram* assumption. To encode this assumption, GGJ use a *hierarchical Dirichlet Process* (Teh et al. 2006). This model additionally tracks the frequencies of two-word sequences and is defined as in (4-5):

$$(4) P(w_i = w | w_{i-1} = w', w_1 \ldots w_{i-2}) = \frac{n_{i-1}(w',w)+\beta P_1(w)}{n_{i-1}(w')+\beta}$$

$$(5) P_1(w_i = w) = \frac{b_{i-1}(w)+\gamma P_0(w)}{b_{i-1}+\gamma}$$

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. Both the β and γ parameters, similar to the α parameter described above, control the bias towards fewer unique bigrams (β) and towards fewer unique lexical items (γ).

Both unigram and bigram generative models implicitly incorporate preferences for smaller lexicons by preferring words that appear frequently (due to (2) and (4)) as well as shorter words in the lexicon (due to (3) and (5)). A Bayesian learner using either model must then infer, based on the data, which lexicon items appear in the corpus (word types) as well as how often and where precisely they appear (word tokens in utterances).

3.3.2 Bayesian inference

To use these Bayesian models to make the inferences about the words in the input, GGJ's ideal learner used an algorithm called Gibbs sampling (Geman & Geman 1984), iterating over the entire corpus and sampling every potential word boundary every iteration. Gibbs samplers are guaranteed to converge, which makes these samplers popular for ideal learner problems, since it means that the true posterior of the model can be examined without the effects of additional constraints imposed by the learning algorithm (PGS). Notably, it often takes many iterations to converge on a reliable answer – for example, GGJ used 20,000 iterations for their ideal learners, meaning every potential boundary was sampled 20,000 times. This is clearly an idealization of the learning process as humans are unlikely to remember a large batch of input data with the precise detail required to conduct this kind of iterative learning process. Nonetheless, it addresses the impact of the Bayesian model's assumptions on word segmentation, assuming Bayesian inference can be carried out somehow (most likely by some kind of heuristic approximation, e.g., see Shi et al. 2010). Because this is a batch learner that finds what it considers the optimal segmentation, I will refer to it as the **BatchOpt** learner.

GGJ found that their bigram BatchOpt learner performed better than their unigram BatchOpt learner, meaning the assumption that words are dependent on previous words was a useful one when Bayesian inference can be carried out perfectly and the basic unit of representation is the phoneme. Given this, I examine this distinction in the syllable-based Bayesian learners. While I can implement the same ideal learning algorithm GGJ used to carry out Bayesian inference, I will also consider the constrained learners that PGS investigated, which incorporate processing and memory constraints into the Bayesian inference process.

The Online Optimal (**OnlineOpt**) learner incorporates a basic processing limitation: Linguistic processing occurs online rather than in a batch after a period of data collection. Thus, the OnlineOpt learner processes one utterance at a time, rather than processing the entire corpus at once. This learner uses a dynamic programming algorithm called the Viterbi algorithm (Viterbi 1967) to converge on the optimal (maximal probability) word segmentation for the current utterance, conditioned on the utterances seen so far. In all other aspects, the OnlineOpt learner is essentially identical to the BatchOpt learner: It has perfect memory for previous utterances and unlimited processing resources.

The Online Sub-Optimal (**OnlineSubOpt**) learner is similar to the OnlineOpt learner in processing utterances incrementally and using a dynamic programming algorithm to estimate segmentation probabilities and select a segmentation (specifically, a forward pass of the forward-backward algorithm (Jurafsky & Martin 2000) to compute all possible segmentations, and then a backward pass to sample from the distribution over segmentations). However, it is additionally motivated by the idea that infants, and human beings in general, may not always make optimal choices. For word segmentation, this could mean that infants do not always select the best segmentation. Instead, infants could select segmentations probabilistically, based on how likely each segmentation is – that is, the learners sample potential segmentations. So, learners will often choose the best segmentation, but will occasionally choose less likely alternatives, based on the probabilities of the various segmentation alternatives.

The Online Memory learner (**OnlineMem**) also processes data incrementally, but uses a Decayed Markov Chain Monte Carlo algorithm (Marthi et al. 2002) to implement a kind of working memory. This learner is similar to the original GGJ ideal learner in that it uses Gibbs sampling. However, the OnlineMem learner does not sample all boundaries; instead, it samples some number (s = 20000) of previous boundaries for every utterance processed.[1] The probability of sampling a boundary b is proportional to the decayed function b_a^{-d}, where b_a is the number of potential boundary locations between b and the end of the current utterance ("how many **b**oundaries **a**way from the end") 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. For example, PGS estimate that the probability of sampling a boundary in the current utterance for a phoneme learner is 0.942 for a learner with d=2, while the probability is 0.323 for a learner with d=1. In this experiment, a set, non-optimized value of d=1.5 was utilized to implement a heavy memory constraint. This resulted in a probability of 0.836 for sampling within the current utterance and a probability of 0.954 for sampling within the current or immediately previous utterance. Having sampled a set of boundaries, the OnlineMem learner can then update its beliefs about those boundaries and subsequently update its lexicon. Because of the decay function, the OnlineMem learner's sampling is heavily biased towards boundaries in recently seen utterances and thus the OnlineMem learner implements a kind of recency effect, where recently seen items receive more processing resources than more distant items. This process crudely mimics the human system of working memory. One can think of this effect as though the learner considers every potential boundary in its limited memory, samples from those boundaries, and changes its mind about boundary decisions only while those items remain in memory; it then moves on to the next utterance.

[1] According to PGS, this works out to approximately 89% less processing than the original ideal (BatchOpt) phoneme-based learner in GGJ, which samples every boundary 20,000 times. For the syllable-based learners, this will work out to approximately 74% less processing than the ideal (BatchOpt) syllable-based learner.

Table 1 summarizes the different learning algorithms used for word segmentation by the Bayesian learners.

Learning algorithm	Parameters	Learning assumptions encoded		
		online processing	sub-optimal decisions	recency effect
BatchOpt	(i) iterations i=20,000	-	-	-
OnlineOpt	N/A	+	-	-
OnlineSubOpt	N/A	+	+	-
OnlineMem	(i) samples per utterance s=20,000 (ii) decay rate d=1.5	+	-	+

Table 1. Summary of learning algorithms used for word segmentation.

3.4. Empirical grounding of the input

I test the syllable-based models using English child-directed speech from the Pearl-Brent derived corpus (PGS) from CHILDES (MacWhinney 2000). This modification of the Brent corpus (Brent & Siskind 2001) contains 100 hours of child-directed speech from 16 mother-child pairs. Because I am investigating word segmentation, I restrict the input to child-directed utterances before 9 months of age, leaving 28,391 utterances (average: 3.4 words per utterance, 10.4 phonemes per utterance, 4.2 syllables per utterance). This subset of the Pearl-Brent derived corpus contained a total of 96,723 word tokens of 3,221 individual word types.

While there are many ways to syllabify a corpus automatically, I opted for a two-pronged approach. I used human judgments of syllabification from the MRC Psycholinguistic Database (Wilson 1988) when available. When human judgments were not available (often due to nonsense words like *badido* and *awfuls* or proper names like *Brenda's* or *Cindy*), I automatically syllabified the corpus in a language-independent way using the Maximum-Onset Principle (Selkirk 1981). 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.[2] Approximately 25% of lexical items were syllabified automatically and only 3.6% of our human judgments differ from automatic syllabification[3]. Each unique syllable is then treated as a single, indivisible unit losing all sub-syllabic phonetic (and phonotactic) information.

3.5 Results

I assess the learners in terms of precision (6), recall (7) and F-score (8), where F-score is the harmonic mean of precision and recall (8):

$$(6)\ Precision = \frac{\#\ correct}{\#\ guessed}$$

$$(7) Recall = \frac{\#\ correct}{\#\ actual}$$

$$(8) F - score = \frac{2 * Precision * Recall}{Precision + Recall}$$

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 boundary precision if that boundary is correct. In comparison, the same learner will have just over 0% boundary recall. Similarly, a learner could posit boundaries at every position, producing 100% boundary recall with very low precision because many of the boundaries were false. As the F-score balances these two measures, a high

[2] Of course, since there is a lack of experimental evidence as to the exact nature of infant syllabification, I take this representation as only an approximation.

[3] Differing segmentations consist primarily of examples such as these: *pos/ter* vs. *po/ster*, *sib/ling* vs. *si/bling*, *es/cape* vs. *e/scape*.

F-score indicates the learner is succeeding at both precision and recall. These measurements can be made over individual word tokens (*the penguin eats the fish* = 5 {*the, penguin, eats, the, fish*}), word boundaries (*the penguin eats the fish* = 4 {*the|penguin, penguin|eats, eats|the, the|fish*}), and lexical items (*the penguin eats the fish* = 4 {*the, penguin, eats, fish*}). Additionally, I also consider the log posterior scores for each learner (9-10), which can be interpreted as the fit between the underlying statistical model (i.e. the unigram or bigram language model) and the learner's output.

$$(9)\ \log(Posterior) \propto \log(\text{Prior} * \text{Likelihood})$$

$$(10)\ \log(P(\theta|X) \propto \log\big(P(\theta) * P(X|\theta)\big))$$

In order to prevent overfitting and to ensure that each model is not unfairly judged based on vagaries of the particular data sets chosen as training and test sets, I created five different training and test sets, where the training set consists of 90% of the corpus, which the learner trained on, and the test set consists of the remaining 10%, which the learner was tested on. Each training-test set pair was a random split of the subset of the Pearl-Brent corpus described in section 3.4. All results presented here are averaged over the results of the five input sets, with standard deviations given in parentheses.

To investigate LiM, I compare ideal learners (BatchOpt) against constrained learners (OnlineOpt, OnlineSubOpt, OnlineMem) for both syllable-based and phoneme-based Bayesian learners. To investigate the effect of using the syllable as the basic unit of representation, I compare syllable-based learners against phoneme-based learners. To investigate the utility of Bayesian inference as a learning strategy, I also compare the Bayesian learners against other learners using simpler strategies that could be viewed as reasonable baselines.

3.5.1 Less is More (LiM): Overview

Table 2 shows the word token F-scores for ideal and constrained learners using different assumptions: (1) a unigram or bigram generative language model, and (2) syllables or phonemes as the basic unit of representation.

	Syl-U	**Pho-U**	**Syl-B**	**Pho-B**
BatchOpt	53.1	54.8	77.1	71.5
OnlineOpt	**58.8**	**65.9**	75.1	69.4
OnlineSubOpt	**63.7**	**58.5**	77.8	39.8
OnlineMem	**55.1**	**67.8**	**86.3**	73.0

Table 2. Word token F-scores for syllable-based (Syl) vs. phoneme-based (Pho) models, comparing unigram (U) and bigram (B) learners. Bold scores indicate that the constrained learner significantly out-performs the ideal learner ($p < 0.05$).

For Bayesian learners using a unigram language model, I find a strong LiM effect: All constrained learners (OnlineOpt, OnlineSubOpt, OnlineMem) significantly out-perform the ideal (BatchOpt) learner, irrespective of the unit of representation. In contrast, learners using a bigram language model only show this effect for the OnlineMem constrained learners that are syllable-based. While online learners occasionally have certain benefits over ideal learners, such as faster convergence and the ability to avoid local minima (Liang & Klein 2009), decreased performance for most constrained learners is probably not unexpected since the constrained learner simply has less data (and so less information) to work with than the ideal learner when making its inferences. The surprising effect is when the constrained learners do better. In particular, the OnlineMem syllable-based learner does show the LiM effect, and a fairly strong one at that (OnlineMem: 86.3 vs. BatchOpt: 77.1). Interestingly, the OnlineMem learner is one of the more constrained learners investigated, as it includes two limiting assumptions: (i) learning is incremental and, (ii) a limited working memory exists. So, the fact that this learner shows a LiM effect is encouraging both for the general hypothesis that cognitive limitations have a beneficial impact on learning and also for the idea that cognitive plausibility, in the form of a modeled working memory, is equally useful for understanding children's language learning abilities.

More generally, these results suggest that constrained learning is more likely to be beneficial if words are assumed to be independent of other words (a unigram model), as PGS found for a phoneme-based learner and I have shown is true for a syllable-based learner. Nonetheless, even if a more sophisticated language model is used where words predict the words that follow them (a bigram model), a LiM effect still arises in syllable-based learner (Syl-B), though only for the potentially more plausible OnlineMem learner. This differs from the phoneme-based learners, which did not show a LiM effect when a bigram language model is used (as PGS found). I discuss these results in more detail in the next section.

3.5.1.1 The effect of cognitive limitations

Focusing first on the unigram learners, we can see a strong and robust LiM effect with both syllable and phoneme-based models. Starting with an ideal, BatchOpt, learner (Syl: 53.1, Pho: 54.8), adding a constraint to process data incrementally (OnlineOpt) increases performance somewhat (Syl: 58.8, Pho: 65.9). Adding a sub-optimal decision making strategy (OnlineSubOpt) also increases performance above the BatchOpt baseline (Syl: 63.7, Pho: 58.5). Finally, when I implement a learner with short-term memory (OnlineMem), I again see a boost in performance compared to the BatchOpt ideal learner (Syl: 55.1, Pho: 67.8).

For the phoneme-based learners, an analysis of the error patterns follows what PGS found: Because the unigram model is unable to account for frequently co-occurring words other than by assuming they are part of the same word, phoneme-based learners tend to undersegment frequent bigrams (e.g., *at the* segmented as *atthe*). The constrained learners appear to avoid this pattern early in the corpus because they have no knowledge of which bigrams are frequent. In this way, the constrained learners tend to segment correctly early on, adding true words into the lexicon which can then be leveraged to avoid undersegmentation later in the corpus.

For syllable-based learners, I find a similar pattern where constrained learners outperform the ideal (BatchOpt) learner when using a unigram assumption. However, an explanation based on the misanalysis of frequent co-occurring words does not account for the syllable-based output.

All syllable-based learners make roughly the same number of mistakes on these kind of frequent bigrams. For instance, on the frequent bigram *come here*, every unigram learner makes the same number of mistakes (22.8 mistakes per run). Current analysis is ongoing to determine the source of the increased performance. For example, it is possible that other kinds of errors are being made by the ideal learner but avoided by the constrained learners.

Turning now to the bigram learners, I find that cognitive limitations do not appear to significantly aid a phoneme-based learner (BatchOpt: 71.5 vs. OnlineOpt: 69.4, OnlineSubOpt: 39.8, OnlineMem: 73.0). However, it is notable that combining online learning with a recency effect (OnlineMem) does not appear to hurt learning, and indeed seems to add somewhat to learning, although not significantly. While this is not quite a LiM effect (since performance did not significantly improve when cognitive limitations were added), it may be viewed as trending towards such an effect since cognitive limitations are not harmful to learning, even though they do not specifically aid learning either.

The syllable-based learner has a different pattern of behavior and shows the LiM effect quite strongly. Adding in sub-optimal segmentations to an online learner (OnlineOpt vs. OnlineSubOpt) increases performance (OnlineOpt: 75.1, OnlineSubOpt: 77.8). This implies that segmentations with lower weights, given the model's naïve assumptions about language structure, may actually be useful sometimes. Combining a recency constraint with online learning yields the best performance of all (OnlineMem: 86.3), and is the most striking example of the LiM effect. Frequency analysis suggests that the OnlineMem learner is identifying slightly more frequent words than the BatchOpt learner (mean frequency = 0.00319 (OnlineMem) vs. 0.00250 (BatchOpt)). To further understand why this LiM effect occurs, I examine the log posterior scores for each of the constrained learners (Table 3), which measure how well the segmented output matches the generative model's assumptions. Because log posteriors range between 0 and negative infinity, scores closer to 0 indicate a better fit to the model's underlying language model.

	Token F-score	Log Posterior
BatchOpt	77.1	-552732
OnlineOpt	75.1	-623216
OnlineSubOpt	77.8	-631540
OnlineMem	86.3	-577879

Table 3: Bigram syllable-based log posterior and token F-scores for each learner averaged over five data sets. Higher F-scores indicate better word segmentation and log posteriors closer to zero indicate a better fit with the model's underlying assumptions about how the corpus data were generated.

Table 3 is useful in that it allows us to compare how each individual learner performs not just in relation to the (adult) gold standard of perfect word segmentation, but importantly how it performs according to the underlying naïve language model (unigram or bigram). As one might expect, the BatchOpt learner brings the data closest to its naïve model, which is apparent by it having the smallest log posterior (-552732). When the corpus is processed incrementally, we see a much larger deviation from the underlying model (OnlineOpt: -623216) and segmenting sub-optimally causes an additional slight decrease in the log posterior (OnlineSubOpt: -631540). The OnlineMem learner, however, is further from the underlying language model than the BatchOpt model (OnlineMem: -577879), but closer than either of the other constrained learners. Nonetheless, it is the OnlineMem learner that shows the LiM effect. This suggests two things. First, segmenting the corpus to match an underlying unigram or bigram model does not necessarily result in increased segmentation performance as compared to the gold standard (BatchOpt vs OnlineSubOpt and OnlineMem). This is not surprising in that we know that language is generated by a process much more complex than a simple n-gram model. Thus, there may be great utility for infant learners in possessing cognitive limitations which keep those learners from segmenting the speech they hear in accordance with a naïve language model. Second, it is more important to be pushed in the right direction than simply to be pushed away from the naïve language model. This is apparent from the OnlineMem learner – while it was not pushed as far from the underlying bigram model as the other constrained learners, it nonetheless seems to be pushed in a better direction since its overall segmentation performance is higher. Both these effects may play a large role in the LiM phenomenon. In particular, if infants begin with naïve assumptions about the language they hear, they must be pushed towards the correct underlying model somehow. These results show that learners using certain cognitively-inspired learning algorithms can not only be pushed away from a naïve language model, but can also be pushed in the right direction. Importantly for the idea that cognitive realism is helpful in computational modeling, the closer our model mimics actual cognitive processes, the better the model performs.

3.5.1.2 The benefit of cognitive limitations

In order to help explain the LiM behavior of our constrained learners, and in particular the OnlineMem learner, I examine the types of words which each learner identifies. One possible explanation for the increased performance of the OnlineMem learner, especially in the bigram case, is that its limited working memory focuses its attention on more frequently occurring units. If this is the case, then these frequent items may account for some of the learner's increased performance, when compared to the ideal (BatchOpt) learner. I determined this to be the case both qualitative and quantitatively. I find (see Table 4) that our BatchOpt learners have higher lexicon recall scores than their OnlineMem equivalents. In contrast, the BatchOpt learners have lower token recall scores. This pattern of results indicates that although the OnlineMem learner is picking out fewer correct word types, these words must be more frequent in order for the OnlineMem learner to have a higher token recall score. Quantitatively, we can measure the frequency of each word within our corpus and determine the average frequency for the correctly identified words from each learner. A 2-tailed, paired t-test shows that the OnlineMem learner does identify true words that are on average more frequent than the BatchOpt learner ($p < .0001$) (BatchOpt: -5.99, OnlineMem: -5.74). This supports the hypothesis that one useful aspect of having a learner with working memory limitations is that it forces the learner to focus on more frequent - and hence more useful - items.

	BatchOpt(U)	OnlineMem(U)	BatchOpt(B)	OnlineMem(B)
Token Recall	45.0	**48.1**	72.5	**85.4**
Lexicon Recall	**73.4**	68.9	**79.7**	76.8

Table 4: Unigram and bigram token recall and lexicon recall results for syllable-based learners. Because fewer types were correctly identified by the OnlineMem learners, yet more tokens were correctly identified, this indicates that the OnlineMem learner identifies more frequent words. Bold values indicate which learner had higher token or lexicon recall.

3.5.2 Syllables vs. phonemes

Clearly our syllable-based learners perform well, but are syllables a better unit of representation than phonemes for word segmentation? Looking again to Table 2, we see that in the unigram case, phoneme-based learners (Pho-U) outperform their syllable-based counterparts (Syl-U), except in the case of the OnlineSubOpt learner. In contrast, in the bigram case, all syllable-based models (Syl-B) outperform their phoneme-based equivalents (Pho-B). This suggests that the bigram assumption is consistently helpful to a syllable-based learner. It may be 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 transitional probabilities (TPs) between syllables are the only source of boundary information. Because there are roughly 2500 unique syllables, there will often be cases where a sparse data problem 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 syllable TP data.

3.5.3 The utility of Bayesian inference as a learning strategy

Table 5 shows the F-scores for word tokens over all the syllable-based and phoneme-based learners, including two additional strategies that can be used as a baseline, the TP-minima learner and the PerceptUnit=Word learner. The first baseline is the Transitional Probability (TP) model, based on Gambell & Yang's (2006) investigation and grounded in empirical infant studies by Saffran and colleagues (Saffran et al. 1996, Aslin et al. 1998). This strategy calculates TPs over perceptual units (like syllables or phonemes) and places boundaries at all local minima. This strategy leverages the observation that TPs tend to be lower between words than within words. Our second baseline is a learner that assumes each basic perceptual unit (e.g., each syllable or each phoneme) is a word (PerceptUnit=Word), a strategy investigated by Lignos (2011). When the perceptual unit is a syllable, this is a strategy that can be very useful in languages containing many monosyllabic words, like English (e.g., the Pearl-Brent corpus averages 1.22 syllables per word).

Notably, all the Bayesian learners out-perform the TP-minima baseline strategy (Syl = 44.0, Pho = 37.4), irrespective of perceptual unit, demonstrating the utility of the Bayesian learning strategy over this more simplistic statistical learning strategy. Clearly, the way in which a statistic, such as TP, is used makes a large difference in terms of outcome, since both the TP-minima learner and the Bayesian learners rely up on TP, but yield markedly different results. Turning to the PerceptUnit=Word strategy, this strategy is clearly a terrible one for a phoneme-based learner, since words are typically comprised of more than one phoneme. And indeed, all phoneme-based learners achieve a better score than the PerceptUnit=Word learner (Pho = 2.2). However, we note again that this strategy is useful for the syllable-based learner because English child-directed speech tends to contain many monosyllabic words (e.g., the Pearl-Brent corpus averages 1.22 syllables per word). Though this may not be a useful strategy cross-linguistically for languages that typically have more syllables per word (e.g., German: 1.60, Japanese: 1.74, Spanish: 1.75, Hungarian: 1.97), it is quite effective for English, achieving an F-score of 72.4. Nonetheless, all the Bayesian syllable-based bigram learners out-perform PerceptUnit=Word, though the unigram learners do not. This provides additional support that the bigram assumption is helpful for syllable-based learners.

	Syl-U	Syl-B	Pho-U	Pho-B
BatchOpt	53.1	77.1	54.8	71.5
OnlineOpt	58.8	75.1	65.9	69.4
OnlineSubOpt	63.7	77.8	58.5	39.8
OnlineMem	55.1	86.3	67.8	73.0
TP-minima	44.0		37.4	
PerceptUnit=Word	72.4		2.2	

Table 5. Word token F-scores across all learning models, including syllable-based (Syl) vs. phoneme-based (Pho) models, unigram (U) vs. bigram (B) models, and the baseline models.

3.5.5 Summary of results

There are four main results of this study. First, I have shown a LiM effect in the task of word segmentation, particularly for Bayesian learners that (i) use a more sophisticated – though still naïve - model of language (a bigram assumption) and (ii) perceive syllables as the basic unit. The fact that this effect was found for bigram learners in particular is new, as previous results suggested a LiM effect arises only for learners using a less sophisticated language model (a unigram assumption). Second, I have demonstrated that syllables are a useful unit for word segmentation. Not only are syllables more psychologically faithful, given what we know about infant speech perception, but learners using them do two useful things: (i) these learners generate more robust LiM effects and (ii) these learners provide support for the utility of the bigram assumption during word segmentation. Third, I find that Bayesian inference is both a useful and useable learning strategy for word segmentation, even if the units of perception are syllables, rather than phonemes. This provides additional support for the viability of Bayesian inference as a learning strategy infants could use.

3.6. Discussion

My results support two broad findings. First, I find that memory-constrained learners outperform their ideal equivalents, which I take as support for the "Less is More" (LiM) hypothesis (Newport 1990). In particular, limited cognitive resources, rather than hurting learner, seem to aid word segmentation. Second, the impact of modeling assumptions is clear, as the LiM effect was obscured in phoneme-based learners but appeared more robustly in syllable-based learners.

This serves to demonstrate that making more cognitively plausible assumptions in computational models of language acquisition may yield answers that do not come from more idealized learning models.

But what exactly is causing the LiM effect here, particularly in the syllable-based bigram learner? PGS found an LiM effect for their phoneme-based learners, but only for learners using a unigram assumption – not for learners using a bigram assumption. One idea is that this 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, online models can actually outperform their batch equivalents. However, this explanation fails to account for my results in two ways: (a) the most direct online equivalent of the ideal syllable-based learner (the OnlineOpt learner) actually performs worse than the ideal syllable-based learner (the BatchOpt learner), and (b) this does not explain the performance boost caused by sub-optimal decision-making (the OnlineSubOpt learner).

Another idea is that the answer lies in the kinds of words these models identify. I find (see Table 4) that the ideal bigram learner correctly segments 72.5% of the words in the input, building a lexicon that contains 79.7% of the actual word-types it encounters. Yet I find that a learner with memory constraints (the OnlineMem learner) can successfully segment 85.4% 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 can be thought of as helping to learn the "important" things. One can think of this memory-constrained learner as one which can retrieve recent knowledge from its memory buffer for later analysis. These processes of retrieval and working memory have been shown to play a crucial role both in language development (Rose et al. 2009) as well as in the way adults regularize input (Hudson

Kam & Chang 2009). Although this has been hypothesized by the literature on LiM in artificial language learning (Kersten & Earles 2001, Cochran et al. 1999), I am unaware of computational support for why constrained processing helps in realistic language acquisition without the presence of an additional bias towards correct generalization (e.g., see Perfors 2011, in press). The fact that this study can help to explain the factors underlying the LiM effect highlights a very major contribution computational modeling can make to developmental research more generally.

For the claim regarding the impact of the unit of representation, I can compare the syllable-based learner results with those of phoneme-based learners. I found a number of crucial distinctions (see Table 2). 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 the 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, I find that a learner using a bigram assumption can have very divergent behavior, depending on the unit of representation. There is a major difference in the performance of the sub-optimal (OnlineSubOpt) learner: The syllable-based OnlineSubOpt learner has comparable performance to the ideal BatchOpt learner (OnlineSubOpt=77.8, BatchOpt=77.1) while its phoneme-based equivalent suffers greatly in comparison with the ideal learner (OnlineSubOpt=39.8, BatchOpt=71.5). I speculate that this is due to the number of potential segmentations the phoneme-based learner considers, compared to the syllable-based learner. In particular, since the OnlineSubOpt 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, I also notice a strong LiM effect in the syllable-based learner that is not present in its phoneme-based counterpart (Syl: BatchOpt=77.1, OnlineMem=86.3; Pho: BatchOpt=71.5, OnlineMem=73.0). More generally, by making more realistic assumptions about the learner's unit of representation, I can 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 I could improve the existing syllable-based Bayesian learners. First, some segmental cue information is likely available to infants, such as phonotactic or articulatory cues. Similarly, suprasegmental cues such as primary stress are known to affect infant word segmentation (Jusczyk et al. 1999b, Thiessen & Saffran 2007) and there is evidence that stressed and unstressed syllables are represented separately in infant memory (Pelucchi, Hay, & Saffran 2009). Finally, the exact form which infants use to represent syllables is unclear. While syllabification must occur, it is unclear precisely how it occurs. When one looks cross-linguistically, languages treat syllabification in very different ways. For instance, it is well documented that in Spanish, syllabification occurs without respect to word boundaries, with any particular syllable potentially containing phonemes from multiple words (Harris 1982). German, on the other hand, tends to avoid these post-lexical resyllabifications (Hall 1992). 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 this pattern of results is truly representative of word segmentation *generally* and not just in English, syllable-based word segmentation models must be tested on data from multiple languages. I am currently investigating these ideal and constrained models of word segmentation in a variety of languages with data available in the CHILDES database (MacWhinney 2000), including German, Spanish, Italian, Japanese, Hungarian, and Farsi.

3.7. Conclusion

This study highlights the benefits of using empirical research from psychology to inform decisions about modeling language acquisition. By combining cognitive limitations with a more realistic unit of representation, I find more robust support for the somewhat counterintuitive “Less is More” (LiM) hypothesis that states cognitive limitations can aid, rather than harm, learning. This demonstrates the utility of adding cognitive plausibility to idealized models of language acquisition. More broadly, this type of research can aid the discovery of language learning strategies that are both useful (which a computational-level modeling approach can identify) and useable (which an algorithmic-level modeling approach can identify). By looking at both types of learning models, I find additional support for Bayesian inference as a learning strategy children may use during language acquisition. Additionally, I have shown computational support for the existence of LiM phenomena outside the traditional realms of morphology and syntax. Because of the computational model, I can not only generate LiM effects, but investigate *why* these effects occur. In this case, it may be that cognitive limitations push learners towards more frequent structures from which later linguistic knowledge may be bootstrapped. More generally, this style of computational work allows us to not only identify the strategies that are likely to be used by children, but also to discover potential explanations for existing, sometimes puzzling, observations about child language acquisition, as with the "Less is More" hypothesis.

4 Investigation 2: Phone learning

4.1 Introduction

At the same time that infants are beginning to segment words out of the fluent speech stream, they are also faced with the task of splitting these words into their constituent units, syllables and phonemes. While syllables may be a natural perceptual unit for infants (Massaro 1974, 1975), phonemes are a representation that must be constructed from the acoustic data. To make things trickier, any given phoneme can have multiple phonetic realizations (for example, the phoneme "t" in *top* [t] is pronounced differently than the "t" in *stop* [t^h] – the latter has a small puff of air accompanying it, called *aspiration*). These phonetic outputs are called phones and represent the basic inventory of sounds in a language (Chomsky & Halle 1968). The relationship between phonemes and phones is further complicated because any given phone may correspond to multiple phonemes, depending on its context (for example, ….[something with the flap sound? *water* /t/ vs. *muddy* /d/]). Still, infants begin learning this phonetic inventory around 6 months (Polka & Werker 1994), have correctly identified most sounds – both phones and phonemes – in their native language by 12 months (Werker & Tees 1984).

As the identification of phones is often assumed to be a precursor to identifying the mapping between phones and phonemes (though see Dillon et al. 2011), we can reasonably wonder how the inventory of phones is first discovered. One popular explanation in recent years has been that infants use distributional learning. Much research has begun to show that infants are aware of the distributional information around them and that they are capable of using that information to make decisions (Saffran et al. 1996, Xu & Garcia 2008, Xu & Denison 2009). Given that infants might have access to this kind of information, they might naturally rely on it when trying to solve the task of phone and phoneme identification. Recent research has investigated Gaussian Mixture Models (GMMs) as a representation of the learning process (Feldman et al. 2009, Vallabha et al. 2007, Dillon et al. 2011). This type of model assumes that there are a set number of phonetic items to be learned, and learning consists of discovering the parameters of a multi-dimensional Gaussian which represents the acoustic realization of that phone. This type of strategy is known to work well for categories with only slight overlap, but has had mixed success with vowels, where categories often overlap heavily (Feldman et al. 2009, Dillon et al. 2011). To combat the failure of the GMM for vowels, Feldman et al. (2009) create a GMM which incorporates information at the word level, simultaneously and successfully learning a lexicon of invariant word forms as well as the phonetic categories that comprise these word forms.

However, an issue with this approach concerns how to represent acoustic data used as input. Typically, measures used by phoneticians such as voice-onset time and formant frequencies are utilized, based on the availability of software to determine such values from the acoustic data and these properties' apparent linguistic relevance. It remains an open question as to how infants know (or learn) to pay attention to such acoustic properties, among all the properties available. Further, by focusing on these kind of measures, there is no way to model the learning of all phonetic categories. For example, vowels are typified by their formant frequencies, yet most consonants have no formants which can be measured or which are relevant to the identity of the consonant. It is also unclear how consonants such as nasals, glides, liquids or fricatives could be quantified through similar acoustic measures.

In order to address this problem, we need some acoustic measure which applies to all phones but which is also more compact (and abstract) than the raw acoustic signal. One measure which has gained wide acceptance within the machine learning community for these kinds of acoustic identification problems is based on the mel-frequency cepstrum (Imai 1983). This cepstrum is the discrete cosine transform of the log Fourier transform of an acoustic signal. Additionally, all of this is calculated on the mel-frequency scale, which is based on the subjective perceptual abilities of humans which is a log function of objective pitch (Stevens et al. 1937). The cepstrum can then be quantified through the set of mel-frequency cepstral coefficients (MFCCs) which collectively define the cepstrum. Some subset of these MFCCs are then used as the basis for learning, having been applied to speech recognition (Viikki & Laurila 1998), speaker identification (Reynolds 1994), and information retrieval concerning music (Logan 2000). MFCCs are based on a perceptual scale which mirrors human abilities, but it is unknown what specific transforms are calculated within the auditory cortex. Nevertheless, one can use MFCCs as a potential stand-in, with the idea that similar methods might also reasonably be employed by the infant brain.

An additional concern with many current models is that they have yet to be tested on realistic data. Because formants and voice-onset time are difficult to measure in fluent speech, most models are tested on acoustic data from experiments (Feldman et al. 2009) or on generated data from Gaussians derived from child-directed speech (Vallabha et al. 2007). These experiments typically include a small number of speakers, potentially only of one gender, producing a small set of words in isolation. This raises the question not only of whether the acoustic patterns of these measures differ in the real world, but also of whether these statistical learning methods are capable of handling the noise which comes from learning in the real world.

4.2 Modeling language learning with infinite Hidden Markov Models

Once a measure is defined to summarize the acoustic input, we can then ask how to model the linguistic system and how to use that model to learn from the input. Because language is in one sense a signal broadcast over time, it often makes sense to model it as a simple hidden Markov model (HMM; Rabiner 1989). HMMs build off of the basic assumption of a Markov chain. The state of any Markov chain is necessarily dependent only on the immediately preceding state. The Markov chain is built off a hidden state sequence, $Z = (z_1, z_2, \ldots, z_N)$. Each value of Z_i corresponds to a particular hidden state $\{1 \ldots K\}$ where K is some finite integer. Every hidden state is associated with an observed variable in the sequence, $Y = (y_1, y_2, \ldots, y_N)$. The HMM is parameterized through a transition matrix which captures the dependency between any two hidden states Z_i and Z_j, where $T_{ij} = p(Z_n = i \mid Z_{n-1} = j)$.The initial state probability is parameterized as $\pi_i = p(Z_1 = i)$. Additionally, for every hidden state Z_i there is some emission probability, parameterized by $\varphi_{Zt} = p(Y_t \mid Z_t)$. We can therefore write the joint distribution over hidden states Z and observed variables Y as:

(11) $p(Z, Y|\pi, T, \varphi, K) = \prod_{i=1}^{N} p(Z_t|Z_{t-1}) p(Y_t|Z_t)$

In order to do Bayesian inference over this model, we must first describe the model priors. The observation paramaters φ are drawn from the arbitrary prior distribution H. Typically, priors for T and π are set as symmetric Dirichlet distributions, given that we have no additional information about what form these parameters might take.

One issue with this type of model is that it assumes that the number of possible hidden states is known *a priori*. This cannot be true for phone learning as languages vary widely in the number of phones that they possess and the size of the phonetic inventory is something infants must learn on their own. Therefore, it is more appropriate to let the model determine an appropriate number of states in an unsupervised fashion. The infinite HMM (iHMM) represents a nonparametric version of the standard HMM where the value of K allows for a countably infinite number of hidden state values. Such a system could be implemented in many different ways. One attempt might be to take K →∞. Such an attempt fails in the context of an HMM because there is no coupling between the transitional probabilities between different states – this is due to the independent priors they are given (Beal et al. 2002). This problem can be solved through a hierarchical Dirichlet process (Teh et al., 2006). We can introduce a coupling between different hidden states by assigning Dirichlet priors with shared parameters:

(12) $T_k \sim Dirichlet(\alpha\beta)$

(13) $\beta \sim Dirichlet(\frac{\gamma}{K} \dots \frac{\gamma}{K})$

As K →∞ this approach begins to approximate the hierarchical Dirichlet process (HDP). The true HDP is a set of Dirichlet processes (DPs) coupled through a base measure which is shared by all DPs and which is itself drawn out of a DP (Teh et al. 2006). That is, each DP is distributed as $G_k \sim DP(\alpha, G_0)$ where G_0 is a shared base measure. One can understand G_0 as the mean of G_k while $\alpha > 0$ controls the variability around G_0 (sometimes called its *concentration*), with smaller values of α leading to larger variability. G_0 is also drawn from a DP, $G_0 \sim DP(\gamma, H)$ where H is again a global base measure for the entire system.

Through the stick-breaking construction for HDPs (Teh et al. 2006), we can identify that G_k describes the transition probabilities between states k and k' as well as the emission probabilities, φ_k'. This allows us to define the iHMM as such:

(14) $\beta \sim GEM(\gamma), \quad T_k|\beta \sim DP(\alpha, \beta), \quad \varphi_k \sim H$

(15) $Z_n|Z_{n-1} \sim Multinomial(T_{Z_{t-1}}), \quad Y_n|Z_n \sim F(\varphi_{Z_n})$

Here GEM(γ) is the stick-breaking construction for DPs (Sethuraman, 1994). Figure 1 captures the graphical model structure of this hierarchical formalism. We can recapture the original HMM structure simply by setting β = (1/K … 1/K, 0, 0 …) where β is non-zero for only the first K entries. Given that we are uncertain about the particular values of α and γ, we place gamma hyperpriors on these variables such that $\alpha \sim Gamma(a_\alpha, b_\alpha)$ and $\gamma \sim Gamma(a_\gamma, b_\gamma)$.

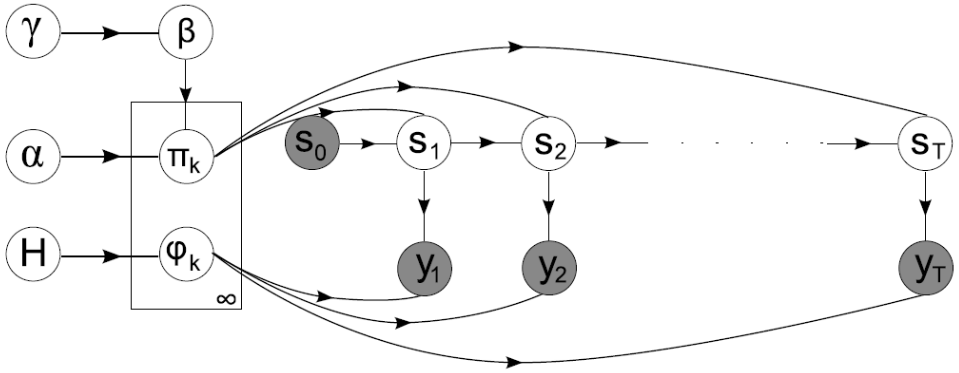

Figure 1. Graphical model implementation of iHMM

4.3 Planned work and discussion

The first question is whether this model is capable of identifying the correct phones (and their acoustic distributions) from acoustic data. If it can, this provides a computational-level existence proof of the viability of such a statistical learning strategy. However, this type of model clearly possesses many abilities which infants lack. To begin with, the data are analyzed all at once (batch learning), rather than in an incremental fashion (online learning). Also, this approach uses an idealized learning algorithm for making the inference about the number and distribution of phones in the language, and so does not include the kind of cognitive constraints infants possess. It may therefore be useful to investigate constrained learning algorithms that implement this statistical learning strategy if it does indeed succeed with an unconstrained learning algorithm.

In addition, there may well be an interaction between phone learning, word segmentation, and lexical development as suggested in Feldman et al. (2009). Experimental evidence suggests that these learning tasks are solved somewhat simultaneously, and it may be that information from one task may usefully inform the other tasks.

5. Conclusion

Any model of language acquisition, whether computational-level or algorithmic-level, is only as worthwhile as the assumptions that it makes about the learning problems children face. I investigate two learning problems infants solve in their first year of life, word segmentation and phone learning, incorporating more realistic assumptions into computational models of these processes. This affects both the framing of the general learning problem as well as the implementation of the learning process itself. The field of language acquisition modeling has seen a growing shift towards this viewpoint, but it is still gaining momentum.

A number of previous studies have shown that modeling more realistic learning is not only possible, but often changes qualitative trends which had previously been seen in other more idealized models; for example, studies of phone identification (Feldman et al. 2009), word segmentation (Pearl et al. 2011, Phillips & Pearl 2012), word learning (Frank et al. 2007), and syntax (Pearl & Sprouse 2012). These results should be taken seriously because these models inform how we understand infant learning generally. More broadly, this can impact the debate about the role of statistical learning in early language development, and the trade-off with innate knowledge of the hypothesis space for solving the language acquisition problem.

6. References

Beal, M.J., Ghahramani, Z., and Rasmussen, C.E. 2002. The infinite hidden markov model. *Advances in Neural Information Processing Systems*, 14:577 -584.

Blanchard, D., Heinz, J., & Golinkoff, R. 2010. Modeling the contribution of phonotactic cues to the problem of word segmentation. *Journal of child language, 37*(3), 487.

Börschinger, B. & Johnson, M. 2011. A particle filter algorithm for Bayesian word segmentation. *Proceedings of Australasian Language Technology Association Workshop*, 10-18.

Bortfeld, H., Morgan, J.L., Golinkoff, R.M. & Rathbun, K. 2005. Mommy and me. *Psychological Science, 16*(4), 298-304.

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

Cascells, W., Schoenberger, A., & Grayboys, T. 1978. Interpretation by physicians of clinical laboratory results. *New England Journal of Medicine, 299*(18), 999-1001.

Chomsky, N. & Halle, M. 1968. The sound pattern of English.

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.

de Boer, B., & Kuhl, P.K. 2003. Investigating the role of infant-directed speech with a computer model. *Acoustics Research Letters Online, 4*(4), 129-134.

Delattre, P.C., Liberman, A.M. & Cooper, F.S. 1955. Acoustic loci and transitional cues for consonants. *Journal of the Acoustical Society of America, 27*, 769-773.

Dillon, B., Dunbar, E., & Idsardi, W. 2011. A single stage approach to learning phonological categories: Insights from Inuktitut. *Manuscript, University of Massachusetts, Amherst and University of Maryland, College Park*.

Ebbinghaus, E. Grundzüge der psychologie. Leipzig: von Veit, 1902.

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.

Elman, J.L. 1993. Learning and development in neural networks: The importance of starting small. *Cognition*, 48, 71-99.

Feldman, N.H., Griffiths, T.L., & Morgan, J.L. 2009. Learning phonetic categories by learning a lexicon. In *Proceedings of the 31st annual conference of the cognitive science society*.

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

Frank, M.C., Goodman, N.D., & Tenenbaum, J.B. 2007. A Bayesian framework for cross-situational word learning. *Advances in neural information processing systems, 20*, 20-29.

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.

Griffiths, T.L. & Kalish, M.L. 2007. Language evolution by iterated learning with Bayesian agents. *Cognitive Science, 31*(3), 441-480.

Hall, T.A. 1992. *Syllable structure and syllable-related processes in German* (Vol. 276). De Gruyter.

Harris, J.W. 1982. *Syllable structure and stress in Spanish: a nonlinear analysis*.

Hay, J.F., Pelucchi, B., Graf Estes, K., & Saffran, J.R. (2011). Linking sounds to meaning: Infant statistical learning in a natural language. Cognitive Psychology, 63, 93-106.

Hoff, E. 2008. Language Development. Belmont, CA: Wadsworth.

Hudson Kam, C.L., & Newport, E.L. 2005. Regularizing unpredictable variation: The roles of adult and child learners in language formation and change. *Language Learning and Development, 1*(2), 151-195.

Hudson Kam, C.L., & Chang, A. 2009. Investigating the cause of language regularization in adults:

Memory constraints or learning effects? *Journal of Experimental Psychology: Learning, Memory and Cognition, 35*(3), 815-821.

Imai, S. 1983. Cepstral analysis synthesis on the mel frequency scale. In *Acoustics, Speech, and Signal Processing, IEEE International Conference on, 8,* 93-96.

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.

Johnson, M. & Goldwater, S. 2009. Improving nonparametric Bayesian inference: experiments on unsupervised word segmentation with adaptor grammars. *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 317-325.

Jurafsky, D. & Martin, J.H. 2000. *Speech & Language Processing*. Pearson Education India.

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 pattern 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., Hohne, E., & Baumann, A. 1999a. Infants' sensitivity to allophonic cues for word segmentation. *Perception and Psychophysics,* 61, 1465-1476.

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

Katz, S.M. 1996. Distribution of context words and phrases in text and language modeling. *Natural Language Engineering, 2*(1):15-59.

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.

Liberman, A.M., Cooper, F.S., Shankweiler, D.P. and Studdert-Kennedy, M. 1967. Perception of the Speech Code. *Psychological Review*, 74(6), 431-461.

Lignos, C. 2011. Modeling infant word segmentation. *Proceedings of the Fifteenth Conference on Computational Natural Language Learning,* 29-38.

Logan, B. 2000. Mel frequency cepstral coefficients for music modeling. In *International Symposium on Music Information Retrieval, 28*(5).

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

Marr, D. 1983. Vision: A computational investigation into the human representation and processing of visual information, Henry Holt and Co. *Inc., New York, N.Y.*

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

Massaro, D.W. 1974. Perceptual Units in Speech Recognition. *Journal of Experimental Psychology*, 102(2), 349-353.

Massaro, D.W. 1975. Understanding language: An information processing analysis of speech perception, reading and psycholinguistics. New York: Academic Press.

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

McMurray, B., Aslin, R.N. & Toscano, J.C. 2009. Statistical learning of phonetic categories: insights from a computational approach. *Developmental science, 12*(3), 369-378.

Mintz, T.H. 2003. Frequent frames as a cue for grammatical categories in child directed speech. *Cognition, 90*(1), 91-117.

Morgan, J., Bonamo, K., & 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.

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.

Pearl, L. & Sprouse, J. 2012. Computational models of acquisition for islands. *Experimental syntax and island effects*.

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

Perfors, A. 2011. Memory limitations alone do not lead to over-regularization: An experimental and computational investigation. In L. Carlson, C. Hoelscher & T.F. Shipley (eds.), *Proceedings of the 33rd Annual Conference of the Cognitive Science Society*. Austin, TX: Cognitive Science Society: 3274-3279.

Perfors, A., Tenenbaum, J.B., Griffiths, T.L., & Xu, F. 2011. A tutorial introduction to Bayesian models of cognitive development. *Cognition, 120*(3), 302-321.

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

Phillips, L. & Pearl, L. 2012. "Less is more" in Bayesian word segmentation: When cognitively plausible learners outperform the ideal. In *Proceedings of the 34th Annual Conference of the Cognitive Science Society*.

Pinker, S. & Ullman, M. 2002. The past and future of the past tense. *Trends in Cognitive Sciences, 6*: 456-463.

Polka, L. & Werker, J.F. 1994. Developmental changes in perception of nonnative vowel contrasts. *Journal of Experimental Psychology: Human Perception and Performance, 20*(2), 421-435.

Rabiner, L.R. 1989. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77: 257-286.

Raphael, L.J. 1972. Preceding vowel duration as a cue to the perception of the voicing characteristics of word-final consonants in American English. *Journal of the Acoustical Society of America, 51*, 1296-1303.

Reynolds, D.A. 1994. Experimental evaluation of features for robust speaker identification. *Speech and Audio Processing, IEEE Transactions on, 2*(4), 639-643.

Rose, S.A., Feldman, J.F. and Jankowski, J.J. 2009. A cognitive approach to the development of early language. *Child Development, 80*(1), 134-150.

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

Selkirk, E.O. (1981) *English Compounding and the Theory of Word-structure*. in: M. Moortgat, H. Van der Hulst & T. Hoestra (eds.) *The Scope of Lexical Rules*, Foris, Dordrecht.

Sethuraman, J. 1994. A constructive definition of dirichlet priors. *Statistica Sinica*, 4: 639-650.

Shi, L., Griffiths, T. L., Feldman, N. H, & Sanborn, A. N. 2010. Exemplar models as a mechanism for performing Bayesian inference. *Psychonomic Bulletin & Review, 17 (4),* 443-464.

Stevens, S.S., Volkman, J. & Newman, E.B. 1937. A scale for the measurement of the psychological magnitude pitch. *The Journal of the Acoustical Society of America, 8*(3), 185-190.

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

Thiessen, E.D. & Saffran, J.R. 2007. Learning to learn: Infant's acquisition of stress-based strategies for word segmentation. *Language Learning and Development, 3*(1), 73-100.

Tversky, A. & Kahneman, D. 1974. Judgment under uncertainty: Heuristics and biases. *Science, 185*(415), 1124-1131

Vallabha, G.K., McClelland, J.L., Pons, F., Werker, J.F., & Amano, S. 2007. Unsupervised learning of vowel categories from infant-directed speech. *Proceedings of the National Academy of Sciences, 104*(33), 13273-13278.

Viikki, O. & Laurila, K. 1998. Cepstral domain segmental feature vector normalization for noise robust speech recognition. *Speech Communication, 25*(1), 133-147.

Viterbi, A.J. 1967. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Transactions on Information Theory, 13*(2): 260-269.

Wang, H. & Mintz, T.H. 2007. A dynamic learning model for categorizing words using frames. *Proceedings of BUCLD, 32*, 525-536.

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.

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

Yang, C.D. 2004. Universal Grammar, statistics or both? *Trends in Cognitive Sciences, 8*(10): 451-456.

Xu, F. & Denison, S. 2009. Statistical inference and sensitivity to sampling in 11-month-old infants. *Cognition, 112*(1), 97-104.

Xu, F. & Garcia, V. 2008. Intuitive statistics by 8-month-old infants. *Proceedings of the National Academy of Sciences, 105*(13), 5012-5015.

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