The utility of cognitive plausibility in language acquisition modeling: Evidence from word segmentation

Lawrence Phillips (lawphill@uci.edu)

Lisa Pearl (lpearl@uci.edu)

Department of Cognitive Sciences

University of California, Irvine

Abstract

The informativity of a computational model of language acquisition is directly related to how closely it approximates the actual acquisition task, sometimes referred to as the model's *cognitive plausibility*. We suggest that though every computational model necessarily idealizes the modeled task, an informative language acquisition model can aim to be cognitively plausible in multiple ways. We discuss these cognitive plausibility checkpoints in general terms, and then apply them to a case study in word segmentation, investigating a promising Bayesian segmentation strategy. We create a more cognitively plausible model of this learning strategy which uses an age-appropriate unit of perceptual representation, evaluates the model output in terms of its utility, and incorporates cognitive constraints into the inference process. Our more cognitively plausible model of the Bayesian word segmentation strategy not only yields better performance than previous implementations but also shows more strongly the beneficial effect of cognitive constraints on segmentation. One interpretation of this effect is as a synergy between the naive theories of language structure that infants may have and the cognitive constraints that limit the fidelity of their inference processes, where less accurate inference approximations are better when the underlying assumptions about how words are generated are less accurate. More generally, these results highlight the utility of incorporating cognitive plausibility more fully into computational models of language acquisition.

Keywords: language acquisition; Bayesian learning; computational modeling; cognitive plausibility; statistical learning; word segmentation

1 Introduction

Language acquisition has long been of interest in cognitive science due to the complexity of the knowledge system acquired and the rapidity of its acquisition. Developmental experiments have revealed much about the precise timeline and certain aspects of the acquisition process, such as the information children are sensitive to in the input (e.g., Saffran, Aslin, & Newport, 1996; Mattys, Jusczyk, & Luce, 1999; Maye, Werker, & Gerken, 2002) and what learning capabilities they possess at different developmental stages (e.g., Thiessen & Saffran, 2003; Bortfeld, Morgan, Golinkoff, & Rathbun, 2005). Computational modeling is a complementary tool that can be used to delve deeper into the acquisition process by evaluating the learning strategies children may use, including both the information they might utilize and how exactly they might utilize it. Importantly, because computational models require us to be explicit about all relevant aspects of the acquisition task being modeled (including input representation, output representation, and the inference process), models make empirical claims about these different components of acquisition. Modeling results can therefore impact our understanding of both the mental representations children have at different points in language development and the process itself (Kol, Nir, & Wintner, 2014; Pearl, in press). Moreover, computational models have a distinct advantage with respect to explaining how acquisition occurs: once we identify a successful strategy via computational modeling, we can then scrutinize the inner workings of the model to understand exactly *why* that strategy works the way it does (something which is of course considerably harder to do experimentally with children's brains).

Notably, the usefulness of a computational model is directly related to how closely it approximates the actual acquisition task, i.e., how "cognitively plausible" it is. Of course, every computational model idealizes and simplifies the modeled task because it is (currently) impossible to include every detail of the acquisition process, and many details may be irrelevant

anyway. However, as Box and Draper (1987) note, "all models are wrong, but some are useful". We suggest that a useful model of language acquisition can strive to be cognitively plausible in multiple ways, relating to both computational-level and algorithmic-level considerations (in the sense of Marr (1982)).

In the remainder of this paper, we first discuss these different cognitive plausibility checkpoints, and subsequently apply them to a case study in word segmentation, where a particular Bayesian word segmentation strategy has shown promise. We then discuss how to create a more cognitively plausible model of the Bayesian learning strategy, including considerations of the input representation, the output evaluation, and the inference process.

Our results suggest that using a more plausible input representation (the syllable), measuring the output in terms of useful units (rather than against adult orthographic segmentation), and incorporating cognitive limitations into the inference process all lead to better word segmentation performance. This provides stronger support for the Bayesian approach to word segmentation, showing that it is robust to changes in the unit of representation as well as changes to the inference process itself. We also discuss the beneficial effect of cognitive constraints on segmentation, suggesting that there may be an advantageous relationship between the naive theories of language that young infants may have and the constraints that limit the fidelity of their inference process. More generally, these results underscore the utility of integrating cognitive plausibility into multiple levels of a model of language acquisition.

2 Cognitive plausibility in computational modeling

Two general classes of computational models have been used most often to understand the language acquisition process: (i) computational-level models (sometimes called "ideal learner" models) that are concerned primarily with examining the learning assumptions that would be

useful for children (e.g., Mintz, 2003; M. Johnson, Griffiths, & Goldwater, 2007; Feldman, Griffiths, & Morgan, 2009; Goldwater, Griffiths, & Johnson, 2009; Dillon, Dunbar, & Idsardi, 2013; Feldman, Griffiths, Goldwater, & Morgan, 2013), and (ii) algorithmic-level models that are concerned primarily with the learning assumptions that are usable by children, who have various cognitive limitations (e.g., Freudenthal, Pine, & Gobet, 2006; Legate & Yang, 2007; Wang & Mintz, 2008; Lignos & Yang, 2010; Blanchard, Heinz, & Golinkoff, 2010; Pearl, Goldwater, & Steyvers, 2011). Because these two model classes have complementary aims, it is often productive to use both model types when examining how a particular acquisition task is solved. For example, if certain learning assumptions are found to be useful for solving a problem when cognitive constraints are not a factor (as in a computational-level model), it can then be worthwhile to investigate if these same assumptions are still useful once cognitive factors impact the learning process (as in an algorithmic-level model). Similarly, if certain learning assumptions yield good results when cognitive constraints hinder the learning process, it can be worthwhile to investigate if these same assumptions are useful in the absence of those constraints or if instead there is some interesting interaction between the learning assumptions and the cognitive constraints. Notably, cognitive plausibility considerations can apply at both the computational and algorithmic level of any model.

2.1 Computational-level considerations

At the computational level, the input and the output must be defined. For input, the most realistic models would learn from data that a child of the age being modeled would learn from, and represent that input in the same manner as the child. For output, the most realistic models would evaluate the modeled learner's acquired knowledge against the knowledge that a child of the age being modeled has acquired. We discuss each in turn.

2.1.1 Input

Traditionally, the choice of input for computational acquisition models has been a delicate balancing act between what input is available to the modeler and what input children would use. For example, in modeling vowel acquisition, one would ideally want acoustic data drawn from naturalistic child-directed speech. Since obtaining this type of data is difficult, one clever approach has been to approximate it using available experimental productions (Vallabha, McClelland, Pons, Werker, & Amano, 2007; Feldman et al., 2009; Toscano & McMurrary, 2010).

Likewise, the choice of input representation often requires this same balancing act, given the availability of various input encodings. For example, several models of the early stages of word segmentation (Brent, 1999; Goldwater et al., 2009; Blanchard et al., 2010) used available phonemic encodings of a child-directed speech corpus from Brent and Cartwright (1996), though infants in the early stages of word segmentation may not encode the input as a sequence of phonemic segments (Werker & Tees, 1984; Werker & Lalonde, 1988; Bertonicini, Bijeljac-Babic, Jusczyk, Kennedy, & Mehler, 1988; Mehler, Dupoux, & Segui, 1990; Jusczyk, 1997; Eimas, 1997, 1999). Importantly, the input representation may well affect acquisition success – a learning strategy that fails with one input representation may succeed with another, and vice versa. Thus, matching the input representation of the child as closely as possible is important for an informative computational acquisition model.

2.1.2 Output

On the output side, computational models of acquisition have traditionally been evaluated against adult knowledge, sometimes called the "gold standard" (Mintz, 2003; Goldwater et al., 2009; Blanchard et al., 2010; Lignos & Yang, 2010; Feldman et al., 2013). Yet, this may not be an accurate representation of the knowledge children achieve. For example, in the first

stages of word segmentation, it is not necessary for children to perfectly segment a fluent stream of speech the way an adult would (and in fact, anecdotal evidence suggests they produce systematic "chunking" errors such as segmenting *that a* as *thata* (Brown, 1973)). Nonetheless, word segmentation results have typically been measured against perfect adult segmentation (e.g., Goldwater et al., 2009; Lignos & Yang, 2010; Blanchard et al., 2010; Pearl et al., 2011). A more realistic evaluation would be against a knowledge state that better matches young children's knowledge at this stage. Because of this, it is helpful to consider what is known about children's learning trajectory when deciding what the model's desired output should be (e.g., the chunking errors mentioned above).

2.2 Algorithmic-level considerations

The above computational-level considerations are relevant for both computational-level models and algorithmic-level models, since both model types need to represent the input and output appropriately. Similarly, algorithmic-level considerations also apply for both model types, since both need to implement the learning process. Bayesian models handily separate these into issues concerning model specification and issues concerning the inference process, but they correspond more generally into how the hypothesis space and learning assumptions are defined vs. how the learner's beliefs are updated.

2.2.1 Hypothesis space & learning assumptions

Both the hypothesis space and the learning assumptions built into the model represent the previous knowledge a child has. The hypothesis space defines the hypotheses under consideration and how relatively probable each one is. Bayesian models include this in the model specification, and encode the relative hypothesis probabilities in the prior. The learning assumptions impact how different hypotheses are evaluated, and may appear in the prior of Bayesian models

if they are in effect before data are encountered or in the likelihood if they occur once data have been encountered. In general, the hypothesis space and learning assumptions are where many theories of language acquisition are directly translated in a computational model, as theories typically describe exactly this aspect of the acquisition process (Mintz, 2003; Yang, 2004; Freudenthal et al., 2006; Legate & Yang, 2007; Frank, Goodman, & Tenenbaum, 2009; Feldman et al., 2013). While both the hypothesis space and learning assumptions must be made explicit in a computational model, it may be that the learning theory on which the model is based is not specific about certain aspects (e.g., for a statistical word segmentation strategy, what syntactic knowledge infants have). Sometimes, there may not be clear empirical data supporting any particular decision about those aspects, and so for these aspects, the computational model can provide the empirical data. If the strategy as instantiated by the model is successful, this is support for those particular decisions. More generally, when a model of a learning strategy is successful, this is an existence proof that this hypothesis space and set of learning assumptions can work, and so the learning strategy itself is viable (Pearl, in press).

2.2.2 Inference process

A learner's beliefs about different hypotheses are updated during learning, given the hypothesis space, learning assumptions, and input data. In a Bayesian model, belief update is accomplished via Bayesian inference, given the prior and likelihood. This is the step where computational-level and algorithmic-level models traditionally diverge, as algorithmic-level models seek to incorporate cognitive constraints into this process while computational-level models do not. Computational-level models typically use an inference process that is known to be optimal (e.g., Gibbs sampling: Feldman et al., 2009; Goldwater et al., 2009; Christodoulopoulos, Goldwater, & Steedman, 2011). Algorithmic-level models typically incorporate cognitive limitations into the inference process (Brent, 1999; Lignos & Yang, 2010; Pearl et al., 2011), sometimes

assuming humans approximate rational inference but do not achieve it due to these limitations and other non-optimal biases humans may have (Tversky & Kahneman, 1974).

3 Incorporating cognitive plausibility into a word segmentation model

We now take these cognitive plausibility considerations and apply them to a case study: a computational model of word segmentation. Our goal is to make a more informative implementation of a Bayesian word segmentation strategy that has previously shown promise (Goldwater et al., 2009; Pearl et al., 2011). We will first describe the learning strategy itself, and then discuss how cognitive plausibility considerations are dealt with at the level of input representation, output evaluation, and inference process.

3.1 The Bayesian word segmentation strategy

Bayesian learning strategies explicitly distinguish between the learner's pre-existing beliefs via the *prior* ($P(h)$) and how the learner evaluates incoming data via the *likelihood* ($P(d|h)$). This information is combined using Bayes' theorem (1) to generate the updated beliefs of the learner via the posterior ($P(h|d)$). 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, Tenenbaum, Griffiths, & Xu, 2011).

$$P(h|d) \propto P(d|h)P(h) \tag{1}$$

The Bayesian word segmentation strategy we investigate is one kind of purely statistical learning strategy that does not rely on language-dependent information, making it a good can-

didate for the early stages of word segmentation when an infant does not yet know (m)any words of the language. The underlying Bayesian models for all of our learners were originally described by Goldwater et al. (2009) (**GGJ** henceforth). These Bayesian models infer a lexicon of word forms from which the observable data are generated (i.e., the observable data are sequences of word tokens drawn from the inferred lexicon). The prior for these models favors "simpler" hypotheses, where simpler translates to two distinct biases: the learner prefers (i) a smaller lexicon and (ii) shorter words in that lexicon. Because these models are generative, meaning that they predict how the words in the utterances of the observable data are produced, the modeled learner must have an explicit idea of how this occurs. Given the limited knowledge of language structure which infants likely possess at the relevant 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 selected with no relation to each other. To encode this assumption in the model, GGJ use a Dirichlet Process (T. Ferguson, 1973), which supposes that the observed sequence of words $w_1 ... w_n$ is generated sequentially using a probabilistic generative process. In the unigram case, the identity of the *i*th word is chosen according to (2), where the probability of the current word is a function of how often it has occurred previously.

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

$n_{i-1}(w_i)$ is the number of times word w_i appears in the previous $i - 1$ words, α is a free parameter of the model which encodes how likely a novel word is encountered, 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) $x_1 ... x_m$.

$$P_0 = P(w = x_1 \dots x_m) = \prod_j P(x_j) \tag{3}$$

P_0 can be interpreted as a parsimony bias, giving the model a preference for shorter words: the more units that comprise a word, the smaller the probability of that word, and so shorter words are more probable. 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. We note that this model does not incorporate information about transitional probabilities, but simply relies on the frequency of lexical items. Because it does not know the true word frequencies (as it is trying to learn the words in the first place), it estimates these based on the number of times it believes the word has appeared previously in the input.

The second model makes a slightly more sophisticated assumption about the relationship between words, where a word is assumed to be related to the previous word (a *bigram* assumption). More specifically, a word is generated based on the identity of the word that immediately precedes it, encoded in a hierarchical Dirichlet Process (Teh, Jordan, Beal, & Blei, 2006). This model additionally tracks the frequencies of two-word sequences and is defined as in (4-5):

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

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

where $n_{i-1}(w', w_i)$ is the number of times the bigram (w', w_i) has occurred in the first $i-1$ words, $n_{i-2}(w')$ is the number of times the word w' occurs in the first $i-2$ words, $b_{i-1}(w_i)$ is the

number of bigram types which contain w_i as the second word, b is the total number of bigram types previously encountered, P_0 is defined as in (3), and β and γ are free model parameters. Both the β and γ parameters, similar to the α parameter in (2), control the bias towards fewer unique bigrams (β) and towards fewer unique lexical items (γ). Like the unigram model, the bigram version tracks the perceived frequency of lexical items, as well as the frequency of bigram pairs. Though it does not calculate transitional probabilities between syllables, it does rely on the transitional probabilities between words that comprise bigrams via (4). In particular, a word w_i which has more frequently followed w' (i.e., has a high transitional probability) is more likely to be generated in bigrams that begin with w'.

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

3.2 Input representation

3.2.1 Empirical grounding

Similar to previous computational models of this Bayesian segmentation strategy (Brent, 1999; Goldwater et al., 2009; Pearl et al., 2011), we use input drawn from samples of American English child-directed speech. Because child-directed speech is known to vary from typical adult-directed speech in many systematic ways (C. Ferguson, 1964; Snow, 1977; Grieser & Kuhl, 1988; Fernald et al., 1989), it is more realistic to use speech directed at children of the age when early word segmentation is occurring.

Our data come from the Pearl-Brent derived corpus (Pearl et al., 2011) from CHILDES

(MacWhinney, 2000). This encoding of the Brent corpus (Brent & Siskind, 2001) contains 100 hours of phonemically-encoded, child-directed speech from 16 mother-child pairs. Because we are investigating early word segmentation, we restrict the input to the subset of utterances directed at children nine months or younger. This leaves 28,391 utterances in the corpus, containing 96,723 word tokens of 3,221 individual word types (average: 3.4 words, 4.2 syllables, and 10.4 phonemes per utterance).

3.2.2 Unit of representation

The most notable early word segmentation model to incorporate statistical learning is the MBPD model of Brent (1999), which used the phoneme as the basic unit of input representation, and many subsequent segmentation models assumed the same (Batchelder, 2002; Fleck, 2008; Goldwater et al., 2009). The ready availability of phonemically encoded corpora and phonological dictionaries certainly played a role in this decision and there has often been little discussion about why the phoneme should be the unit of representation for modeling early word segmentation.

To identify the appropriate unit of input representation for the first stages of word segmentation, it is important to determine when that process occurs and what the infant representation of the input is likely to be like at that point. The earliest evidence for word segmentation occurs at six months, when infants are able to use highly frequent and familiar words, such as their own names, to segment adjacent words (Bortfeld et al., 2005). We assume here that this represents a lower bound for when word segmentation begins.

For statistical word segmentation strategies like the Bayesian learning strategy we investigate, a number of studies show that infants between the ages of seven and nine months are capable of using statistical cues to segment unfamiliar words from fluent speech (Saffran et al., 1996; Mattys et al., 1999; Thiessen & Saffran, 2003; Pelucchi, Hay, & Saffran, 2009).

Interestingly, when language-dependent cues such as lexical stress patterns are pitted against statistical cues, eight- and nine-month-olds prefer to use the language-dependent cues (E. Johnson & Jusczyk, 2001; Thiessen & Saffran, 2003) while seven-month-olds prefer to use statistical cues (Thiessen & Saffran, 2003). This suggests that statistical word segmentation likely begins somewhere around seven months and is supplanted by more precise language-specific strategies over the course of the next two months. We assume then that the input representation we want to use for a purely statistical learning strategy is the one infants possess around seven months.

3.2.2.1 Infant input representation. The question of what constitutes the basic unit of infant speech perception at this age is a controversial one. The two options most frequently proposed are that infants perceive words as sequences of syllables or phonetic segments. Generally, evidence from infants younger than six months suggests that they perceive syllables more easily (Jusczyk & Derrah, 1987; Bertonicini et al., 1988; Bijeljac-Babic, Bertoncini, & Mehler, 1993; Eimas, 1999), while there is some evidence that phonetic segments are being learned after this (Polka & Werker, 1994; Pegg & Werker, 1997; Jusczyk, Goodman, & Baumann, 1999; Maye, Weiss, & Aslin, 2008). However, it should be noted that no study to date has conclusively shown that infants perceive only syllables (or only segments), with most experimental results compatible with either perceptual unit. We review the experimental literature below to determine that it is possible that infants at seven months perceive speech as a sequence of syllables, rather than phonemes.

First, very young infants are capable of distinguishing syllables which differ by a single segment, but are not able to group syllables based on their internal segments (Jusczyk & Derrah, 1987; Bertonicini et al., 1988; Jusczyk, Jusczyk, Kennedy, Schomberg, & Koenig, 1995). For example, Jusczyk et al. (1995) tested two-month-olds on their ability to remember a set of syllables after a two minute delay, where the three syllables heard during training either shared a

common segment (e.g. [bi], [ba], [bu]) or not (e.g., [si], [ba], [tu]). It turned out that infant reactions did not depend on whether the training syllables shared a common segment, and Jusczyk et al. (1995) interpreted this as two-month-olds not recognizing the segmental similarities between syllables. These results align with previous work with two-month-olds by Jusczyk and Derrah (1987) and Bertonicini et al. (1988), where infant reactions to familiar vs. novel syllables also did not depend on whether the familiar syllables shared a common segment. Thus, it is unlikely that the two-month-olds represented the speech as a sequence of segments.

Additionally, much research has been conducted on young infants' ability to perceive the similarity of words as determined by the number of shared syllables or segments. Bijeljac-Babic et al. (1993) found that newborns can categorize utterances by the number of syllables the utterances possess, but not by the number of segments. Jusczyk et al. (1995) found that two- and three-month-olds are better able to detect novel bisyllables that do not possess a common initial syllable (e.g., /balo/ and /pamʌl/), which suggests that infants perceive bisyllabic utterances sharing an initial syllable (e.g., /balo/ and /banʌl/) as more similar. Notably, when these infants were trained on a set of bisyllables that shared the phonemes /a/ and /b/, always in that order (e.g., /labo/ and /zabi/), infants showed no sign of recognizing novel bisyllables that fit (or did not fit) that pattern (e.g., /nabʌl/ vs. /banʌl/). Eimas (1999) found that three- to four-month-olds are able to categorically represent bisyllabic utterances sharing a common initial syllable (e.g., /bapi/ and /batad/), and suggestive evidence that this is also true for bisyllabic utterances sharing a common final syllable (e.g., /piba/ and /tadba/). In contrast, no evidence was found that monosyllabic utterances sharing a common initial consonant (e.g., /bɪd/ and /bæd/) were categorically represented. This inability to recognize the similarity of syllables with shared internal structure suggests that segmental similarity is not salient at this age. Jusczyk (1997) summarized the state of the field at the time by noting that "there is no indication that infants under six months of age represent utterances as strings of phonetic segments".

However, there is some evidence that eight-month-olds possess the ability to recognize featural similarities between segments. Maye et al. (2008) found that eight-month-olds can discriminate a non-native voice-onset time (VOT) contrast after exposure to a bimodal distribution of tokens, and will generalize this contrast across phonetic features (e.g., infants were trained on VOT distributions over dental sounds, but tested on bilabial sounds). This suggests that eight-month-olds can perceive phonetic features, and would seem to indicate that eight-month-olds possess enough knowledge of their language's phonology to perceive the input as segments. It has also been argued that the gradual loss of non-native phonetic contrasts corresponds to the acquisition of segmental phonology in infants (Werker & Lalonde, 1988), and this loss occurs for various consonants between ten and twelve months (Werker & Tees, 1984; Werker & Lalonde, 1988; Best, McRoberts, & Sithole, 1988; Best, McRoberts, LaFleur, & Silver-Isenstadt, 1995) but for various vowels around six months (Kuhl, Williams, Lacerda, Stevens, & Lindblom, 1992; Polka & Werker, 1994). Still, Jusczyk (1997) notes that these results "do not necessarily require that speech be represented as strings of phonetic segments", as the particular experimental results are consistent with a syllabic unit of representation. More generally, Jusczyk (1997) noted that "considerably more research is needed to determine the level of detail present in any long-term storage of speech information by older infants", and that remains true today.

So, while no clear conclusions can be drawn from the experimental literature about the basic unit of seven-month-old speech perception, there is evidence that younger infants find it easier to represent commonalities between syllables. Moreover, from a utilization standpoint, infants have been shown to use statistical cues based on syllables at seven to eight months (Saffran et al., 1996; Thiessen & Saffran, 2003; Pelucchi et al., 2009). In contrast, infants do not use phonotactic cues, which depend on sequences of phones or phonemes, until nine months (Jusczyk, Friederici, Wessels, Svenkerud, & Jusczyk, 1993). Given this, we suggest that syllables are a cognitively plausible basic unit of representation at this stage of development,

though there is certainly much perceptual reorganization occurring at the same time. Because of this, to determine an early word segmentation strategy's viability, it seems necessary to consider its performance when the syllable is the basic unit of perception.

3.2.2.2 Cognitively plausible model input. We converted the Pearl-Brent derived corpus (Pearl et al., 2011) into a syllabic form using a two-pronged approach. First, we 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 *ba-dido* or proper names like *Brenda's*), we automatically syllabified our 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 approximating the kind of syllabification that infants might use, since there is a lack of experimental evidence regarding the exact nature of infant syllabification. Approximately 25% of lexical items were syllabified automatically using this approach.[1] Each unique syllable was then treated as a single, indivisible unit, losing all sub-syllabic phonetic (and phonotactic) information. This is most similar to what experimental results from two- to four-month-olds suggest, and so seemed most cognitively plausible given that it is unknown what changes to the unit of perception are occurring in the few months afterwards.

However, we do note that there are imperfections in the way the syllabic unit of perception is instantiated in our model. First, it is unknown how infants actually syllabify naturalistic speech. We used adult syllabification and the Maximum-Onset Principle as an approximation, but this does not necessarily correspond to infant syllabification.

Second, one critique of phoneme-based models is that infants perceive phones rather than

[1] Of the words syllabified using human judgments, only 3.6% of the human judgments from the MRC database differ from what automatic syllabification would predict.

phonemes. This problem is not alleviated by representing the input as syllables the way we do, since syllabification occurs over phonemically-encoded rather than phonetically-encoded speech. Experimental support for phone-based perception comes from Seidl, Cristià, Bernard, and Onishii (2009), who trained English and French infants on a pattern where fricatives followed nasal vowels while stops followed oral vowels. Notably, the oral/nasal vowel distinction is phonemic in French, but not in English. Seidl et al. (2009) found that both eleven-month-old French infants and four-month-old English infants could learn the pattern, while eleven-month-old English infants could not. They interpreted this as the French eleven-month-olds succeeding by perceiving the phonemic contrast of oral vs. nasal vowels while the English eleven-month-olds failed because they could not perceive this contrast any longer. However, the four-month-old English infants could succeed because they still perceived the phonetic detail of the oral and nasal vowels, and thus could learn the pattern by perceiving it as a sequence of phones. Notably, phonetic detail is rarely included in word segmentation models (although see Batchelder (2002) and Fleck (2008)) because that level of detail is not encoded in most child-directed speech corpora.

Third, syllabification occurs within words in our model (e.g., *who's afraid* becomes *who's a fraid*, rather than *who 'sa fraid*). This may not pose a problem for languages such as German, which disallow post-lexical resyllabification (Hall, 1992), but English is likely not one of these languages (Labov, 1997). So, this is another place where future work may be able to better approximate infant perception of the input.

3.3 Output evaluation

Early word segmentation models have predominantly been evaluated against how an adult would segment a fluent speech stream. Typically, the segmentation assumed is orthographic words (rather than phonological words, for example), because this encoding is readily available

for many suitable corpora. However, this seems an unlikely outcome for the early stages of the word segmentation, which may be more oriented towards bootstrapping the word segmentation process. For example, successful early segmentation could yield units that are useful for subsequent stages of word segmentation (such as identifying language-specific cues to segmentation) and for subsequent language acquisition tasks (such as grammatical categorization or syntactic learning).

So, one way to assess a model's output in a more cognitively plausible manner is to see if it generates useful units, rather than adult orthographic words. For example, some oversegmentation "errors" may result in real words that are fairly frequent (e.g., *alright* /ɑl ɹajt/ segmented as *all* /ɑl/ and *right* /ɹajt/). Notably, the Bayesian learner relies on a word's perceived frequency to make its segmentation decisions, and this kind of error impacts the perceived frequencies of the words involved: the true word (*alright*) has a lower frequency than it should, but the "error" words *all* and *right* have a higher frequency than they should. Although this means that the learner is less likely to accurately segment *alright* when it is later encountered, the learner is more likely to accurately segment *all* and *right* when either word is seen. So, this type of "reasonable error" is likely to boost performance if the oversegmented words are frequent.

Additionally, some oversegmentation "errors" may result in morphological components (typically productive morphology) that are useful for grammatical categorization and could help segment other root forms (e.g., segmenting off *–ly* /li/ or *–ing* /iŋ/). Similar to the oversegmentation error above, the perceived frequency of the affixes will make a Bayesian learner more likely to identify both the morphology and the root forms as separate units. Since both of these are useful units, this will boost the learner's performance for other tasks. Notably, because we use syllables as the basic unit of representation, only syllabic morphology can be successfully identified this way (e.g., the learner cannot segment the plural morpheme /-s/ from the word *cats* because the single syllable /kæts/ cannot be subdivided).

As another example, undersegmentation errors may produce function word collocations that act as coherent pieces in syntactic rules (e.g., segmenting *Could I* as *couldI*, or *is that a* as *isthata*). Anecdotal evidence suggests older children do produce errors of this kind, where a phrase is treated as single unanalyzed chunk, e.g., *thata* and *what'sthat* from Brown (1973) and *lookatthat* and *openthedoor* from Peters (1983). Since older children make these errors, it seems likely that infants would as well.

To evaluate our model of early word segmentation, we first use the adult orthographic segmentation as a "gold standard", as this allows for easy comparison to prior segmentation studies. We then adjust our definition of the desired segmentation output to better match the likely target segmentation for early word segmentation, and count the three "error" types mentioned above as valid segmentations: (i) real words, (ii) morphological units, and (iii) function word collocations.

3.4 Inference process

Many previous models of Bayesian learning strategies use ideal learners to investigate the utility of the learning assumptions encoded by the model when inference is guaranteed to be optimal (Johnson 2008, Frank et al., 2009; Goldwater et al., 2009; Dillon et al., 2013; Feldman et al., 2013). However, as human inference may not be optimal (though it may approximate optimal inference in some cases), it is worth seeing whether the learning assumptions of the word segmentation strategy we investigate here are as useful when inference is not optimal.

We begin with a standard inference algorithm used by ideal learner models, called Gibbs sampling (Geman & Geman, 1984). Gibbs sampling is a Markov chain Monte Carlo algorithm, and operates by first guessing the value of every hidden variable in the model (in this case, whether a word boundary exists or not). It then iterates through every variable (i.e., potential boundary position) and a value for that variable is chosen (i.e., actual boundary or not) condi-

tioned on the current value of all other variables. This is repeated and will, in almost all cases, converge in the limit on the underlying joint distribution. 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 assumptions of the Bayesian word segmentation strategy, assuming Bayesian inference can be carried out in some manner (e.g., using exemplar models (Shi, Griffiths, Feldman, & Sanborn, 2010) or sparse-distributed memory systems (Abbott, Hamrick, & Griffiths, 2013) to implement importance sampling, or using a particle filter (Sanborn, Griffiths, & Navarro, 2010)). Because this learner processes the input in a batch and finds what it considers the optimal segmentation, we refer to it as the **BatchOpt** learner. To make the modeled learner's inference process more cognitively plausible, we then include different cognitively-inspired processing constraints implemented by the three constrained learners of Pearl et al. (2011).

3.4.1 Adding incremental processing

The first constraint is to make input processing incremental, so that data are processed as the learner encounters them rather than being saved up into a large batch that is processed all at once. The Dynamic Programming with Maximization (DPM) learner of Pearl et al. (2011) is the most direct translation of the ideal learner into an incremental learner, and is essentially equivalent to the online learner presented in Brent (1999). It processes each utterance in the corpus one at a time and attempts to choose at each point the optimal segmentation (i.e., the one with maximum probability) using the Viterbi algorithm (a kind of dynamic programming method), given what it has seen before. This behavior is a form of greedy optimization, po-

tentially leading the learner to a locally-optimal point in the segmentation hypothesis space, rather than converging on the globally-optimal segmentation, in the manner that Gibbs sampling would. Because of this online process of finding the locally optimal segmentation, we refer to this learner as the Online Optimal learner (**OnlineOpt**).

3.4.2 Adding sub-optimal decision-making

The second constraint is to sample probabilistically from the set of possible segmentations, rather than always choosing the one that has the maximum probability. This type of sampling of hypotheses is consistent with evidence that children make inferences by sampling from the posterior distribution (Denison, Bonawitz, Gopnik, & Griffiths, 2013). So, for instance, if there are two possible segmentations, A and B, and the model decides that $p(A|data) = .75$ and $p(B|data) = .25$, this learner will choose segmentation A with probability 0.75 and segmentation B with probability 0.25. This process is similar to a single-particle particle filter, where a probability distribution is represented as a single point estimate. Interestingly, while the model itself is quite different, Sanborn et al. (2010) show that for models of category judgments, a single-particle particle filter approximates human judgment patterns quite reasonably. This lends some support to the plausibility of this type of inference algorithm for modeling human learning. This algorithm also potentially avoids the pitfall of the OnlineOpt learner by allowing locally sub-optimal decisions that may turn out to be more optimal globally. This learner also uses dynamic programming to incrementally process the input (in this case, the Forward algorithm to calculate segmentation probabilities), and so was called the Dynamic Programming with Sampling (DPS) learner by Pearl et al. (2011). We will refer to it as the Online Sub-Optimal (**OnlineSubOpt**) learner, since it processes the data incrementally but may choose a locally sub-optimal segmentation.

3.4.3 Adding a recency effect

The third constraint is to implement a recency effect, which may be thought of as a kind of short-term memory. A learner with this recency effect focuses its processing resources on more recent data, rather than giving all data equal attention. This effect is implemented using a version of Gibbs sampling called Decayed Markov Chain Monte Carlo (Marthi, Pasula, Russell, & Peres, 2002), and the learner using this was called the DMCMC learner by Pearl et al. (2011). We refer to it as the Online Memory (**OnlineMem**) since it processes utterances online and has memory biases that constrain where it focuses its attention.

For every utterance, the OnlineMem learner samples some number s of previous potential word boundaries. We set s to 20,000 for the simulations reported below; this amounts to 74% less processing than the syllable-based BatchOpt learner using regular Gibbs sampling and so represents a significant processing reduction. The probability of sampling a boundary b is proportional to the decay 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 a boundary is from the end of the current utterance, the less likely it is to be sampled, and larger values of d indicate a stricter memory constraint. For our simulations, a set, non-optimized value of d=1.5 was utilized to implement a heavy memory constraint. This results in 83.6% of boundary samples occurring in the current utterance, with only 11.8% in the previous utterance and the remaining 4.6% located in other utterances, predominantly in the next-most-previous utterance.

For each sampled boundary, the learner updates its beliefs about whether a boundary exists, conditioned only on the utterances already encountered. Because of the decay function, this learner's sampling is heavily biased towards boundaries in recently seen utterances, thus implementing a recency effect that crudely mimics short-term memory. Intuitively, this can be

thought of as the learner having not only the current utterance in memory, but some decaying version of previously heard utterances. Thus, something heard in the current utterance can potentially allow the learner to change its mind about something in the previous utterance.[2]

3.4.4 Learner summary

Table 1 summarizes the differences between the different Bayesian learners we investigate. Of the three constrained learners, we suggest that the OnlineMem learner is the most cognitively plausible since it incorporates both online processing and a type of short-term memory. So, we will be particularly interested in the performance of the modeled learner using this inference procedure, since it is the one we believe is most realistic of the ones we investigate and therefore most informative for understanding infant word segmentation.

	Parameters	Learning assumptions		
		online processing	sub-optimal decisions	recency effect
BatchOpt	iterations $= 20,000$	-	-	-
OnlineOpt	N/A	+	-	-
OnlineSubOpt	N/A	+	+	-
OnlineMem	samples per utterance $= 20,000$ decay rate $= 1.5$	+	-	+

Table 1: Summary of modeled learners used for Bayesian word segmentation, including the relevant parameters and learning assumptions encoded by each. All learners use the following Bayesian model parameters: $\alpha = 1$, $\beta = 1$, $\gamma = 90$.

[2]The OnlineOpt and OnlineSubOpt learners can also be thought of as having recency effects insofar as they are constrained to processing only the most recent utterance. In this case, what is held in memory is the entire utterance rather than some set of previous boundary positions. Notably, this contrasts with the recency effect of the OnlineMem learner as the OnlineOpt and OnlineSubOpt learners do not have a preference to process items that occur at the end of the current utterance.

4 Results

To evaluate the modeled learners' performance, we compare the segmentations they produce against a target segmentation (i.e., the adult knowledge gold standard segmentation or the adjusted segmentation that includes certain "reasonable" errors). The metrics we use are standard for word segmentation model evaluation: precision (6), recall (7), and F-score (8), where F-score is the harmonic mean of precision and recall and provides a convenient summary statistic indicating how accurate and complete a segmentation is (as measured by precision and recall, respectively).

$$Precision = \frac{\#\, correct}{\#\, guessed} = \frac{\#\, true\, postives}{\#\, true\, positives + \#\, false\, positives} \tag{6}$$

$$Recall = \frac{\#\, correct}{\#\, true} = \frac{\#\, true\, positives}{\#\, true\, positives + \#\, false\, negatives} \tag{7}$$

$$F\text{-}score = \frac{2 * Precision * Recall}{Precision + Recall} \tag{8}$$

Due to length considerations, we consider F-scores only over word tokens, though we note that these metrics can be used for different units relevant for word segmentation: 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}). Word token F-score corresponds to how well the modeled learner was able to recover the target words from a fluent stream of speech, and frequent words impact the segmentation F-score more than infrequent words.

Because we are interested in how well the acquired segmentation knowledge generalizes to

data the learner has not encountered before, we use separate training and test sets. In particular, the learner uses the training set to learn what lexicon items tend to appear and how often they tend to appear, and then applies this knowledge to segment the test set. Five training and test sets are used to ensure that any vagaries of a particular data set are averaged out, where the training set consists of 90% of the corpus and the test set consists of the remaining 10%. Each training-test set pair was a random split of the subset of the Pearl-Brent corpus described in section 3.2.1. All results presented here are averaged over the results of the five data sets, with standard deviations given in parentheses.

4.1 Unit of representation

4.1.1 The Bayesian segmentation strategy

The first question is whether the Bayesian segmentation strategy can succeed when syllables are the unit of representation, since it previously succeeded when using phonemes. Table 2 shows the segmentation performance for Bayesian learners using either a unigram or bigram assumption, comparing the segmentations of phoneme-based and syllable-based learners against the gold standard segmentation.

		Unigram		Bigram	
		Phoneme	Syllable	Phoneme	Syllable
Bayesian	**BatchOpt**	55.0 (1.5)	53.1 (1.3)	69.6 (1.6)	77.1 (1.4)
	OnlineOpt	52.6 (1.5)	58.8 (2.5)	63.2 (1.9)	75.1 (0.9)
	OnlineSubOpt	46.5 (1.5)	63.7 (2.8)	41.0 (1.3)	77.8 (1.5)
	OnlineMem	60.7 (1.2)	55.1 (0.3)	71.8 (1.6)	86.3 (1.2)
Other	**Lignos2012**	7.0 (1.2)	87.0 (1.4)		
	TPminima	52.6 (1.0)	13.0 (0.4)		

Table 2: Phoneme-based and syllable-based segmentation results compared against the adult gold standard for Bayesian and non-Bayesian learners, showing average word token F-score with standard deviations in parentheses.

We first replicate the phoneme-based results in Pearl et al. (2011): (i) bigram learners typically perform better than their unigram counterparts (except the OnlineSubOpt learner), and (ii) the OnlineMem constrained unigram learner significantly outperforms the BatchOpt ideal unigram learner. Turning to the syllable-based results, we again find that bigram learners outperform their unigram counterparts, but much moreso (e.g., OnlineMem bigram=86.3 vs. OnlineMem unigram=55.1). In addition, the syllable-based bigram learners significantly outperform their phoneme-based counterparts (e.g., BatchOpt syllable-based=77.1 vs. phoneme-based=69.6; OnlineMem syllable-based=86.3 vs. phoneme-based=71.8). This suggests that the utility of the bigram assumption is heightened for learners perceiving the input as a stream of syllables. The unigram assumption's utility is not consistently strengthened when going from phoneme-based to syllable-based learners, however: only the OnlineOpt and OnlineSubOpt syllable-based learners outperform their phoneme-based counterparts. Still, the performance gain is quite substantial for those two learners (e.g., OnlineSubOpt phoneme-based=46.5 vs. syllable-based=63.7). Interestingly, we also find that several constrained learners outperform their ideal counterparts (e.g., unigram OnlineSubOpt=63.7 vs. BatchOpt=53.1, bigram OnlineMem=86.3 vs. BatchOpt=77.1). We discuss this effect more in section 4.3, but note here that this indicates this interesting behavior is more robust for a syllable-based learner than for a phoneme-based learner.

4.1.2 A comparison with other segmentation strategies

While we currently believe syllables are a more cognitively plausible unit of representation, it may turn out that future research identifies phoneme-like units as more likely. In that case, the Bayesian strategy seems to fare about as well (i.e., it is generally still successful). But what about other segmentation strategies that are currently designed to operate over syllables? Table 2 shows the results from two other segmentation strategies: the subtractive segmentation

strategy from Lignos (2012) and a learner who posits boundaries at transitional probability minima (TPminima), investigated by Yang (2004). These two strategies provide a comparative baseline for the Bayesian segmentation strategy, and illustrate that only the Bayesian strategy is successful across different units of representation.

We replicate results showing that the subtractive segmentation strategy of Lignos (2012) does quite well over syllables. To make a fair comparison, we evaluate the Lignos2012 learner that does not use stress information. This learner operates by segmenting words that appear adjacent to already known words (i.e., subtractive segmentation), similar to infants in Bortfeld et al. (2005). In cases where there are multiple words that might be segmented, the learner uses beam search to evaluate the different options, choosing the segmentation whose constituents appear most frequently elsewhere. While the syllable-based version performs as well as the best implementation of the Bayesian strategy (Lignos2012=87.0 vs. OnlineMem bigram=86.3), the phoneme-based version noticeably suffers, having an average segmentation token F-score of 7.0. This suggests that this subtractive segmentation strategy is only useful if infants do not represent fluent speech as a sequence of segments.

In contrast, the TPminima strategy does almost as well as some of the Bayesian learners when operating over phonemes (TPminima=52.6 vs. BatchOpt unigram=55.0), but terribly when operating over syllables (TPminima=13.0). We thus replicate the syllable-based failure that Yang (2004) discovered, though we find considerably more success when the phoneme is the relevant unit, presumably because this learner is able to leverage phonotactic cues, similar to the phoneme-based learner in Blanchard et al. (2010). So, unlike the subtractive segmentation strategy of Lignos (2012), the TPminima strategy is only useful if infants *do* represent fluent speech as a sequence of segments.

4.1.3 Summary: Unit of representation

Our results suggest that a Bayesian segmentation strategy not only performs well when using syllables as the basic unit of representation, but in fact performs better than when phonemes are the basic unit. However, the fact that it performs well regardless of unit of representation sets it apart from another successful syllable-based segmentation strategy, the subtractive segmentation strategy of Lignos (2012). So, the Bayesian learning strategy is viable whether future experimental research determines that infants perceive speech syllabically or segmentally.

4.2 Evaluation Metrics

Now that we have shown that a syllable-based Bayesian learner is successful when its output is compared against adult knowledge of English segmentation, we turn to comparison against a more cognitively plausible target segmentation, which includes the "reasonable errors" discussed in section 3.3:

1. Errors that produced real words (e.g., *alright* /ɑl ɹajt/ segmented as *all* /ɑl/ and *right* /ɹajt/)

2. Errors that produced productive morphological units (e.g., segmenting off *–ing*)

3. Errors that produced function word collocations (e.g., segmenting *is that a* as *isthata*)

To account for these reasonable errors, we examined the segmentations generated by the Bayesian learners and adjust the word token precision and recall accordingly. For all reasonable errors, only the portion of the error which satisfied the reasonable error description was counted as correct. For instance, if a learner missegmented *oopsie* as *oop* and *see*, *see* would be counted as correct, adding to both precision and recall, while *oop* would still be counted as incorrect since it is not an English word. For the reasonable real word errors, we wanted to exclude

babbling (e.g. *lalalalala*) or transcription error matches (e.g. *wh*), so only real word errors that resulted in words that appeared at least 10 times in the corpus were counted as correct.

For productive morphology errors, we created a list of English syllabic morphology by hand to match against (see Appendix A). Morphological segmentations were only counted as correct if they were made in the correct location, i.e., either at the beginning or end of the word. For instance, segmenting the prefix /ɹi/ out of *redo* would be counted as a reasonable morphology error since *re–* is a legitimate English prefix. However, if *very* were segmented as /vɛ ɹi/, the /ɹi/ would not count as a reasonable morphology error, as it occurs word-finally and *–re* is not a legitimate English suffix.

For function word collocations, we created a list of function words that appeared in the corpus (see Appendix B), and matched any potential function word collocation errors against that list. To be counted as this kind of reasonable error, the entire undersegmentation needed to include only function words.

Table 3 shows the results of this reasonable error evaluation for syllable-based learners and Table 4 shows the percentage of all errors made that are of each reasonable error type.

	Unigram		Bigram	
	Gold	Adjusted	Gold	Adjusted
BatchOpt	53.1 (1.3)	55.7 (1.2)	77.1 (1.4)	80.2 (1.4)
OnlineOpt	58.8 (2.5)	60.7 (2.6)	75.1 (0.9)	78.1 (1.4)
OnlineSubOpt	63.7 (2.8)	65.8 (2.8)	77.8 (1.5)	80.4 (1.7)
OnlineMem	55.1 (0.3)	58.7 (0.7)	86.3 (1.2)	89.6 (0.7)
Lignos2012	87.0 (1.4)	91.2 (1.2)		
TPminima	13.0 (0.4)	24.3 (0.5)		

Table 3: Word token F-scores for all syllable-based learners when compared against the adult segmentation (Gold) or when adjusted to include "reasonable errors" (Adjusted). Standard deviations are shown in parentheses.

First, unsurprisingly, we see that all adjusted F-scores are higher than their gold standard equivalents because the reasonable error adjustment is a relaxation of the criteria for correctness.

	Unigram			Bigram		
	Real	Morph	Func	Real	Morph	Func
BatchOpt	0.73	0.13	4.40	4.19	0.69	6.37
OnlineOpt	2.15	0.47	3.17	6.44	0.90	4.85
OnlineSubOpt	2.59	0.45	3.38	8.77	2.08	2.87
OnlineMem	2.19	0.31	5.02	14.41	3.20	3.64
Lignos2012	19.00	3.59	0.03			
TPminima	0.01	0.00	7.33			

Table 4: Percentage of errors producing at least one real word (Real), one productive morpheme (Morph), or composed entirely of function words (Func).

In all cases, token F-scores increase by 2 to 3 points. Thus, the same general trends we observed in section 4.1.1 hold, where bigram learners significantly outperform unigram learners and the Bayesian strategy is quite successful. More interestingly, different Bayesian learner variants seem to produce different distributions of reasonable error types. Comparing unigram learners to bigram learners, we see that the unigram learners tend to produce fewer real word errors than the bigram learners, with as few as 0.73% for the unigram learners (BatchOpt) and as many as 14.41% for the bigram learners (OnlineMem). A similar pattern appears for morphology errors, with the bigram learners all producing more syllabic morphemes than their unigram equivalents. These two findings are intuitively reasonable, as the unigram model tends to group frequently occurring words together, having no other way of dealing with frequent sequences (unlike the bigram model). This means that the unigram learner is in general less likely to oversegment words and so less likely to segment a real word or a morpheme out of a larger word. Given this, we might then expect unigram learners to generate more function word collocations, since those are undersegmentations of (presumably) frequently occurring words. However, only two of the four learners show more function word collocations for their unigram variant (OnlineSubOpt, OnlineMem). This may be due to these bigram learners' relatively high rate of oversegmentation (see Table 5).

We also see a division between ideal and online learners, where online learners tend to make more real word errors, both for the unigram and bigram models (e.g., unigram BatchOpt: 0.73% vs. OnlineOpt: 2.15%, OnlineSubOpt: 2.59%, OnlineMem: 2.19%). This is especially true for the bigram OnlineMem learner which produces real words as 14.41% of its errors versus 4.19% from the bigram BatchOpt learner. The same pattern holds for morphological errors, with the online learners always producing more than their ideal counterpart. However, the reverse is generally true for function word collocations, where the ideal learner always produces more of these undersegmentation errors than either the OnlineOpt, OnlineSubOpt, and bigram OnlineMem learners. Only the unigram OnlineMem learner produces more of these errors than the unigram BatchOpt learner. In general, the trend towards more function word collocations and fewer real word and morphological errors correlates with the tendency to undersegment, while fewer function word collocations and more real word and morphological errors correlates with the tendency to oversegment. This can be seen in the over- and undersegmentation rates for each learner in Table 5.

Turning to our two baseline strategies, we see that both the Lignos2012 and TPminima learners also significantly benefit from the adjusted target segmentation criteria, though the syllable-based TPMinima's performance is still very poor. Interestingly, these two learners show very different patterns of reasonable errors. The Lignos2012 learner very rarely creates function word collocations, but often identifies real words and morphology. As with the Bayesian learners, this seems related to the tendency to oversegment (oversegmentation rate = 95.63%). In contrast, the TPminima learner essentially only makes function word collocation errors. This is due to the TPminima learner's tendency to almost always undersegment rather than oversegment (undersegmentation rate = 99.78%). In general, these learners appear to have more extreme oversegmentation (Lignos2012) or undersegmentation (TPminima) tendencies when compared against the Bayesian learners.

	Unigram		Bigram	
	Over %	Under %	Over %	Under %
BatchOpt	3.25	96.05	24.65	74.19
OnlineOpt	9.57	89.23	23.42	75.21
OnlineSubOpt	12.46	85.86	31.32	67.10
OnlineMem	16.84	80.87	60.71	36.71
Lignos2012	95.63	4.37		
TPminima	0.22	99.78		

Table 5: Undersegmentation and oversegmentation rates for all learners as a percentage of all errors made by each learner. Undersegmentations are defined as any error which conjoins at least two real words (e.g., *isthata*). Oversegmentations are defined as any error which takes a real word and split it into smaller pieces (e.g., *a fraid*).

4.3 The role of the inference process

We next investigate the effect of more cognitively plausible inference algorithms. Looking at the results in Table 3, we see that learner performance varies quite dramatically based on the specific algorithm chosen to approximate the inference process as well as the underlying generative model the learner is using. Among the unigram learners, we find that the BatchOpt learner actually performs the worst (Adjusted=55.7), highlighting that optimal inference does not yield the best performance. Instead, the OnlineSubOpt learner, one of the more constrained learners, yields the best segmentation results, at roughly 10 points higher (Adjusted=65.8). Among the bigram learners, the BatchOpt learner fares better relatively (Adjusted=80.2), on par with two of the constrained learners (OnlineOpt=78.1, OnlineSubOpt=80.4). This is likely because the bigram generative assumption, while still inaccurate, is a better generative model than the unigram assumption. So, finding the optimal segmentation that fits the bigram assumption, which the BatchOpt learner does, is a better approach than finding the optimal segmentation that fits a unigram assumption. Interestingly, however, the OnlineMem learner is far and away the best (Adjusted=89.6), around 10 points higher than all other learners. We explore the cause of this effect below in more detail, but note here that these results with constrained learners demon-

strate that more cognitively plausible inference algorithms do not generally harm segmentation performance for a syllable-based Bayesian learner – on the contrary, they often help it.

Notably, these results are in line with the previous phoneme-based modeling investigations of Pearl et al. (2011), who also found that constrained learners can perform as well as or better than the BatchOpt learner. However, this interesting behavior was limited to the unigram phoneme-based learners, which underscores how the choice of input representation affects the results. As noted above, we found this behavior in both unigram and bigram syllable-based learners, rather than only unigram learners. Pearl et al. (2011) explained this unigram behavior by noting that the unigram BatchOpt learner made many function word collocation errors (e.g. *canyou, doyou, itsa*) while the OnlineMem learner made far fewer of these errors. While this explanation seems to hold for some syllable-based learners (bigram BatchOpt=6.37% function errors, OnlineMem=3.64%), it does not explain why other learners producing fewer of these errors don't also perform better (e.g. bigram OnlineSubOpt = 2.87% function errors).

What then is the cause of the better segmentation performance for the constrained learners, whose inference process is less powerful than the ideal learner's inference process? The first thing we investigated was whether the BatchOpt learner was indeed converging on a more optimal segmentation than its constrained counterparts. It could be that some quirk of the inference algorithms causes the constrained learners to instead converge on segmentations more in line with the model's generative assumptions (e.g., the unigram or bigram assumption). To test this possibility, we calculated the log posteriors for each learner, where the log posterior represents the probability of the segmentation given the data (9).

$$log(Posterior) \propto log(Prior * Likelihood) = log(P(H) * P(D|H)) \tag{9}$$

If the inference algorithms are performing as expected, we should see that the BatchOpt

learner should have a better log posterior than any of the other learners, since it is meant to converge on the globally optimal segmentation via Gibbs sampling. We would also expect that the OnlineOpt learner should have a better log posterior than the OnlineSubOpt learner since the OnlineOpt learner always chooses the segmentation with the best log posterior. We find that both these predictions are borne out for the syllable-based learners, with the relative ranking of learners based on log posterior as BatchOpt $>$ OnlineMem $>$ OnlineOpt $>$ OnlineSubOpt. So, it is not the case that our constrained learners are finding segmentations that better correspond to the generative assumptions than the BatchOpt learner – the inference algorithms are behaving in this respect as they are supposed to. Yet they still find "sub-optimal" segmentations that are a better match to the target segmentation. We examine each constrained learner in turn.

For the OnlineOpt learner, we see from Table 3 that the unigram OnlineOpt learner outperforms the BatchOpt by roughly 5 points (58.8 vs. 53.1). In the bigram case, the OnlineOpt learner performs slightly worse, around 2 points lower than the BatchOpt learner (78.1 vs. 80.2). This implies that for a syllable-based learner, using an "optimal" online inference process provides some benefit for a unigram model, but harms performance somewhat for a bigram model (though the learner still performs quite well).

Since all online learners tend to make errors that produce real words, it is likely that not having access to later utterances is beneficial, particularly for unigram learners (Pearl et al., 2011). To understand why this is, recall that the Bayesian learner's decisions are based on the perceived frequency of the segmented items. For a unigram learner, a frequent sequence of words like *what's that* can only be explained as a single word *whatsthat*. The BatchOpt learner, which will learn that this is a frequent sequence because it is learning over all utterances at once, therefore undersegments this frequent sequence. In contrast, an online learner does not know this is a frequent sequence when it is first encountered, and so the online learner is less likely to undersegment it at that point. If the learner does not make this undersegmentation

error, the perceived frequencies of *whats* and *that* are higher – and so the online learner is less likely to undersegment *what's that* in future utterances. In this way, the online unigram learners (such as the OnlineOpt learner) have an advantage over the BatchOpt unigram learner. We therefore expect the BatchOpt unigram learner to undersegment more often than the constrained unigram learners, which is exactly what Table 5 shows. The poorer performance of the bigram OnlineOpt learner may be due to this learner's relatively high rate of undersegmentation, as it is the only online learner which undersegments more than the BatchOpt learner (bigram OnlineOpt=75.21%, BatchOpt=74.19%).

In contrast, the OnlineSubOpt learner outperforms the OnlineOpt learner for both unigram and bigram language models, yielding the best unigram segmentation result (65.8) and a bigram result equivalent to the BatchOpt learner's (OnlineSubOpt=80.4, BatchOpt=80.2). Here, the undersegmentation bias again correlates with segmentation performance: the OnlineSubOpt learner has fewer undersegmentation errors and higher F-scores than the BatchOpt learner. It may be that the noise introduced by this learner's inference algorithm (to sample the hypothesis space of segmentations rather than always choosing the locally best one) leads it to undersegment sequences less often even if it perceives those sequences as being frequent.

Turning to the OnlineMem learner, we see that it also outperforms the BatchOpt learner for both unigram and bigram language models (unigram OnlineMem=58.7 vs. BatchOpt=55.7, bigram OnlineMem=89.6 vs. BatchOpt=80.2). Most striking is the bigram performance, as noted above, which is significantly higher than all other learners. This again correlates with undersegmentation behavior, with the bigram OnlineMem learner having the weakest undersegmentation tendency of all Bayesian learners. This is likely what causes it to identify more real words and morphological units in its errors.

More generally for syllable-based learners, we find that incorporating some cognitively realistic assumptions into a learner's inference process can significantly improve segmentation

performance in some cases. Importantly, such cognitive limitations never drastically decrease segmentation performance. The Bayesian segmentation strategy therefore seems robust across different cognitively constrained approximations of "ideal" Bayesian inference, which is crucial if infants – who likely do approximate such inference – are meant to use this strategy.

5 Discussion

By incorporating more cognitively plausible assumptions into different aspects of the word segmentation modeling process, we have found a number of useful results. Using syllables as the unit of input representation, we found (i) that performance improved for the Bayesian strategy and (ii) a more robust effect of inference process, where cognitively constrained learners outperform the ideal learner. By having a more nuanced definition of segmentation success, we uncovered interesting differences in the units identified by each Bayesian learner, based on how the inference process was constrained. We discuss the impact of each of these results in turn.

5.1 Unit of representation

A longstanding question in developmental speech perception has been the nature of infant representation of the speech stream. While no strong conclusions can be drawn from the current state of the literature, our results suggest that syllables may be a better unit of representation for models of early word segmentation, in line with prior syllabic word segmentation models (Swingley, 2005; Gambell & Yang, 2006; Lignos & Yang, 2010; Lignos, 2012). Interestingly, for the Bayesian segmentation strategy we examined, syllables are not the only viable unit of representation – phonemes will do as well. This contrasts with other segmentation strategies that require the unit to be the syllable (Lignos2012) or the phoneme (TPminima). Therefore, this Bayesian strategy appears resilient across both potential units of representation, and so can

continue as a viable learning strategy for early word segmentation, regardless of whether future research definitively rules out one or the other. Still, as other units of representation are possible (as well as more realistic instantiations of the syllabic unit we use here), a useful question for future research is how the Bayesian strategy fares across other potential units. More broadly related to this, computational modelers can provide empirical evidence about infant input representation by demonstrating which representational units coupled with which learning strategies yield acquisition success. These can then serve as specific proposals about the acquisition process that can be experimentally investigated.

5.2 Inference Process

Since we find (as Pearl et al., 2011 did) beneficial impacts of cognitive constraints on the Bayesian learning strategy, it is unlikely to be a fluke that is specific to a phoneme- or syllable-based Bayesian learner. Therefore, it is worth thinking more broadly about why the specific cognitive constraints we incorporated into the inference process (as represented by the OnlineOpt, OnlineSubOpt, and OnlineMem learners) had such a helpful effect.

It is sometimes the case that online learning algorithms perform inference better than batch algorithms (Liang & Klein, 2009), but that does not seem true here, as the log posterior analysis clearly indicates that the ideal learner produces a "better" segmentation with respect to the underlying generative assumption (unigram or brigam). Notably, either assumption is a very inaccurate representation for how words are actually generated by adult speakers (though perhaps reasonable as a naive theory a six-month-old might have). Therefore, we believe the main problem is that the ideal learner follows its inaccurate generative assumption too closely. Approximated inference that yields a poorer match to this underlying assumption should therefore yield better segmentation results – and this is exactly what we find: for example, the unigram OnlineSubOpt learner achieves the best segmentation performance but has the worst fit to the

underlying unigram generative model.

This highlights an interesting interaction between the correctness of the underlying generative assumptions and the fidelity of the inference process. If the underlying generative assumption is rather inaccurate, it may be better to have a less accurate inference process precisely because it pushes the learner away from that inaccurate assumption. Thus, there is a potential synergy between the naive theories of language young infants may possess and the cognitive constraints that limit the power of their inference processes. Interestingly, this pattern of results fits the general intuition of the "Less is More" hypothesis (Newport, 1990), which posits that children's cognitive limitations are the key to what makes them such efficient language learners. Our results suggest that cognitive limitations are helpful for word segmentation specifically when the learner's underlying theory of what language should look like is inaccurate. Whether this is true about cognitive limitations for other acquisition tasks remains to be explored, but seems an intriguing avenue of research given the plethora of Bayesian learning strategies currently shown to be successful using ideal learners (e.g., phonetic category learning (Feldman et al., 2013), simultaneous phone & phonological rule learning: (Dillon et al., 2013), word-meaning learning (Frank et al., 2009), grammatical categorization (Christodoulopoulos et al., 2011), and hierarchical structure identification (Perfors, Tenenbaum, & Regier, 2011)).

5.3 Evaluation Measures

An additional concern for models of language acquisition is how to evaluate a model's performance. In fact, this is generally a problem for evaluating unsupervised learning approaches in machine learning (von Luxburg, Williamson, & Guyon, 2011). For language acquisition, it is common practice to compare against adult knowledge, as represented by a "gold standard" of some kind. However, this presents two potential problems.

First, the gold standard may not be an accurate representation of adult knowledge. For

word segmentation, we might wonder how closely orthographic segmentations match the actual adult cognitive representation of the core language units (e.g., see Blanchard et al., 2010 for discussion of orthographic vs. phonological vs. grammatical words).

Second, we may not expect children to reach adult competence using a given strategy. For early word segmentation, it is known that statistical cues are not the only ones utilized by infants (Jusczyk et al., 1993; Jusczyk, Houston, & Newsome, 1999; E. Johnson & Jusczyk, 2001; Thiessen & Saffran, 2003; Bortfeld et al., 2005). It would therefore be surprising if statistical word segmentation were capable of producing adult-like segmentations given that its true objective is likely to create a proto-lexicon which later segmentation strategies may be bootstrapped from.

In this paper, we focused on addressing the second concern by considering different kinds of units as reasonable outputs of early segmentation. In particular, we allowed three kinds of "reasonable errors", two which were produced by oversegmentation (real words, productive morphology) and one which was produced by undersegmentation (function word collocations). This led to what we feel is a more fair evaluation of a segmentation strategy, and improved segmentation performance for all learners. Notably, qualitatively different patterns emerged, depending on a particular learner's tendency to undersegment or oversegment, with constrained learners oversegmenting more than the ideal learner. This suggests that very young infants may make more oversegmentation "errors" in English (such as identifying productive morphology) if they are using the Bayesian strategy described here. Some experimental evidence suggests that 7.5-month-old infants have indeed keyed into productive morphology and can use that morphology as a segmentation cue (Willits, Seidenberg, & Saffran, 2009). One way they might have learned these useful morphological units is by using the Bayesian segmentation strategy described here with some version of constrained inference.

More generally, this evaluation approach can allow us to assess the utility of each learner's

segmentations, as one reasonable goal for unsupervised models is the utility of the output they generate. In particular, are the segmentations produced useful for future acquisition stages? For example, since one goal is to generate a proto-lexicon from which to bootstrap language-specific segmentation cues like stress pattern (Swingley, 2005), does the proto-lexicon produced yield the correct language-specific cue (in English, stress-initial)? As another example, is the proto-lexicon produced good enough to yield useful word-object mappings (learning word meaning) or clusters of similar-behaving words (learning grammatical categories)?

A related evaluation approach is to assess the utility of a strategy cross-linguistically – that is, is it useful for as many languages as we can test it on, or does it only work if certain properties are true of a language? For early word segmentation strategies, the goal is certainly for the basic strategy to succeed on any language. For acquisition tasks that involve language-specific knowledge (e.g., later segmentation strategies involving language-specific cues), it may be that the different variants of a strategy succeed, depending on the language properties. Still, we typically look for a language acquisition strategy that can succeed for any language, under the assumption that core aspects of the language acquisition process are universal. Thus, evaluating any proposed strategy on multiple languages should be encouraged.

6 Conclusion

This study highlights the benefits of using experimental research to inform decisions about modeling language acquisition, with the goal of creating more informative models. We can incorporate cognitive plausibility at both the computational and algorithmic levels of model design by considering the most plausible unit of input representation, evaluating model output against what is more likely to be a human learner's target knowledge, and modifying the inference process to include cognitively-inspired limitations. We demonstrated this approach

for a promising Bayesian word segmentation strategy, and made several important discoveries. First, this strategy can succeed with a more cognitively plausible unit of input representation. Second, this strategy in fact performs better when using the more cognitively plausible unit. Third, a more nuanced perspective on what good segmentation is reveals several "reasonable errors" that a learner using this strategy may make, which we can then use experimental methods to look for in infants. Fourth, a learner with a naive theory of language performs better when inference is not optimal, which underscores one reason why infants may be able to succeed at early acquisition tasks even when their cognitive resources are so limited. By making a firmer commitment to cognitive plausibility in computational modeling simulations, we can create models with more explanatory power that can both make empirically testable predictions and impact our theories about how language acquisition occurs.

References

Abbott, J., Hamrick, J., & Griffiths, T. (2013). Approximating bayesian inference with a sparse distributed memory system. In *Proceedings of the 35th annual conference of the cognitive science society* (pp. 1686–1691).

Batchelder, E. (2002). Bootstrapping the lexicon: A computational model of infant speech segmentation. *Cognition*, *83*(2), 167–206.

Bertonicini, J., Bijeljac-Babic, R., Jusczyk, P., Kennedy, L., & Mehler, J. (1988). An investigation of young infants' perceptual representations of speech sounds. *Journal of Experimental Psychology*, *117*(1), 21–33.

Best, C., McRoberts, G., LaFleur, R., & Silver-Isenstadt, J. (1995). Divergent developmental patterns for infants' perception of two nonnative consonant contrasts. *Infant Behavior and Development*, *18*, 339–350.

Best, C., McRoberts, G., & Sithole, N. (1988). Examination of perceptual reorganization for nonnative speech contrasts: Zulu click discrimination by english-speaking adults and infants. *Journal of Experimental Psychology: Human Perception and Performance*, *14*(3), 345–360.

Bijeljac-Babic, R., Bertoncini, J., & Mehler, J. (1993). How do 4-day-old infants categorize multisyllabic utterances? *Developmental Psychology*, *29*(4), 711–721.

Blanchard, D., Heinz, J., & Golinkoff, R. (2010). Modeling the contribution of phonotactic cues to the problem of word segmentation. *Journal of child language*, *37*, 487–511.

Bortfeld, H., Morgan, J., Golinkoff, R., & Rathbun, K. (2005). Mommy and me: Familiar names help launch babies into speech-stream segmentation. *Psychological Science*, *16*(4), 298–304.

Box, G., & Draper, N. (1987). *Empirical model building and response surfaces*. New York,

NY: John Wiley & Sons.

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

Brent, M., & Cartwright, T. (1996). Distributional regularity and phonotactic constraints are useful for segmentation. *Cognition*, *61*, 93–125.

Brent, M., & Siskind, J. (2001). The role of exposure to isolated words in early vocabulary. *Cognition*, *81*, 31–44.

Brown, R. (1973). *A first language: The early stages*. Harvard University Press.

Christodoulopoulos, C., Goldwater, S., & Steedman, M. (2011). A bayesian mixture model for part-of-speech induction using multiple features. In *Proceedings of the conference on empirical methods in natural language processing.*

Denison, S., Bonawitz, E., Gopnik, A., & Griffiths, T. (2013). Rational variability in children's causal inferences: The sampling hypothesis. *Cognition*, *126*, 285–300.

Dillon, B., Dunbar, E., & Idsardi, W. (2013). A single-stage approach to learning phonological categories: Insights from inuktitut. *Cognitive Science*, *37*(2), 344–377.

Eimas, P. (1997). Infant speech perception: Processing characteristics, representational units, and the learning of words. In R. Goldstone, P. Schyns, & D. Medin (Eds.), *The psychology of learning and motivation* (Vol. 36). San Diego: Academic Press.

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

Feldman, N., Griffiths, T., Goldwater, S., & Morgan, J. (2013). A role for the developing lexicon in phonetic category acquisition. *Psychological Review*, *120*(4), 751–778.

Feldman, N., Griffiths, T., & Morgan, J. (2009). Learning phonetic categories by learning a lexicon. In *Proceedings of the 31st annual conference of the cognitive science society* (pp. 2208–2213).

Ferguson, C. (1964). Baby talk in six languages. *American Anthropologist*, *66*, 103–113.

Ferguson, T. (1973). A bayesian analysis of some nonparametric problems. *Annals of Statistics*, *1*(2), 209–230.

Fernald, A., Taeschner, T., Dunn, J., Papousek, M., de Boysson-Bardies, B., & Fukui, I. (1989). A cross-language study of prosodic modifications in mothers' and fathers' speech to preverbal infants. *Journal of Child Language*, *16*(3), 477–501.

Fleck, M. (2008). Lexicalized phonotactic word segmentation. In *Association for computational linguistics.*

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

Freudenthal, D., Pine, J., & Gobet, F. (2006). Modeling the development of children's use of optional infinitives in dutch and english using mosaic. *Cognitive Science*, *30*(2), 277-310.

Gambell, T., & Yang, C. (2006). *Word segmentation: Quick but not dirty.*

Geman, S., & Geman, D. (1984). Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, *6*, 721–741.

Goldwater, S., Griffiths, T., & Johnson, M. (2009). A bayesian framework for word segmentation. *Cognition*, *112*(1), 21–54.

Grieser, D., & Kuhl, P. (1988). Maternal speech to infants in a tonal language: Support for universal prosodic features in motherese. *Developmental Psychology*, *24*(1), 14–20.

Hall, T. (1992). *Syllable structure and syllable-related processes in german* (Vol. 276). Walter de Gruyter.

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. (2008). Using adaptor grammars to identify synergies in the unsupervised acqui-

sition of linguistic structure. In *Acl* (pp. 398–406).

Johnson, M., Griffiths, T., & Goldwater, S. (2007). Adaptor grammars: A framework for specifying compositional nonparametric bayesian models. *Advances in Neural Information Processing Systems*, *19*, 641–648.

Jusczyk, P. (1997). *The discovery of spoken language*. Cambridge, MA: MIT Press.

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

Jusczyk, P., Friederici, A., Wessels, J., Svenkerud, V., & Jusczyk, A. (1993). Infants' sensitivity to the sound patterns of native language words. *Journal of Memory and Language*, *32*, 402–420.

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

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

Jusczyk, P., Jusczyk, A., Kennedy, L., Schomberg, T., & Koenig, N. (1995). Young infants' retention of information about bisyllabic utterances. *Journal of Experimental Psychology: Human Perception and Performance*, *21*(4), 822–836.

Kol, S., Nir, B., & Wintner, S. (2014). Computational evaluation of the traceback method. *Journal of Child Language*, *41*(1), 176–199.

Kuhl, P., Williams, K., Lacerda, F., Stevens, K., & Lindblom, B. (1992). Linguistic experience alters phonetic perception in infants by 6 months of age. *Science*, *255*, 606–608.

Labov, W. (1997). Resyllabification. In F. Hinskens & R. van Hout (Eds.), *Variation, change, and phonological theory.* John Benjamins Publishing.

Legate, J., & Yang, C. (2007). Morphosyntactic learning and the development of tense. , *14*(3), 315-344.

Liang, P., & Klein, D. (2009). Online em for unsupervised models. In *Human language technologies: The 2009 annual conference of the north american chapter of the acl* (pp. 611–619).

Lignos, C. (2012). Infant word segmentation: An incremental, integrated model. In *Proceedings of the 30th west coast conference on formal linguistics.*

Lignos, C., & Yang, C. (2010). Recession segmentation: Simpler online word segmentation using limited resources. In *Proceedings of the fourteenth conference on computational natural language learning* (pp. 88–97).

MacWhinney, B. (2000). *The childes project: Tools for analyzing talk* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. New York, N.Y.: Henry Holt and Co. Inc.

Marthi, B., Pasula, H., Russell, S., & Peres, Y. (2002). Decayed mcmc filtering. In *Proceedings of 18th uai* (p. 319-326).

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

Maye, J., Weiss, D., & Aslin, R. (2008). Statistical phonetic learning in infants: facilitation and feature generalization. *Developmental Science*, *11*(1), 122-134.

Maye, J., Werker, J., & Gerken, L. (2002). Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition*, *82*, B101–B111.

Mehler, J., Dupoux, E., & Segui, J. (1990). Constraining models of lexical access: The onset of word recognition. In G. Altman (Ed.), *Cognitive models of speech processing: Psycholinguistic and computational perspectives.* Cambridge, MA: MIT Press.

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

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

Pearl, L. (in press). Evaluating learning strategy components: Being fair. *Language*.

Pearl, L., Goldwater, S., & Steyvers, M. (2011). Online learning mechanisms for bayesian models of word segmentation. *Research on Language and Computation*, *8*(2), 107–132. (special issue on computational models of language acquisition)

Pegg, J., & Werker, J. (1997). Adult and infant perception of two english phones. *Journal of the Acoustical Society of America*, *102*(6), 3742–3753.

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

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

Perfors, A., Tenenbaum, J., & Regier, T. (2011). The learnability of abstract syntactic principles. *Cognition*, *118*, 306–338.

Peters, A. (1983). *The units of language acquisition*. New York: Cambridge University Press.

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

Saffran, J., Aslin, R., & Newport, E. (1996). Statistical learning by 8-month-old infants. *Science*, *274*, 1926–1928.

Sanborn, A., Griffiths, T., & Navarro, D. (2010). Rational approximations to rational models: Alternative algorithms for category learning. *Psychological Review*, *117*(4), 1144–1167.

Seidl, A., Cristià, A., Bernard, A., & Onishii, K. (2009). Allophonic and phonemic contrasts in infants' learning of sound patterns. *Language Learning and Development*, *5*, 191–202.

Selkirk, E. (1981). English compounding and the theory of word structure. In M. Moortgat,

H. van der Hulst, & T. Hoestra (Eds.), *The scope of lexical rules.* Dordrecht: Foris.

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

Snow, C. (1977). The development of conversation between mothers and babies. *Journal of Child Language*, *4*(1), 1–22.

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

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

Thiessen, E., & Saffran, J. (2003). When cues collide: Use of stress and statistical cues to word boundaries by 7- to 9-month-old infants. *Developmental Psychology*, *39*(4), 706–716.

Toscano, J., & McMurrary, B. (2010). Cue integration with categories: Weighting acoustic cues in speech using unsupervised learning and distributional statistics. *Cognitive Science*, *34*, 434–464.

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

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

von Luxburg, U., Williamson, R., & Guyon, I. (2011). Clustering: Science or art? In *Jmlr workshop and conference proceedings 27* (pp. 65–79). (Workshop on Unsupervised Learning and Transfer Learning)

Wang, H., & Mintz, T. (2008). A dynamic learning model for categorizing words using frames. In *Proceedings of bucld* (Vol. 32, pp. 525–536).

Werker, J., & Lalonde, C. (1988). Cross-language speech perception: Initial capabilities and

developmental change. *Developmental Psychology*, *24*(5), 672–683.

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

Willits, J., Seidenberg, M., & Saffran, J. (2009). Verbs are looking good in language acquisition. In *Proceedings of the 31st annual conference of the cognitive science society* (pp. 2570–2575).

Wilson, M. (1988). The mrc psycholinguistic database machine readable dictionary. *Behavioral Research Methods, Instruments and Computers*, *20*, 6–11.

Yang, C. (2004). Universal grammar, statistics or both? *Trends in Cognitive Sciences*, *8*(10), 451–456.

Appendices

A Productive syllabic morphology

Prefixes	*be-, un-, non-, re-*
Suffixes	*-ing, -full, -ness, -er, -est, -ly, -ity, -es, -ble, -wise*

Table 6: Productive syllabic morphology identified as reasonable errors.

B Function words

a, about, across, again, against, ah, ahead, ahoy, all, also, although, always, am, among, an, another, any, anyhow, anyone, anyway, anywhere, are, around, as, at, bah, be, because, before, beneath, beside, besides, but, by, can, cause, contrary, could, damn, down, during, either, else, even, every, everyone, everywhere, except, few, for, forward, from, further, gee, goodbye, gosh, had, he, her, hers, herself, hi, his, hooray, hurray, I, in, indeed, inside, instead, into, is, it, itself, it's, its, less, many, me, meanwhile, might, more, must, my, myself, near, never, no, nobody, none, nope, nor, north, nothing, o'clock, of, off, ok, okay, on, onto, opposite, or, other, otherwise, ouch, ought, our, ourself, out, outside, own, per, please, probably, right, second, self, several, she, should, some, somebody, somehow, someplace, something, sometimes, somewhere, south, ten, that, the, theirs, them, themselves, there, there's, they, third, this, though, to, too, today, together, tomorrow, tonight, towards, two, um

Table 7: Function words identified in collocations as reasonable errors.