Evaluating language acquisition models: A utility-based look at Bayesian segmentation

Lisa Pearl (lpearl@uci.edu)
Lawrence Phillips (lawphill@uci.edu)
Department of Cognitive Sciences
University of California, Irvine

Abstract

Computational models of language acquisition often face evaluation issues associated with unsupervised machine learning approaches. These acquisition models are typically meant to capture how children solve language acquisition tasks without relying on explicit feedback, making them similar to other unsupervised learning models. Evaluation issues include uncertainty about the exact form of the target linguistic knowledge, which is exacerbated by a lack of empirical evidence about children's knowledge at different stages of development. Put simply, a model's output may be good enough even if it doesn't match adult knowledge because children's output at various stages of development *also* may not match adult knowledge. However, it's not easy to determine what counts as "good enough" model output. We consider this problem using the case study of speech segmentation modeling, where the acquisition task is to segment a fluent stream of speech into useful units like words. We focus on a particular Bayesian segmentation strategy previously shown to perform well on English, and discuss several options for assessing whether a segmentation model's output is good enough, including cross-linguistic utility, the presence of reasonable errors, and downstream evaluation. Our findings highlight the utility of considering multiple metrics for segmentation success, which is likely also true for language acquisition modeling more generally.

1 Introduction

A core issue in machine learning is how to evaluate unsupervised learning approaches (von Luxburg, Williamson, & Guyon, 2011), since there is no *a priori* correct answer the way that there is for supervised learning approaches. Computational models of language acquisition commonly face this problem because they attempt to capture how children solve language acquisition tasks without explicit feedback, and so typically use unsupervised learning approaches. Moreover, evaluation is made more difficult by uncertainty about the exact nature of the target linguistic knowledge and a lack of empirical evidence about children's knowledge at specific stages in development. Given this, how do we know that a model's output is "good enough"? How should success be measured? To create informative cognitive models of acquisition that offer insight into how children acquire language, we should consider how to evaluate acquisition models appropriately (Pearl, 2014; Phillips, 2015; Phillips & Pearl, 2015b).

As a case study, we investigate the initial stages of speech segmentation in infants, where a fluent stream of speech is divided into useful units, such as words. For example, the acoustic signal

transcribed via IPA as /ajlʌvðizpɛŋgwɪnz/ (*Ilovethesepenguins*) might be segmented as /aj lʌv ðiz pɛngwɪnz/ (*I love these penguins*). A particular Bayesian segmentation strategy has been shown to be quite successful on English (Goldwater, Griffiths, & Johnson, 2009; Pearl, Goldwater, & Steyvers, 2011; Phillips & Pearl, 2012, 2014a, 2014b; Phillips, 2015; Phillips & Pearl, 2015b), particularly when cognitive plausibility considerations have been incorporated at both the computational and algorithmic levels of Marr (1982). One way to evaluate if this strategy is "good enough" is to see how it fares cross-linguistically. This is based on the premise that core aspects of the language acquisition process – such as the early stages of segmentation occurring in six- to seven-month-olds (Thiessen & Saffran, 2003; Bortfeld, Morgan, Golinkoff, & Rathbun, 2005) – are universal. So, a viable learning strategy for early segmentation should succeed on any language infants encounter.

Traditionally, a segmentation model's output has been compared against a "gold standard" derived from adult orthographic segmentation (e.g., Brent, 1999; M. Johnson, 2008; Goldwater et al., 2009; Blanchard, Heinz, & Golinkoff, 2010; M. Johnson, Demuth, Jones, & Black, 2010; M. Johnson & Demuth, 2010; Pearl et al., 2011; Lignos, 2012; Fourtassi, Börschinger, Johnson, & Dupoux, 2013). Notably, orthographic segmentation assumes the desired units are orthographic words. However, if we look at the world's languages, it becomes clear that it's also useful to identify morphemes – the smallest meaningful linguistic units – particularly for languages with regular morphology. That is, infants might reasonably segment morphemes from fluent speech rather than entire words. Notably, models that identify sub-word morphology are penalized by the gold standard evaluation, and this highlights the need for a more flexible metric of segmentation performance. More generally, a segmentation strategy that identifies units useful for later linguistic analysis shouldn't be penalized.

Still, how do we know that the segmented units are truly useful? If we believe that the output of early segmentation scaffolds later acquisition processes, a useful segmentation output should simply enable these later processes to successfully occur (Phillips & Pearl, 2015a). For example, one goal of early segmentation is to generate a proto-lexicon in order to bootstrap language-specific segmentation cues like stress pattern (e.g., in English, words in child-directed speech tend to begin with stress (Swingley, 2005), and the same is true for child-directed German and Hungarian (Phillips & Pearl, 2015a)). Does the inferred proto-lexicon of units yield the appropriate language-specific cue? As another example, infants begin to learn mappings from word forms to familiar objects as early as six months (Tincoff & Jusczyk, 1999; Bergelson & Swingley, 2012; Tincoff & Jusczyk, 2012). Can the inferred proto-lexicon be used successfully for this process?

In the remainder of this chapter, we first review relevant aspects of infant speech segmentation, including what is known about the developmental trajectory, the cues infants are sensitive to, and how infants perceive the input. This forms the empirical basis for the modeled Bayesian segmentation strategy, which we then discuss in terms of its underlying generative assumptions and the algorithms used to carry out its inference. We turn then to the cross-linguistic evaluation of this strategy over input derived from child-directed speech corpora in seven languages from the CHILDES database (MacWhinney, 2000): English, German, Spanish, Italian, Farsi, Hungarian, and Japanese. This section includes comparison to the gold standard as well as a more flexible metric derived from the gold standard that's consistent with children's imperfect early segmentation behavior. We find that the Bayesian strategy seems to be "good enough" cross-linguistically. This is especially true once

we use this more nuanced output evaluation that considers potentially useful non-word units valid. This serves as a general methodological contribution about the definition of segmentation success, especially when we consider that useful units may vary across the world's languages.

We conclude with an evaluation metric that is even more utility-driven: whether the output of the segmentation strategy is helpful for subsequent acquisition processes, such as inferring a language-specific stress-based segmentation cue and learning early word-meaning mappings. Interestingly, just because a strategy yields more accurate segmentations when compared against the gold standard doesn't mean it's always more useful for subsequent acquisition processes. This underscores the value of considering multiple metrics for segmentation success, in addition to the traditional comparison against the gold standard.

2 Early speech segmentation

Segmentation isn't easy – words blur against one another, making speech more like a stream of sound rather than something divided into discrete, separable chunks (Cole & Jakimik, 1980). Yet, speech segmentation is one of the first tasks infants accomplish in their native language, and the resulting segmented units underlie subsequent processes such as learning word meanings, syntactic categories, and syntactic structure. In order to accurately model the segmentation process, we need to know the empirical data which form the basis for decisions regarding the model's learning assumptions, input, inference, and evaluation.

2.1 When does early speech segmentation begin?

The first behavioral evidence for speech segmentation in infants comes at six months (Bortfeld et al., 2005), when infants seem to know a small set of very frequent words (Bergelson & Swingley, 2012). They recognize these words in speech, and can use them to segment new utterances. Between seven and nine months, infants learn to utilize language-specific cues such as stress pattern (Jusczyk, Cutler, & Redanz, 1993; Jusczyk, Houston, & Newsome, 1999; Thiessen & Saffran, 2003), phonotactics (Mattys, Jusczyk, & Luce, 1999), allophonic variation (Hohne & Jusczyk, 1994; Jusczyk, Hohne, & Baumann, 1999), and coarticulation (E. Johnson & Jusczyk, 2001). These language-specific cues are typically more reliable than language-independent cues like transitional probability between syllables. However, it turns out that infants around seven months old prefer to rely on transitional probability information alone rather than language-specific cues like stress patterns (Thiessen & Saffran, 2003), even though transitional probability is a less reliable cue. Neuroimaging evidence from neonates suggests that this sensitivity to statistical cues like transitional probabilities is present at birth (Teinonen, Fellman, Näätänen, Alku, & Huotilainen, 2009). This in turn suggests that the initial stages of speech segmentation rely on cues that are independent of the particular language being segmented (e.g., the process of tracking transitional probabilities does not vary from language to language, though the probabilities themselves clearly do). It's only after this initial stage that infants harness the more powerful language-dependent cues that do vary between languages (e.g., the specific stress-based seg-

mentation cues which differ between English and French). So, a model of early speech segmentation should also likely rely only on language-independent cues.

2.2 The unit of infant speech perception

The basic unit of infant speech perception has been a source of significant debate for some time (see Phillips (2015) and Jusczyk (1997) for a more detailed review of this debate). Experimental studies have typically focused on whether the basic representational unit for infants is syllabic or segmental (Jusczyk & Derrah, 1987; Bertonicini, Bijeljac-Babic, Jusczyk, Kennedy, & Mehler, 1988; Bijeljac-Babic, Bertoncini, & Mehler, 1993; Jusczyk, Jusczyk, Kennedy, Schomberg, & Koenig, 1995; Eimas, 1999). Jusczyk (1997) summarizes several studies by noting that "there is no indication that infants under six months of age represent utterances as strings of phonetic segments". Instead, evidence for segmental representations of speech is mostly present in infants older than six months: infants first begin to ignore vowel contrasts that aren't relevant for their native language around six months (Kuhl, Williams, Lacerda, Stevens, & Lindblom, 1992; Polka & Werker, 1994), while irrelevant consonant contrasts are ignored between eight and twelve months (Werker & Tees, 1984; Werker & Lalonde, 1988; Best, McRoberts, & Sithole, 1988; Best, McRoberts, LaFleur, & Silver-Isenstadt, 1995).

More generally, as infants get older, they're better able to represent information about segments; in contrast, they appear relatively comfortable with syllables from early on. This can be seen in infants' ability to track statistical relationships: while transitional probabilities over syllables can be tracked at birth (Teinonen et al., 2009), transitional probabilities over segments first seems to occur around nine months (Mattys et al., 1999). Given this, a reasonable assumption for a model meant to capture segmentation strategies being used by six-month-olds would be that the input is perceived as a stream of syllables.

2.3 Constraints on infant inference

One thing that makes child language acquisition so impressive is that it's accomplished despite the many cognitive constraints imposed by the developing brain. Though there's been increasing interest in cognitively-constrained language acquisition models (e.g., Anderson, 1990; Shi, Griffiths, Feldman, & Sanborn, 2010; Bonawitz, Denison, Chen, Gopnik, & Griffiths, 2011; Pearl et al., 2011; Phillips & Pearl, 2015b), there isn't very much experimental evidence to suggest exactly what kinds of constraints should be imposed. There are many possibilities, but we focus on three that have been incorporated into past acquisition models and which seem reasonable starting points: online processing, non-optimal decision-making, and recency effects.

Online processing refers to the idea that data are processed as they are encountered, rather than being stored in explicit detail for later batch processing. It's generally accepted that this is a reasonable constraint for human language processing, and commonly used as justification that a model operates at the algorithmic level in the sense of Marr (1982), rather than being idealized (e.g., Lignos & Yang, 2010; Pearl et al., 2011; Lignos, 2012; Phillips & Pearl, 2014b, 2014a, 2015b). So, this is likely reasonable to incorporate into infant inference.

For decision-making, experimental evidence from infants and children suggest that they don't always choose the highest probability option available, which would be considered the optimal decision (Köpcke, 1998; C. H. Kam & Newport, 2005; C. L. H. Kam & Newport, 2009; Davis, Newport, & Aslin, 2011; Denison, Bonawitz, Gopnik, & Griffiths, 2013). Instead, children sometimes appear to probabilistically sample the available options (Davis et al., 2011; Denison et al., 2013). Other times, they appear to simply generalize to a single option, which might in fact be a lower probability option (Köpcke, 1998; C. H. Kam & Newport, 2005; C. L. H. Kam & Newport, 2009). This suggests that infant inference may involve non-optimal decision-making.

With respect to memory constraints, experimental evidence suggests that a recency bias exists in infants (Cornell & Bergstrom, 1983; Gulya, Rovee-Collier, Galluccio, & Wilk, 1998; Rose, Feldman, & Jankowski, 2001), where the most recently encountered data have privileged status. So, this is also reasonable to incorporate into infant inference.

2.4 What we know about segmentation output

As mentioned before, one empirical checkpoint for segmentation is that a strategy ought to be successful for any human language. Beyond that, we also have some evidence about the kinds of errors children produce – this underscores that successful early segmentation doesn't necessarily mean adult-like segmentation. For example, Brown (1973) and Peters (1983) find that even three-year-old children still produce missegmentations. These errors can be broadly split into two types: function word undersegmentations (e.g. *that'sa*, *it'sa*) and function word oversegmentations (e.g. *a nother*, *be have*). So, segmentation error patterns may provide a useful qualitative benchmark for model output, and have been previously used this way (Lignos, 2012; Phillips & Pearl, 2012, 2015b).

3 A Bayesian segmentation strategy

Bayesian segmentation strategies combine the prior probability of a potential segmentation s for an utterance u with the likelihood in order to generate the posterior probability of s ($P(s|u)$) using Bayes' rule, as shown in (1).

$$P(s|u) \propto P(s)P(u|s) \tag{1}$$

The Bayesian strategy we investigate builds off of two fundamental insights about the infant's inferred proto-lexicon, both of which were used in an earlier segmentation strategy by Brent (1999). First, frequent words should be preferred over infrequent words. Second, shorter words should be preferred over longer words. These parsimony biases were incorporated by Goldwater et al. (2009) into a Bayesian strategy that used a Dirichlet Process (Ferguson, 1973) to determine the prior probability of a segmentation.

The Dirichlet Process (DP) is a non-parametric stochastic process resulting in a probability distribution often used in Bayesian modeling as a prior because it has properties well-suited to language modeling. First, because it's non-parametric, it doesn't need to pre-specify the number of items (e.g.,

word types) which might be encountered. Second, the DP facilitates "rich-get-richer" behavior, where frequent items are more likely to be encountered later. This is useful because word frequencies in natural languages tend to follow a power-law distribution which the DP naturally reproduces due to this behavior (Goldwater, Griffiths, & Johnson, 2011).

Goldwater et al. (2009) implemented two versions of the DP segmentation strategy that differed in their generative assumptions. Both versions use the likelihood function, i.e., the probability of an utterance given its segmentation $P(u|s)$, to simply rule out potential segmentations that don't match the observed utterance. For example, a possible segmentation /ðə pɛŋgwɪn/ (*the penguin*) doesn't match an observed speech stream /ðəkɪri/ (*thekitty*) when the possible segmentation is concatenated (*thepenguin≠thekitty*). So, this possible segmentation would have a likelihood of 0. In contrast, the possible segmentation /ðəki ri/ (*theki tty*) does match (*thekitty=thekitty*), and so would have a likelihood of 1. Where the DP strategy versions differ is how the prior probability for a segmentation is determined. We describe each version in turn before reviewing the inference algorithms paired with each generative model.

3.1 DP segmentation: Unigram assumption

The first version of the DP segmentation strategy uses a unigram language model (DP-Uni), with the modeled learner naively assuming that each word is chosen independently of the words around it. The prior of the potential segmentation is calculated using this generative assumption. To do this, the model must define the probability of every word in the utterance, and so a DP-Uni learner assumes that for any utterance, each segmented word w_i is generated by the following process:

1. If w_i is not a novel lexical item, choose an existing form ℓ for w_i.

2. If w_i is a novel lexical item, generate a form (e.g., the syllables $x_1 \ldots x_M$) for w_i.

Because the model doesn't know whether w_i is novel, it has to consider both options when calculating the probability of the word. We note that deciding whether w_i is novel is not the same as deciding whether the form of w_i has been previously encountered. As an example, consider the first time the modeled learner encounters the sequence /et/ (such as in the word *ate*). Because $count_{/et/} = 0$, the word must be novel ($count_{/et/}$ then = 1). Now, suppose the same sequence is encountered again, but from the word *eight*. The learner must decide if this sound sequence is a second instance of the /et/ it saw before (*ate*) or instead the first instance of a novel /et/ type (such as in the word *eight*). In the first case, the count might be updated to $count_{/et/} = 2$; in the second case, the counts might be updated to $count_{/et/} = 1$ and $count_{/et/_2} = 1$. This particular aspect of the DP distribution naturally allows for the existence of homophones (e.g., *ate*/*eight*) without requiring any additional machinery.

Returning to the DP generation process, generating either non-novel or novel items is fundamental to the DP. In a classic DP, the probability of generating a non-novel item is proportional to the number of times that item has been previously encountered. This is shown in (2), where n_ℓ refers to the number of times lexicon item ℓ has been seen in the set of words previously encountered, denoted as w_{-i}. In the denominator, i represents the total number of words encountered thus far, including the word

previously under consideration. Because the current word is not included in n_ℓ, 1 is subtracted from it.

$$P(w_i = \ell, w_i \neq novel | w_{-i}) = \frac{n_\ell}{i - 1 + \alpha} \tag{2}$$

Equation (2) gives higher probability to word types that have been encountered before. So, the more a word type is encountered by the modeled learner, the more often the modeled learner will prefer it in the future. This will end up generating the power-law frequency distribution found in natural languages.

When the DP instead generates a novel word, the word is not represented as an atomic whole but rather constructed from its individual parts, such as syllables. To model this, we make use of the P_0 in (3) to describe the probability that any word might be made up of a particular string of sub-word units $x_1 \ldots x_M$. Each sub-word unit x_j is generated in turn for all M units in the word, with the probability of x_j treated as a uniform choice over all possible sub-word units in the corpus.[1]

$$P_0(w_i = x_1 \ldots x_M) \propto \prod_{j=1}^{M} P(x_j) \tag{3}$$

The probability of generating a novel item in a DP is weighted by the free model parameter α, also known as the DP concentration parameter. This parameter has an intuitive interpretation, where higher values of α lead to a preference for generating novel words in the proto-lexicon.[2] The full probability of generating a novel word is therefore described by equation (4).

$$P(w_i = \ell, w_i = novel | w_{-i}) = \frac{\alpha P_0(w_i = x_1 \ldots x_M)}{i - 1 + \alpha} \tag{4}$$

Both equations 2 and 4 can be combined to generate the full probability of a word being produced either as a non-novel or novel item.

$$P(w_i = \ell | w_{-i}) = \frac{n_\ell + \alpha P_0(w_i = x_1 \ldots x_M)}{i - 1 + \alpha} \tag{5}$$

3.2 DP segmentation: Bigram assumption

The second version of the DP segmentation strategy uses a bigram language model (DP-Bi), with the learner assuming (slightly less naively) that each word is chosen based on the word preceding it. Goldwater et al. (2009) model this using a hierarchical Dirichlet Process (Teh, Jordan, Beal, & Blei, 2006), with the generative process selecting bigrams, words, and sub-word units as follows:

[1] The model additionally includes the generation of word and utterance boundaries, with a word ending with some probability $p_\#$ and an utterance ending with some probability $p_\$$. See Goldwater et al. (2009) for discussion of these model components in the unigram and bigram versions of this strategy.

[2] A more thorough treatment of the role of various model parameters for both the unigram and bigram DP segmentation models can be found in Goldwater et al. (2009).

1. If the pair $< w_{i-1}, w_i >$ is not a novel bigram, choose an existing form ℓ for w_i from those that have been previously generated after w_{i-1}.

2. If the pair $< w_{i-1}, w_i >$ is a novel bigram:

   If w_i is not a novel lexical item, choose an existing form ℓ for w_i.

   If w_i is a novel lexical item, generate a form $(x_1 \ldots x_M)$ for w_i.

As with the DP-Uni model, the DP-Bi model must make a decision between an item being novel or not; the main difference is that the DP-Bi model considers bigrams first. If a bigram isn't novel, the DP-Bi model gives higher probability to bigrams that have been encountered before. If a bigram is instead novel, then the individual lexical item (the second word in the bigram) must also be generated. This is done with a DP in the same fashion as the unigram DP, making this a hierarchical DP. The probability of any bigram $< w_{i-1}, w_i >$ is then determined using equations (6), (7), and (3).

$$P(< w_{i-1}, w_i = \ell > |w_{-i}) = \frac{n_{<w_{i-1}, w_i = \ell>} + \beta P_1(w_i = \ell | w_{-i})}{n_{w_{i-1}} + \beta} \tag{6}$$

Equation (6) calculates the probability of the bigram $< w_{i-1}, w_i >$, given that the second word of the bigram w_i takes the form ℓ and considering all the words observed previously except w_i, denoted by w_{-i}. This includes the number of times ℓ appears as the second word of bigrams beginning with word w_{i-1} ($n_{<w_{i-1}, w_i = \ell>}$), as well as the total number of bigrams beginning with w_{i-1}, denoted by $n_{w_{i-1}}$. The concentration parameter β determines how often a novel second word is expected, with higher values indicating a general preference for more novel bigrams.

Equation (7) describes the process for generating a novel second word in the bigram.

$$P_1(w_i = \ell | w_{-i}) = \frac{t_{w_i = \ell} + \gamma P_0(w_i = x_1 \ldots x_M)}{t + \gamma} \tag{7}$$

A novel second word with form ℓ is based on the number of times any bigram with second word ℓ has been generated, $t_{w_i = \ell}$. The total number of novel bigrams is represented by t. The concentration parameter γ determines how often this novel second word is itself a novel lexical item, constructed from its constituent sub-word units $x_1 \ldots x_M$ using P_0, as in Equation (3).

3.3 DP segmentation inference

Pearl et al. (2011) used a variety of inference algorithms with these two DP segmentation strategies, including both idealized and constrained procedures. Idealized inference procedures provide a computational-level analysis in the sense of Marr (1982), and offer a best-case scenario of how useful the learning assumptions of the model are. Constrained inference procedures provide a more algorithmic-level analysis in the sense of Marr (1982). They in turn offer a more cognitively plausible assessment of how useable the learning assumptions are by humans, who have cognitive limitations on their inference capabilities (particularly infants). Here we focus on one inference algorithm of each kind.

3.3.1 Idealized inference

The original inference algorithm used by Goldwater et al. (2009) for DP segmentation was Gibbs sampling (Geman & Geman, 1984), a batch procedure commonly used for idealized inference, due to its guaranteed convergence on the optimal solution given the model constraints. Gibbs sampling initializes the model parameters (in this case potential word boundaries), and then updates each parameter value one at a time, conditioned on the current value of all other parameters. This process is repeated for a number of iterations until convergence is reached (e.g., Goldwater et al. (2009) and Pearl et al. (2011) used 20,000 iterations).

For DP segmentation, each possible boundary location is a parameter which either has a boundary or not. Boundaries are initialized randomly, and the inference procedure goes through each boundary location in the corpus, deciding whether to place/remove a boundary, given the other current boundary locations. In particular, for each possible boundary, there is a choice between creating a single word (H_0) or two words (H_1) out of the two adjoining pieces. For example, H_0 might be /ðəkɪɾi/ (*thekitty*) while H_1 is /ðə kɪɾi/ (*the kitty*) for the potential boundary location between /ðə/ and /kɪɾi/. The probability of inserting a boundary (H_1) can be defined as the normalized probability of H_1, shown in (8), with $P(H_0)$ and $P(H_1)$ defined by the DP-Uni or DP-Bi generative models:

$$Normalized\ P(H_1) = \frac{P(H_1)}{P(H_1) + P(H_0)} \tag{8}$$

The inference procedure then probabilistically selects either H_0 or H_1, based on their normalized probabilities. If no boundary is placed (H_0), only a single word has to be generated; if a boundary is placed (H_1), two words have to be generated. Intuitively, the model may prefer either H_0 because it requires fewer words or H_1 because it requires shorter words. The exact trade-off depends on the model parameters and, most importantly, on the frequency each word (or bigram) is currently perceived to have.

3.3.2 Constrained inference

Pearl et al. (2011) described several inference procedures that incorporate one or more of the cognitive limitations relevant for infant speech segmentation mentioned before: (i) online processing, (ii) non-optimal decision-making, and (iii) a recency bias. We focus on the one that incorporates all three of these constraints to some degree (called the DMCMC constrained learner by Pearl et al. (2011)). This inference procedure performs inference online, segmenting each utterance as it's encountered. It can also be thought to involve non-optimal decision-making because it probabilistically samples whether to insert a boundary rather than always selecting the highest probability option.[3]

It additionally uses the Decayed Markov Chain Monte Carlo method (Marthi, Pasula, Russell, & Peres, 2002) to implement a recency bias. In particular, similar to the idealized inference procedure,

[3]Unlike Gibbs sampling, which also probabilistically samples whether to insert a boundary, there is no guarantee of convergence on the optimal solution.

it samples individual boundary locations and updates them conditioned on the value of all other currently encountered potential boundaries. However, instead of sampling all currently known potential boundaries equally, the locations to sample are selected based on a decaying function anchored from the most recently encountered potential boundary location (at the end of the current utterance). The probability of sampling potential boundary b_a, which is a potential boundaries away from the end of the current utterance, is determined by (9):

$$P(b_a) = \frac{a^{-d}}{\sum a_i^{-d}} \tag{9}$$

The parameter d implements the recency effect: larger values of d indicate stronger biases, concentrating the boundary sampling efforts on more recently encountered data. We discuss results using d = 1.5, which implements a strong recency bias: using this d value, Phillips and Pearl (2015b) found that 83.6% of sampled boundaries occurred in the current utterance in their corpus of English child-directed speech, 11.8% in the previous utterance, and only 4.6% in any other previous utterance.

4 How well does this work cross-linguistically?

4.1 Cross-linguistic corpora

Phillips and Pearl (2014b, 2014a) evaluated the DP segmentation strategy on seven languages: English, German, Spanish, Italian, Farsi, Hungarian, and Japanese. These languages vary in many ways, including their morphology: some are more agglutinative and have more regular morphology systems (Hungarian, Japanese) while the others are fusional to different degrees and have less regular morphological systems (English, German, Spanish, Italian, and Farsi). Syllabified child-directed speech corpora were derived from the CHILDES database (MacWhinney, 2000; Gervain & Erra, 2012; Phillips & Pearl, 2015b)[4] and relevant summary statistics for them are shown in Table 1.

We can make a few observations. First, not all languages had corpora available of speech directed at children younger than a year old, so the age range does vary. Second, the corpora vary in size, though this doesn't appear to negatively impact the results for smaller corpora. Third, the number of unique syllables in each corpus varies considerably by language (e.g., Spanish: 522, Hungarian: 3029). While some of this variation is due to the size of the corpus itself (more utterances allow more syllable types to appear), there are also language-specific phonotactic restrictions on syllables that impact the number of syllable types observed. For example, in Japanese only the phoneme /ɴ/ may appear after a vowel, and in Spanish only the phoneme /s/ can appear as the second consonant in a coda. In contrast, languages such as English, German, and Hungarian allow much more complex syllable types (e.g., consider the English coda in *warmth*, /ɹmθ/).

The average number of syllables per utterance also varies, which is partially due to speech directed at older children containing longer utterances (e.g., Farsi utterances have 6.98 syllables per utterance and are directed at children up to age five). However, the number of syllables per utterance is also

[4] See Phillips (2015) for details of this process.

Language	Corpora	Age range	# Utt	# Syl types	Syls/Utt	B Prob
English	Brent	0;6-0;9	28391	2330	4.16	76.26
German	Caroline	0;10-4;3	9378	1682	5.30	68.60
Spanish	JacksonThal	0;10-1;8	16924	522	4.80	53.93
Italian	Gervain	1;0-3;4	10473	1158	8.78	49.94
Farsi	Family, Samadi	1;8-5;2	31657	2008	6.98	43.80
Hungarian	Gervain	1;11-2;11	15208	3029	6.30	51.19
Japanese	Noji, Miyata, Ishii	0;2-1;8	12246	526	4.20	44.12

Table 1: Summary of the syllabified child-directed speech corpora, including the CHILDES database corpora they are drawn from (Corpora), the age ranges of the children they are directed at (Age range), the number of utterances (# Utt), the number of unique syllables (# Syl types), the average number of syllables per utterance (Syls/Utt), and the probability of a word boundary appearing between syllables (B Prob).

impacted by the number of syllables per word - languages that tend to be more monosyllabic will tend to have fewer syllables per word, and so their utterances will tend to have fewer syllables as well. Boundary probability captures this monosyllabic tendency, where higher probabilities indicate that syllables tend to be followed by boundaries (i.e., words are more likely to be monosyllabic). For example, the English and German data have higher boundary probabilities than the other languages, and therefore tend to have more monosyllabic words.

4.2 Gold standard evaluation

4.2.1 Evaluation metrics

We first present the DP segmentation's ability to match the gold standard, the adult-level knowledge typically represented via the orthographic segmentation. There are multiple units a segmentation can be measured on: word tokens, lexical items, and word boundaries. For example, the utterance *The penguin is next to the kitty* might be segmented as *Thepenguin is nextto the kitty*. There are seven word tokens (individual words) in the original sentence, but the segmentation only identifies three of those tokens (*is*, *the* and *kitty*). The same utterance has six lexical items, which correspond to the unique words: *the, penguin, is, next, to, kitty* (*the* appears twice). Again, the segmentation only correctly identifies three lexical items (*is*, *the*, and *kitty*). This utterance also has six word boundaries (excluding the utterance boundaries) and the segmentation correctly identifies four of those (*thepenguin is*, *is nextto*, *nextto the*, and *the kitty*). This highlights the differences between the units.

Word tokens are impacted by word frequency, while lexical items factor out word frequency. Measuring boundary identification tends to yield better performance than measuring word tokens or lexical items. Intuitively, this is because identifying a boundary requires the model to be correct only once. On contrast, correctly segmenting words requires being correct twice, both in inserting the word-initial and word-final boundaries (unless the word is at an utterance edge).

No matter which unit we use for comparison, they are measured with the same metrics: precision and recall, which are often combined into a single summary statistic, the F-score. Precision captures how accurate the segmentation is: for each unit identified, is that unit correct in the segmentation? Recall captures how complete the segmentation is: for each unit that should have been identified, is that unit identified in the segmentation? In signal detection theoretic terms, these correspond to (10), which involve *Hits* (units correctly identified in the segmentation), *False Alarms* (units identified in the segmentation that aren't correct), and *Misses* (units not identified in the segmentation that are nonetheless correct).

$$\textit{Precision} = \frac{\textit{Hits}}{\textit{Hits} + \textit{False Alarms}} \qquad \textit{Recall} = \frac{\textit{Hits}}{\textit{Hits} + \textit{Misses}} \tag{10}$$

High precision indicates that when a unit is identified in a segmentation, it is often correct. High recall indicates that when a unit should be identified in a segmentation, it often is. Because both of these are desirable properties, they are often combined into the F-score via the harmonic mean. Precision, recall, and F-scores all range between 0 and 1 (though sometimes this is represented as a percentage between 0 and 100), with higher values indicating a better match to the gold standard.

$$\textit{F-score} = \frac{2 * \textit{Precision} * \textit{Recall}}{\textit{Precision} + \textit{Recall}} \tag{11}$$

4.2.2 Model training and parameter estimation

Because the algorithms used for inference are probabilistic in nature, each modeled strategy (DP-Uni and DP-Bi) was trained and evaluated five times, with the results averaged. Although the learning process modeled is unsupervised, Phillips and Pearl (2014a, 2014b) nonetheless separated the data into training and test sets to better determine how each modeled strategy adapted to new data. Each corpus was randomly split five times so that the training set consisted of 90% of the corpus and the remaining 10% became the corresponding test set. The corpora themselves are temporally ordered, so utterance order captures the order an infant might encounter the utterances. This relative ordering was preserved in both the training and test sets.

The free parameters for the DP-Uni (α) and DP-Bi strategies (β, γ) were set by searching ranges derived from those explored in Goldwater et al. (2009) and Pearl et al. (2011): α, $\beta \in$ [1, 500], $\gamma \in$ [1, 3000]. For each language and each strategy, a learner using the idealized inference algorithm was used to determine which free parameter values resulted in the best word token F-score. The values identified by this process are shown in Table 2.

When we look at the best parameter values, it turns out that the DP-Uni strategy fares best on all languages when α = 1. This indicates a strong bias for small proto-lexicons, since novel words are dispreferred. In contrast, the DP-Bi strategy has more variation, roughly breaking into three classes. The first class, represented by English, Italian, and German, has β = 1 and γ between 90 and 100. This indicates a very strong bias for small proto-lexicons, since novel bigrams are strongly dispreferred (β) and novel words as the second word in a bigram are somewhat dispreferred (γ). The second class, represented by Spanish and Japanese, has β between 200 and 300 and γ between 50 and 100. This

	DP-Uni	DP-Bi	
	α	β	γ
English	1	1	90
Italian	1	1	90
German	1	1	100
Spanish	1	200	50
Japanese	1	300	100
Farsi	1	200	500
Hungarian	1	300	500

Table 2: Best free parameter values for all unigram and bigram Bayesian segmentation strategies across each language.

indicates a weaker bias for small proto-lexicons, since novel bigrams are only somewhat dispreferred (β) and novel words as the second word in a bigram are also only somewhat dispreferred (γ). The third class, represented by Farsi and Hungarian, has β between 200 and 300 and γ = 500. This indicates an even weaker bias for small proto-lexicons, since novel bigrams are again only somewhat dispreferred (β) and novel words as the second word in a bigram are even less dispreferred (γ).

Since we are setting these parameter values for our analyses, this translates to the modeled infant already knowing the appropriate values for each language. For the DP-Uni strategy, this may reflect a language-independent bias, since the values are all the same. However, for the DP-Bi strategy, the infant would need to adjust the values based on the ambient language. For either strategy, one potential way to converge on the best values is to have hyperparameters on them and simply learn their values at the same time as segmentation is learned. Incorporating hyperparameter inference into the DP segmentation strategy would be a welcome avenue for future segmentation modeling work, particularly if it turns out infants are using something like the DP-Bi strategy. Here we discuss the results obtained by manually setting the free parameters to their respective optimized values in each language.

4.2.3 Baseline strategy: Random Oracle

A baseline strategy first explored by Lignos (2012) is a random oracle strategy (RandOracle). The random aspect refers to the strategy treating each possible boundary location as a Bernoulli trial. The oracle aspect comes from the strategy already knowing the true probability of a boundary occurring in the corpus (B Prob in Table 1). Boundaries are then randomly inserted with this probability.

4.2.4 Cross-linguistic performance

Table 3 presents the gold standard word token F-score results for each learner on all seven languages. First, we can see that bigram assumption is generally helpful, though the degree of helpfulness varies cross-linguistically. For example, it seems very helpful in English and German (e.g., English Idealized DP-Uni: 53.1 vs. DP-Bi: 77.1; German Constrained DP-Uni: 60.3 vs. DP-Bi: 82.6) and not

		Eng	Ger	Spa	Ita	Far	Hun	Jpn
DP-Uni	**Idealized**	53.1	60.3	55.0	61.9	66.6	59.9	63.2
	Constrained	55.1	60.3	56.1	58.6	59.6	54.5	63.7
DP-Bi	**Idealized**	77.1	73.1	**64.8**	**71.3**	**69.6**	**66.2**	**66.5**
	Constrained	**86.3**	**82.6**	60.2	60.9	62.5	59.5	63.3
Baseline	**RandOracle**	56.4	47.5	27.0	22.8	20.3	26.4	26.1

Table 3: Word token F-scores for learners across English, German, Spanish, Italian, Farsi, Hungarian, and Japanese. Higher token F-scores indicate better performance, with the best score for each language in bold.

particularly helpful at all in Japanese (Japanese Idealized DP-Uni: 63.2 vs. DP-Bi: 66.5; Japanese Constrained DP-Uni: 63.7 vs. DP-Bi: 63.3). Still, with the exception of the English DP-Uni learner, every single Bayesian learner does better than the random oracle baseline. Interestingly, in English, the DP-Bi constrained learner has the highest score of all learners in all languages. Altogether, this suggests that the DP segmentation strategy is generally a very good one for identifying words in fluent speech.

Nonetheless, something seems to be going on cross-linguistically. Why do we see such variability in segmentation performance (DP-Uni: 53.1-66.6; DP-Bi: 59.5-86.3; RandOracle: 20.3-56.4)? The variability in the random oracle baseline is particularly suggestive that there is something about the languages themselves, rather than something specific to the DP segmentation strategy. More specifically, English and German seem inherently easier to segment than the other languages.

Fourtassi et al. (2013) suggested that some languages are inherently more ambiguous with respect to segmentation than others. Specifically, even if all the words of the language are already known, some utterances can *still* be segmented in multiple ways (e.g., /gɹejtfʌl/ segmented as *great full* and *grateful* in English). The degree to which this occurs varies by language, with the idea that languages with high inherent ambiguity are harder to correctly segment. If this is true, we might expect that low inherent segmentation ambiguity correlates to high performance by segmentation strategies. With this in mind, perhaps English and German have lower inherent segmentation ambiguity than the other languages (RandOracle English: 56.4, German: 47.5, Other languages: 20.3-26.4).

In order to quantify this ambiguity, Fourtassi et al. (2013) proposed the normalized-segmentation entropy (NSE) metric:

$$NSE = -\sum_{s} P_s \log_2(P_s)/(N-1) \tag{12}$$

Here, P_s represents the probability of a possible segmentation *s* of an utterance and N represents the length of that utterance in terms of potential word boundaries (which is determined by the number of syllables for our learners). To calculate the probability of an utterance, we use the unigram or bigram DP generative model equations described above in section 3, since these represent the probability of generating that utterance under a unigram or bigram assumption. As an example, to calculate the NSE of a single utterance /aimgɹejtfʌl/, we use the unigram and bigram model equations to gen-

erate the probability of every segmentation comprised of true English words (P_s above). In this case, two segmentations are possible: *I'm grateful* and *I'm great full*. The probabilities for each segmentation are then used in Equation 12 above, with $N = 2$ since there are two potential word boundaries among the three syllables.

Because a low NSE represents a true segmentation that is less inherently ambiguous for the learners using the n-gram assumptions tested here, English and German should have lower NSE scores if inherent segmentation ambiguity was the explanation for the better segmentation performance. Table 4 shows the NSE scores for both unigram and bigram learners for all seven languages, with token F-scores for the respective idealized inference learners for comparison.

DP-Uni	NSE	F-score	**DP-Bi**	NSE	F-score
German	0.000257	60.3	**German**	0.000502	73.0
Italian	0.000348	61.9	**Italian**	0.000604	71.3
Hungarian	0.000424	59.9	**Hungarian**	0.000694	66.2
English	0.000424	53.1	**English**	0.000907	77.1
Farsi	0.000602	66.6	**Spanish**	0.00103	64.8
Japanese	0.00126	55.0	**Farsi**	0.00111	69.6
Spanish	0.00128	63.2	**Japanese**	0.00239	66.5

Table 4: Average NSE scores across all utterances in a language's corpus, ordered from lowest to highest NSE and compared against the idealized inference token F-score for the language. Results are shown for both the DP-Uni and DP-Bi models. Lower NSE scores represent less inherent segmentation ambiguity and higher token F-scores indicate a better segmentation performance.

From Table 4, we see that German fits with the hypothesis that low NSE predicts higher segmentation performance, having in both cases the lowest NSE scores. Yet English doesn't fit this pattern, ranking third/fourth overall for a unigram DP learner and fourth overall for a bigram DP learner. This is despite English having the lowest token F-scores for the DP-Uni learner and highest token F-scores for the DP-Bi learner. Because of this, the high segmentation performance on both German and English can't simply be due to both having lower inherent segmentation ambiguity.

More generally, it becomes clear by looking at all seven languages that low NSE doesn't always lead to higher token F-scores. If it did, we would expect to find a significant negative correlation between NSE score and token F-score – but this doesn't happen (DP-Uni: r = -0.084, p = 0.86; DP-Bi, r = -0.341, p = 0.45). Examining individual languages in Table 4, this lack of correlation is apparent. The DP-Uni Farsi NSE score is ranked fifth lowest, but in fact has the highest F-score, while the DP-Uni Spanish NSE score is actually the worst, though it has the second best F-score. When we turn to the DP-Bi learners, we see that Hungarian has the third best NSE score but the next to worst F-score, while English has the fourth worst NSE score but the best F-score. So, NSE isn't the principal factor determining segmentation performance, though it may play some role.

An alternative factor comes from considering how often words of the language tend to be monosyllabic. This is captured by the boundary probability in Table 1, where English and German both have a much higher probability of having a boundary appear after a syllable (76.26% and 68.60%,

respectively, compared to the next highest language Spanish, with boundary probability 53.93%). These are precisely the languages that especially benefit – though only for the DP-Bi and RandOracle learners.

One possible explanation for why boundary probability especially impacts these learners relates to the types of errors these learners make. More specifically, if a learner tends to oversegment (i.e., incorrectly insert word boundaries), languages that tend to be more monosyllabic already may benefit more – this bias to insert word boundaries may yield true words more often by sheer luck in these languages. We can get a sense of whether a strategy tends to oversegment by looking at the errors it produces. Table 5 shows the percentage of all segmentation errors that were oversegmentations for each learner, where possible error types are undersegmentations like *thekitty*, oversegmentations like *the ki tty*, and other segmentation errors like *theki tty*.

		Oversegmentation Errors (%)						
		Eng	Ger	Spa	Ita	Far	Hun	Jpn
DP-Uni	**Idealized**	1.7	9.1	8.7	39.9	47.7	45.3	39.0
	Constrained	9.0	15.9	25.8	53.8	68.0	55.3	53.5
DP-Bi	**Idealized**	13.8	26.0	33.0	73.1	59.8	58.0	58.4
	Constrained	44.8	60.6	72.8	89.9	93.4	82.7	79.9
Baseline	**RandOracle**	51.7	60.0	57.7	57.7	58.7	56.9	54.5

Table 5: Percentage of errors which resulted in an oversegmentation as compared to adult orthographic segmentation.

If we look at the DP-Uni learners, we see that there *isn't* an oversegmentation tendency for English and German, though there is for most of the rest of the languages. Because English and German are the only two languages where an oversegmentation tendency is specifically beneficial (because they tend to have more monosyllabic words), this may be why the DP-Uni learners don't do much better on English and German compared with the rest of the languages. This contrasts noticeably with the DP-Bi and RandOracle oversegmentation tendencies, which are comparatively much higher for English and German (e.g., DP-Bi Constrained English: 44.8% and RandOracle English: 51.7% vs. DP-Uni Constrained English: 9.0%). So, English and German, which tend to have more word boundaries per utterance anyway, yield better performance for learners that have a stronger tendency to insert word boundaries.

This makes an interesting testable prediction about infant segmentation more generally. If the segmentation strategy infants use leads them to oversegment more often (such as the DP-Bi strategy here, particularly when coupled with constrained inference), we might expect infant segmentations to better match adult-like segmentations in languages whose child-directed speech contains more monosyllabic words (e.g., English and German). In contrast, for languages with fewer monosyllabic words, we would expect infant segmentation to match adult-like segmentation less well.

4.3 A more flexible metric: Reasonable errors

We know that infant segmentation certainly isn't a perfect match to adult-like segmentation, given the available observational and behavioral data. With this in mind, Phillips and Pearl (2014a, 2014b, 2015b) considered an output evaluation that allowed the following "reasonable errors" as legitimate early segmentations:

1. Oversegmentations that result in real words (e.g., *grateful* /gɹejt fʌl/ segmented as *great* /gɹejt/ and *full* /fʌl/)
2. Oversegmentations that result in productive morphology (e.g., segmenting off *–ing* /iŋ/)
3. Undersegmentations that produce function word collocations (e.g., segmenting *that a* as *thata*)

Table 6 offers some examples of each reasonable error type in different languages, while Table 7 shows how frequently each error type is made by the modeled learners.

		True	**Segmented**
Real words	**Spa**	*porque* 'because'	*por que* 'why'
	Jap	*moshimoshi* 'hello'	*moshi moshi* 'if if'
Morphology	**Ita**	*devi* 'you must'	*dev i* 'must' 2^{nd}-PL
	Far	*miduni* 'you know'	*mi dun i* PRES 'know' 2^{nd}-SG
Function words	**Ita**	*a me* 'to me'	*ame* 'to-me'
	Far	*mæn hæm* 'me too'	*mænhæm* 'me-too'

Table 6: Examples of reasonable errors (with English glosses) made in different languages. *True* words refer to the segmentation in the original corpus, while *Segmented* output represents the segmentation generated by a modeled learner.

For the errors resulting in at least one true word, Phillips and Pearl (2014a, 2014b, 2015b) included only those occurring at least five times in the corpus because of the reason these types of errors are helpful – namely, they boost the perceived frequency of the true word. For example, if a model segments *grateful* as *great* and *full*, then the next time the word *great* is encountered, it is more likely to be segmented because it has been previously seen. So, only true word errors occurring with some frequency are likely to have this beneficial effect. Additionally, this ensures that typos and nonsense words in the corpus are unlikely to be treated as true words.

We can see in Table 7 that certain modeled learners tend to produce more real word errors: learners using constrained inference and learners in Italian, Farsi, and Japanese. Additionally, the DP-Bi

		Eng	Ger	Spa	Ita	Far	Hun	Jpn
		Real Word Errors (%)						
DP-Uni	**Idealized**	1.0	3.3	2.8	23.7	20.1	11.9	17.7
	Constrained	3.4	4.5	6.3	27.6	25.5	14.9	21.4
DP-Bi	**Idealized**	5.8	7.9	11.2	38.3	24.6	17.8	26.7
	Constrained	29.6	17.6	15.1	57.0	41.5	27.7	34.6
Baseline	**RandOracle**	17.5	7.7	13.6	14.9	10.0	8.6	12.6
		Morphology Errors (%)						
DP-Uni	**Idealized**	0.2	2.7	2.8	3.3	5.0	2.5	9.1
	Constrained	0.6	4.6	7.5	4.8	7.5	3.4	10.5
DP-Bi	**Idealized**	1.0	7.7	10.4	6.3	8.4	3.3	10.4
	Constrained	2.6	24.9	20.4	6.7	13.0	4.6	16.9
Baseline	**RandOracle**	2.2	12.1	10.6	3.0	5.1	3.0	10.1
		Function Word Errors (%)						
DP-Uni	**Idealized**	8.8	27.2	8.9	6.4	4.3	2.3	6.9
	Constrained	10.2	26.7	7.7	5.8	3.1	2.3	7.7
DP-Bi	**Idealized**	15.7	28.2	6.4	5.8	3.7	2.9	5.2
	Constrained	9.9	10.8	2.3	1.1	0.2	1.7	2.6
Baseline	**RandOracle**	3.6	8.7	4.2	3.4	1.0	1.1	2.1

Table 7: Percentage of model errors which produced reasonable errors of different kinds. Real Word Errors represent true words in the corpus occurring at least five times. Morphology Errors represent morphological affixes occurring in the correct location (e.g. suffixes after the main word, prefixes at the beginning). Function Word Errors represent collocations of function words.

learners tend to yield more than the DP-Uni learners. This is likely due to the oversegmentation tendency – these are the same learners that also tend to oversegment more often. Because real word errors occur due to oversegmentation, this makes this error type more likely to occur for learners that oversegment.

For the errors resulting in productive morphology, Phillips and Pearl (2014a, 2014b, 2015b) referenced lists of morphemes for each language produced by linguistically-trained native speakers. Similar to the true word errors, oversegmentations are more likely to produce productive morphology because morphological affixes smaller than words can only be produced through oversegmenting. However, we do note that sub-syllabic morphology was excluded, due to the modeled learners perceiving the speech stream as a sequence of atomic syllables. This ruled out much of the common inflectional morphology in Indo-European languages (e.g., *-s* in English). Still, German, Spanish, Farsi, and Japanese have a number of these errors, regardless of the specific modeled learner.

Function word collocation errors, in contrast to the other two types, occur due to undersegmentation. So, learners with a stronger oversegmentation bias, like the constrained DP-Bi learner, produce these relatively less frequently. Looking across the languages, we can see that English and German tend to have more of these errors, perhaps because of the frequency of monosyllabic function words.

For example, many of these errors are combinations of the form MODAL VERB+PRONOUN (e.g., *can you*), COPULA+PRONOUN (e.g., *are you*), or PREPOSITION+DETERMINER (e.g., *in a*). The other languages tend to also have richer morphology which can negate the need for separate function words.

Table 8 shows the evaluation when we consider all three reasonable error types as acceptable segmentation. The main difference (of course) is that the average token F-score is higher than before (e.g., unadjusted DP-Bi average: 68.9; adjusted DP-Bi average: 77.4). Perhaps more interestingly, we can see that the constrained learners perform as well as or better than the idealized bigram learner on most languages. In cases where they perform noticeably below the idealized learner (DP-Uni: Italian, Hungarian; DP-Bi: Spanish, Italian), they don't perform so poorly that we would consider them unsuccessful (e.g., all are above a word token F-score of 60). This indicates that constrained inference doesn't necessarily hinder this Bayesian segmentation strategy – and in fact, may be helpful for some languages – especially once we incorporate a more nuanced standard of segmentation success.

		Eng	Ger	Spa	Ita	Far	Hun	Jpn
DP-Uni	**Idealized**	56.8	71.1	59.2	68.0	72.1	64.8	71.1
	Constrained	60.0	71.3	64.6	65.1	75.5	61.4	72.7
DP-Bi	**Idealized**	81.5	82.9	**74.7**	**76.8**	**76.4**	**72.0**	**76.3**
	Constrained	**90.1**	**88.4**	71.5	71.2	75.1	71.4	75.1
Baseline	**RandOracle**	64.6	59.4	42.4	31.0	31.7	33.1	41.8

Table 8: Adjusted word token F-scores, counting reasonable errors as acceptable output, for learners across English, German, Spanish, Italian, Farsi, Hungarian, and Japanese. Higher token F-scores indicate better performance, with the best score for each language in bold.

More generally, many errors made by each learner may not be so harmful. For example, they can potentially be useful for later segmentation (real word errors), identifying productive morphology (morphology errors), grammatical categorization (morphology errors), and the early stages of syntactic bootstrapping (morphology errors, function word collocations). So, the output units of this strategy may be useful for the acquisition process, even if they're not the adult-like segmentation. But how do we tell if they really *are* useful for acquisition?

5 How useful are the units?

Model output can be measured in two core ways: intrinsically and extrinsically (Galliers & Jones, 1993). Intrinsic evaluations are concerned with the model's direct objective, i.e., the task it was trained for. Intrinsic measures for speech segmentation include comparisons against the gold standard (e.g., F-score), measures of model fit (e.g., log posterior probability), and comparison against behavioral results from experimental setups (e.g., see Frank, Goodman, and Tenenbaum (2009) and Kolodny, Lotem, and Edelman (2015)). In contrast, extrinsic measures are concerned with how the model output is used for alternative tasks. For speech segmentation, this relates to how the segmented items are used in the overall acquisition process. That is, because the segmented units are used

by other processes during acquisition, it makes sense to try to measure their effectiveness for these secondary tasks. Two ways to do this are (1) incorporate the secondary task into the segmentation model (*joint modeling*), and (2) use the output of the segmentation process to perform the secondary task in isolation (*downstream evaluation*).

Joint modeling for acquisition makes sense when we have reason to believe the two tasks are solved at the same time by infants. So, for example, while joint modeling of early segmentation and phonotactics may be possible, experimental evidence suggests that phonotactic learning begins significantly after early segmentation (Bortfeld et al., 2005; Thiessen & Saffran, 2003; Mattys et al., 1999). This makes it seem cognitively implausible to model both processes jointly.

However, for tasks that may be solved simultaneously, the idea is that the two tasks can bootstrap off each other, resulting in better overall performance in both than if the tasks were solved in isolation. Still, this requires the modeler to make assumptions about what the infant knows with respect to how the two tasks relate to each other, so that the connections between them can be leveraged. More specifically, because these assumptions are typically built into the joint model directly, the implication is that infants already know them *a priori* (either innately or learned very rapidly in the first months of life). Depending on the particular joint relationship, these assumptions may be plausible – or not.

Downstream evaluation, on the other hand, assumes that there is *no* feedback between the first task and the second – the first task happens first, and its output is used as input to the second task. This may be appropriate when there's a known gap between the tasks, such as early segmentation and phonotactic learning. It may also be appropriate if the tasks have some overlap, but there's reason to think infants don't leverage the synergies between the tasks.

In a sense, joint modeling represents a *best* case scenario for infants. The infant knows that the two tasks are connected and knows specifically how they are related, allowing for joint learning to take advantage of the two processes. In contrast, downstream evaluation represents a *worst* case scenario because the infant doesn't realize the two tasks are connected and so can't leverage the additional information available. In reality, infants may often fall somewhere in between these two endpoints, depending on the specific acquisition tasks. By considering both perspectives, we can set upper and lower bounds on what acquisition success looks like. This is particularly helpful for tasks happening early in acquisition, because we are often unsure exactly what infant knowledge representations look like. So, instead of relying only on intrinsic measures that assume these representations take a certain form, we can also use extrinsic measures to evaluate the utility of the inferred representations.

5.1 Learning the right stress-based segmentation cue

In languages with lexical stress, where stress is placed on syllables within a word, stressed syllables can provide a cue to word boundaries in fluent speech particularly when stressed syllables reliably occur at word edges. In languages with fixed lexical stress, like Hungarian, this cue is essentially deterministic: every Hungarian word has stress on the initial syllable. In languages with variable lexical stress, like English, this cue can be probabilistic: the exact position of stressed syllables varies from word to word (e.g., *ápple* vs. *banána*), though there is a reliable tendency for words in English child-directed speech to begin with stressed syllables (Pearl et al., 2011; Phillips & Pearl, 2015b).

Importantly, the existence of these reliable stress cues and their exact implementation depends on the language. Given this, there's been significant interest about when infants learn to identify and leverage these stress-based cues to speech segmentation (e.g, Jusczyk et al., 1993; Jusczyk, Houston, & Newsome, 1999; E. Johnson & Jusczyk, 2001; Thiessen & Saffran, 2003; Thiessan & Saffran, 2007).

Because these stress-based cues are language-specific, it's usually thought that infants have to first identify enough words in the language to determine the specific instantiation of the stress-based cue (assuming there is one). The earliest evidence for infant sensitivity to stress cues is at 7.5 months (Jusczyk, Houston, & Newsome, 1999) while the early stages of segmentation begin around 6 months (Bortfeld et al., 2005). However, given how close in time these two processes occur (stress-cue identification and early segmentation), it's quite possible they overlap. That is, infants could be learning to segment using statistical cues while simultaneously trying to identify if there's a dominant stress pattern in the segmented words. So, a joint model of acquisition seems reasonable. Of course, it's also possible that in reality early segmentation provides a seed pool of segmented words very quickly, and infants subsequently use these to infer a stress-based segmentation cue. So, a downstream evaluation also seems reasonable. We discuss studies using each as extrinsic evaluation metrics for the DP segmentation strategy.

5.1.1 Stress cue identification: Joint modeling

Doyle and Levy (2013) investigate a joint model of the DP-Bi segmentation strategy (the best-performing DP segmentation variant) and stress pattern identification, using idealized inference. In particular, this model assumes that the learner knows lexical stress exists in the language and attempts to identify the dominant pattern for every sequence of M syllables (i.e., it assumes all two syllable words have a dominant pattern, and this pattern may be different from the one all three syllable words have, and so on). This is accomplished via an update to the P_0 probability from (3), as shown in (13), which now depends both on the probability of the syllables in the word w_i (P_W) and the stress pattern s_i observed for a word with M syllables (P_S).

$$P_0(w_i, s_i) = P_W(w_i)P_S(s_i|M) \tag{13}$$

More specifically, P_W is calculated using P_0 from the original DP-Bi implementation. P_S is calculated as a multinomial over all possible stress patterns of length M (given a uniform prior), with the parameter values derived from observed frequency and plus-one smoothing. This places minimal restrictions on stress patterns for words: a word might possess multiple stressed syllables or no stressed syllables at all.

Doyle and Levy (2013) evaluated their joint model on the Korman corpus of English child-directed speech (Korman, 1984), as modified by Christiansen, Allen, and Seidenberg (1998). This sample contains speech directed at infants between 1.5 and 4 months old, with 24493 word tokens. Notably, this corpus highlights the monosyllabic bias in English child-directed speech, as the corpus consists of 87.3% monosyllabic words. Phonemic forms, syllabification, and stress patterns were all derived from the MRC Psycholinguistic Database (Wilson, 1988), and demonstrated that this English speech

sample had a strong word-initial stress bias (89.2% of all multisyllabic tokens had stress on the first syllable).

Doyle and Levy (2013) compared modeled learners using a joint DP-Bi+Stress strategy against learners using the original DP-Bi strategy. From the perspective of extrinsic measures of segmentation, the question is whether there is bootstrapping observed for both processes. If so, this means the segmented units provide useful information for inferring stress-based cues and stress-based cues provide useful information for segmentation.

To determine whether the segmented units are helpful for inferring stress-based cues, Doyle and Levy (2013) compared the word-initial stress bias in the segmented words generated by both strategies. Given that the English stress-based cue is that words begin with stressed syllables, modeled learners who succeed should have a bias for stress-initial words over stress-final words. It turned out that while both strategies generated this bias for the segmented units (as shown in Table 9), the joint DP-Bi+Stress strategy did so slightly more strongly (DP-Bi+Stress: 87.3% vs. DP-Bi: 86.4% word-initial stress on bisyllabic words). So, there is some synergistic value in the segmented units for the learner – they are useful for more easily inferring the correct stress-based cue in English, with the idea that a stronger bias in the correct direction makes inference easier.

To determine whether the stress-based cues are helpful for segmentation, Doyle and Levy (2013) compared the word token and word type F-scores for both strategies. If there is useful segmentation information contained in the developing representation of the English-specific stress cue, this should be reflected in the F-scores of the DP-Bi+Stress being higher than the DP-Bi alone. As Table 9 shows, this does indeed occur (though again, the benefit is somewhat slight, given the excellent segmentation performance the DP-Bi already achieves on its own).

	DP-Bi+Stress	**DP-Bi**
Stress bias	87.3%	86.4%
Word token F-score	68	67
Word type F-score	80	77

Table 9: Extrinsic measure evaluations for the joint model from Doyle & Levy 2013 (DP-Bi+Stress) compared against the original DP-Bi model. Stress bias indicates the bias towards the correct English stress-based cue to segmentation, with higher percentages indicating a stronger bias. F-scores are shown either including frequency (Word token) or factoring it out (Word type). Higher scores indicate better segmentation performance compared against the gold standard.

These results indicate that jointly inferring boundaries and stress cues does yield a bootstrapping effect for both tasks. However, because this effect is rather small, it may well be that infants can do just fine even if these processes aren't occurring simultaneously.

5.1.2 Stress cue identification: Downstream evaluation

Here we consider the downstream evaluation for the DP segmentation strategy explored by Phillips and Pearl (2015a), where the segmentation process yields a proto-lexicon. We can then evaluate that

proto-lexicon with respect to inferring the correct stress-based cue. Phillips and Pearl (2015a) did this for several DP segmentation variants, some of which had significantly lower F-score performance than others. This can give us a sense of how good segmentation needs to be in order for the units in the proto-lexicon to be helpful for inferring the correct stress-based cue.

The evaluation process itself is similar to the approach taken by Doyle and Levy (2013): examine the proto-lexicon yielded by different DP segmentation strategy variants and calculate the bias towards word-initial stress for each. Table 10 shows the adjusted segmentation F-scores achieved by each strategy against the gold standard on English child-directed speech from the Brent corpus, when reasonable errors are counted as acceptable. It additionally shows the strength of the bias in the inferred proto-lexicon towards word-initial stress. To determine the stress patterns for segmented words, Phillips and Pearl (2015a) referenced the English Callhome Lexicon (Kingsbury, Strassel, McLemore, & MacIntyre, 1997). Child-register words not found in standard dictionaries (e.g., *moosha*) were manually coded when the proper stress could be reasonably inferred. Additionally, to better approximate the stress of words in fluent speech, monosyllabic words were left unstressed. Similar stress analyses were done for German and Hungarian, using the Caroline (German) and Gervain (Hungarian) corpora of child-directed speech, and determining stress patterns by using the Callhome German Lexicon (Karins, MacIntyre, Brandmair, Lauscher, & McLemore, 1997) and the stress rules of Hungarian (all words are stress-initial).

		Adjusted F-score			**% Stress-initial items**		
		English	German	Hungarian	English	German	Hungarian
	Adult seg	100	100	100	88.4%	90.3%	100%
Unigram	**Idealized**	56.8	71.1	64.8	87.3%	90.8%	96.9%
	Constrained	60.0	71.3	61.4	85.3%	87.6%	93.1%
Bigram	**Idealized**	81.5	82.9	72.0	88.4%	90.9%	97.9%
	Constrained	90.1	88.4	71.4	90.6%	92.6%	96.4%
Baseline	**RandOracle**	64.6	59.4	33.1	46.7%	49.6%	52.5%

Table 10: Downstream evaluation for different DP segmentation strategy variants, compared against the adult orthographic segmentation and the random oracle baseline in English, German, and Hungarian. Adjusted word token F-scores are shown, which count reasonable errors as acceptable segmentation. The percentage of bisyllabic word types in the inferred proto-lexicon with word-initial stress are shown, given all bisyllabic word types the learner identified containing a single stressed syllable.

We can first observe that if segmentation matches adult orthographic segmentation (with an F-score of 100), the stress-initial bias is quite strong: 88.4% (English), 90.3% (German), or 100% (Hungarian) of bisyllabic word types with a single stressed syllable have that syllable at the beginning of the word. This should make inferring the stress-based segmentation cue straightforward. Interestingly, every single DP segmentation learner – regardless of its F-score performance – achieves a stress-initial bias that's almost as strong (English, German, Hungarian), as strong (English, German), or stronger (English, German). That is, even strategies whose segmentation seems poorer (DP-Uni learners) or actually worse than the baseline on F-score (e.g., DP-Uni learners in English) achieve a

very strong stress-initial bias in their proto-lexicons. It doesn't matter that their segmentation doesn't match the adult orthographic standard; it's good enough from the perspective of inferring the correct stress cue to segmentation.

This contrasts notably with the baseline random oracle strategy, which is the only strategy to identify a proto-lexicon that fails to have a bias (German), has only a very slight bias in the correct direction (Hungarian), or actually has a bias in the wrong direction (English, where less than 50% of its bisyllabic word types are stress-initial). This occurs even though its adjusted English F-score is higher than the F-scores of the DP-Uni learners. This underscores that even if a strategy's output doesn't match adult segmentation, that output can still be quite useful – as we see here especially with the DP-Uni learners. The important stress pattern property is preserved in the inferred proto-lexicon.

5.1.3 Stress cue identification: Summary

Both the joint modeling and downstream extrinsic metrics suggest that the DP segmentation strategy is quite useful for identifying the stress-based segmentation cue for a language. It's particularly notable that this can occur for the downstream evaluation even if the segmentation doesn't match the adult segmentation very well, as indicated by the standard metric of the F-score. So, even lower quality proto-lexicons may be good enough for subsequent acquisition processes like identifying the language-specific stress cue to speech segmentation. We turn next to another acquisition process that relies on the output of segmentation.

5.2 Learning the meaning of concrete nouns

Infants begin associating word forms from their proto-lexicons with concrete objects such as *cookie* and *nose* when they're between six and nine months old (Bergelson & Swingley, 2012) and so this process could very well overlap with early speech segmentation. Frank et al. (2009) proposed a Bayesian word-mapping strategy for concrete objects that incorporates an infant's developing ideas about referential intention, which they called the Intentional strategy. The Intentional strategy identifies a comparatively accurate lexicon of word-meaning mappings, and accounts for several well-known phenomena in the developmental word-mapping literature, including cross-situational word learning (Yu & Smith, 2007; Smith & Yu, 2008; Yu & Smith, 2011), mutual exclusivity (Markman & Wachtel, 1988; Markman, 1989; Markman, Wasow, & Hansen, 2003), one-trial learning (Carey, 1978; Markson & Bloom, 1997), object individuation (Xu, 2002), and intention reading (Baldwin, 1993). One assumption Frank et al. (2009)'s implementation makes is that speech is represented in its adult orthographic segmentation. Given the Intentional strategy's success at matching developmental data and its reliance on segmented speech, it seems reasonable to use it either within a joint model of segmentation and word learning or as a downstream evaluation of a segmentation strategy's output. We first give an overview of the Intentional strategy and then discuss each evaluation in turn.

5.2.1 The Intentional word learning strategy

Because the Intentional strategy is a Bayesian strategy that relies on a generative model, its components can be represented with a plate diagram as shown in Figure 1. The modeled infant using this strategy observes the individual words uttered (W_s) in a particular situation s and the concrete objects O_s in the vicinity. The learner then infers which objects I_s the speaker intends to refer to as well as the lexicon L the speaker is drawing from in order to make the words W_s refer to those objects. That is, the lexicon L is a set of word-meaning mappings, drawn from the word forms W and the available objects O in all observed situations.

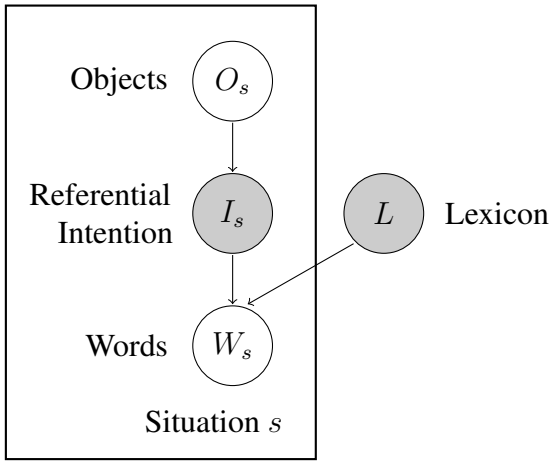

Figure 1: Plate diagram of the Intentional strategy's generative model.

The probability of a corpus of situations S, given an adult speaker's lexicon L, can be represented as in (14). It is the product of the probability of each individual situation s, where the probability of each possible intended object I_s is summed (because the actual intention of the speaker isn't observed).

$$P(C|L) = \prod_{s \in S} \sum_{I_s \in O_s} P(I_s|O_s) P(W_s|I_s, L) \tag{14}$$

The probability of intending to speak about any available object, $P(I_s|O_s)$, is treated as uniform, such that all objects are equally likely to be referred to. The difference between intended objects comes from the probability of selecting the words W_s given the speaker's intentions in that situation I_s and the speaker's lexicon ($P(W_s|I_s, L)$). The modeled learner is aware that some words may be spoken non-referentially (i.e., they don't map to a concrete object) and so this probability is calculated as in (15). More specifically, there is some probability γ that the word is used referentially and some probability (1-γ) that it isn't. The best-performing variant of the Intentional strategy reported in Frank et al. (2009) used $\gamma = 0.1$, implementing a strong bias for words to be used non-referentially. This makes intuitive sense as only a few words (and often only a single word) in an utterance actually refers to a concrete noun (e.g., In *Look at the kitty!*, only *kitty* is referential in this sense).

$$P(W_s|I_s, L) = \prod_{w \in W_s} [\gamma \sum_{o \in I_s} \frac{1}{|I_s|} P_R(w|o, L) + (1 - \gamma) P_{NR}(w|L)] \quad (15)$$

The probability for all observed words W_s is the product of each individual word w. If a word is used referentially, the learner assumes it's chosen uniformly from all words linked in the lexicon to the intended object in question ($P_R(w|o, L)$). For instance, if *duck* is the only word linked to the object DUCK, then it has probability 1. If both *duck* and *bird* are linked to the object DUCK, then each has probability 0.5. This probability is summed across all potential intended objects I_s and averaged.

In contrast, if a word is used non-referentially (in the sense that it doesn't refer to an available concrete noun in O_s), the learner then distinguishes within $P_{NR}(w|L)$ between words that are in the lexicon and words that aren't. A word can be used non-referentially even if it's in the lexicon because a word form may map to more than one meaning (e.g., consider the concrete noun *duck* vs. the action verb *duck*), and the learner allows this possibility. So, if a word in the lexicon is used non-referentially, it's selected with probability proportional to κ. In contrast, if a word not in the lexicon is used non-referentially, it's selected with probability proportional to 1. If $\kappa < 1$, words with entries in the lexicon are less likely to be used non-referentially than words that aren't. The best-performing variant of the Intentional strategy reported in Frank et al. (2009) used $\kappa = 0.05$, implementing a strong bias for words in the lexicon not to be used non-referentially. That is, if a word has an entry in the lexicon mapping the word form to one or more concrete objects, this learner strongly prefers the word to be used to refer to a concrete object.

5.2.2 Learning concrete nouns: Joint modeling

M. Johnson et al. (2010) explore one approach to a joint model of segmentation and word-object mapping. They implement these models using Adaptor Grammars (AGs), an extension of probabilistic context free grammars (PCFGs) that can be used to model segmentation. Using AGs, they implement both the DP-Uni and DP-Bi segmentation strategies, as well as joint models that incorporate the Intentional strategy for learning a concrete noun lexicon. They implement two variants of the Intentional strategy: the original version (+Intention), which they refer to as *reference*, and one that builds in a hard constraint that a word can only refer to a single concrete object within an utterance (+Intention+Only1), which they refer to as *reference1* (for DP-Uni) and *referenceC1* (for DP-Bi). In effect, in (15), the +Intention+Only1 learner has an additional constraint on P_R that considers whether the word has already been used referentially for an intended object in I_s. If it has, P_R=0.

M. Johnson et al. (2010) used idealized inference and evaluated these strategies on the Fernald-Morikawa corpus (Fernald & Morikawa, 1993), which consists of 22,000 words (5,600 utterances) of mother-child play sessions involving pairs of toys. These utterances were phonemically transcribed using the VoxForge dictionary and segmentation assumed phonemes were the basic units of representation. Given the empirical data about infant speech representation, it turns out this may not be the most plausible assumption. Nonetheless, M. Johnson et al. (2010) found synergies between speech segmentation and concrete object learning, as shown below in Table 11, and so future studies may wish to replicate these investigations using syllables as the basic unit of representation.

		Segmentation F	Lexicon F
DP-Uni	Base	53.3	0.0
	+Intention	53.7	14.9
	+Intention+Only1	54.7	14.7
DP-Bi	Base	69.5	0.0
	+Intention	72.6	22.0
	+Intention+Only1	75.0	63.6

Table 11: Extrinsic measure evaluations for the joint model (+Intention, +Intention+Only1) from Johnson et al. (2010) compared against the original DP segmentation strategies in isolation (Base). Segmentation F-score indicates how well the modeled learner segmented the utterances, when compared to the adult orthographic gold standard. Lexicon F-score indicates how well the modeled learner identified the word-object mappings. For both F-scores, higher scores indicate better performance.

First, we can ask if knowing that some of the units in speech refer to available concrete objects improves segmentation. Whether the modeled learner is using the DP-Uni or DP-Bi strategy variant, the answer is clearly yes, though the improvement is more substantial for the DP-Bi learner (e.g., DP-Uni Base=53.3 vs. +Intention+Only1=54.7; DP-Bi Base=69.5 vs. +Intention+Only1=75.0). Moreover, hard-wiring in the constraint that words within an utterance can only refer to at most a single concrete object is helpful (DP-Uni +Intention=53.7 vs. +Intention+Only1=54.7; DP-Bi +Intention=72.6 vs. +Intention+Only1=75.0), though segmentation performance is already quite good for all DP segmentation strategy variants.

Second, we can ask whether knowing the segmented units improves identification of the mappings from word forms to concrete objects in the lexicon. M. Johnson et al. (2010) used a default assumption that no words in the utterance are referential, and so the baseline performance for the DP-Uni and DP-Bi segmentation strategies alone is 0.0. As a point of reference, Frank et al. (2009) achieved a lexicon F-score of 55.0 on the corpus they used, and this was twice as good as the next closest strategy, which had a lexicon F-score of 22.0. Looking at the joint model results, it seems that only the DP-Bi joint model with the constraint restricting a word form's referent within an utterance (DP-Bi +Intention+Only1) achieves a noteworthy lexicon F-score (63.6). Still, it's higher than Frank et al. (2009)'s previous results that used the adult orthographic segmentation (though again, we note that was on a different corpus). This suggests that the imperfectly segmented units are indeed helpful, but only the more sophisticated segmentation strategy (DP-Bi) generates segmentations that are good enough for a joint model to benefit from them, and only if the word-mapping portion of the joint model contains that additional restriction on word reference within an utterance.

More generally, if infants are using joint learning strategies of this kind, these results indicate that the imperfect segmentations generated by the DP segmentation strategy may be sufficient to bootstrap word-object mapping for concrete nouns. In particular, the more sophisticated DP-Bi assumption is the most promising of the DP segmentation variants. Interestingly, this is also supported by the results of the downstream evaluation we discuss in the next section, though whether the learner uses idealized vs. constrained inference turns out to matter.

5.2.3 Learning concrete nouns: Downstream evaluation

Phillips and Pearl (2015a) investigated a downstream evaluation of the syllable-based DP segmentation strategy using the Intentional strategy for learning concrete nouns. First, each modeled learner was trained on a subsection of the Brent English corpus of child-directed speech (Brent & Siskind, 2001), which is naturalistic speech directed at children between six and nine months and consists of 28391 utterances. Using the proto-lexicon it had inferred, each modeled learner then segmented the corpus used by Frank et al. (2009), which is derived from two video files from the Rollins corpus (Rollins, 2003) directed at six-month-olds, where caretakers were asked to play with their infants in an experimental setup. These files were hand-annotated for concrete objects in the immediate vicinity, yielding 619 utterances. The segmented Rollins utterances were then used as input to the Intentional strategy, along with the annotations of available concrete objects in each utterance. The results for both segmentation and inferring the lexicon of concrete nouns are shown in Table 12.

		Segmentation Token F		**Mapping Lexicon P**
		Brent	**Rollins**	**Rollins**
	Adult Segmentation	100	100	58.3
DP-Uni	**Idealized**	53.1	51.4	46.5
	Constrained	55.1	52.4	47.3
DP-Bi	**Idealized**	77.1	74.6	54.4
	Constrained	86.3	81.3	38.8
Baseline	**RandOracle**	56.4	57.6	40.6

Table 12: Segmentation and word-object mapping results for modeled learners in Phillips & Pearl (2015a). Token F-scores are shown for segmentations both on the original Brent corpus and the Rollins corpus. Lexicon precision is shown for the Rollins corpus, representing the accuracy of the word-object mapping model trained on each learner's segmentation.

The first thing to notice is that segmentation performance transfers quite well between the Brent and Rollins corpora, no matter which learner we look at (e.g., DP-Bi Idealized: Brent=77.1 vs. Rollins=74.6). This highlights the generalizability of the segmentation knowledge each learner has internalized by first encountering the Brent data.

Turning to the lexicon of word-mappings inferred by the learners, Phillips and Pearl (2015a) chose to focus on the precision only, with the idea that it is more important for the very early stages of word learning to find a highly accurate set of mappings, even if not all correct mappings are identified. Moreover, certain principles of word learning infants follow, such as mutual exclusivity (Markman & Wachtel, 1988; Markman et al., 2003), would prevent them from learning all possible lexicon mappings in the Rollins corpus (e.g. both *rabbit* and *bunny* can refer to RABBIT). So, Phillips and Pearl (2015a) reasoned precision was the more relevant metric, rather than the F-score which additionally incorporates recall.

While the Intentional strategy trained on the adult orthographic segmentation did the best (58.3),

all of the DP segmentation strategy variants yielded segmentations the Intentional strategy found more useful than a random segmentation (40.6), except for the constrained DP-Bi learner. Strikingly, the constrained DP-Bi learner yielded the least accurate lexicon (38.8), despite having the highest token F-scores among the modeled learners (Brent: 86.3, Rollins: 81.3). This surprising results underscores the importance of extrinsic evaluation metrics – just because a strategy fares well on intrinsic measures like token F-score doesn't mean it will be useful, as assessed by extrinsic measures like downstream evaluation.

On the other hand, of course, good performance on an intrinsic measure doesn't automatically yield poor performance on extrinsic measures. The idealized DP-Bi learner, in contrast to the constrained version, yields the most accurate lexicon (54.4) among the modeled learners and has the second highest segmentation performance on the Rollins corpus (74.6). This is in line with the joint modeling results from M. Johnson et al. (2010), who also found that the best lexicon resulted from the idealized DP-Bi learner. Also, both DP-Uni learners yield better lexicons than the random oracle baseline (DP Uni Idealized=46.5, Constrained=47.3), despite having significantly poorer segmentation performance.

5.2.4 Learning concrete nouns: Summary

The results of both the joint and downstream evaluations using word-object mapping suggest that the DP segmentation strategy (particularly the DP-Bi version) yields segmentations that are useful for learning a lexicon of concrete nouns. Like the findings for stress cue identification, it's notable that yielding a less adult-like segmentation doesn't necessarily mean the segmentation isn't useful for word learning. So, as before, even lower quality proto-lexicons (as measured against adult orthography) may be good enough for acquisition processes that depend on those segmented units.

5.3 Extrinsic evaluations: Summary

More generally, these extrinsic evaluations suggest that judging the quality of an acquisition strategy from multiple perspectives, both intrinsic and extrinsic, is worth doing. Output that may not intrinsically seem so good could very well be good enough to accomplish what it needs to for the language acquisition process. In terms of extrinsic evaluation options, a joint model may be a good option when two tasks are closely interrelated and overlap in development. However, implementing a joint model can prove technically challenging, depending on the modifications required, and requires assumptions about infant learning which may or may not be well-founded. In contrast, downstream evaluation can be applied without altering the learning strategy for either task, though it may underestimate the synergistic information available to children.

6 Closing thoughts

In this chapter, we hope to have shown that the output of early acquisition processes, such as speech segmentation, is not necessarily the knowledge that an adult has. Nonetheless, this output may very

well be useful for infants by providing units that scaffold other acquisition processes. Given this goal, the Bayesian segmentation strategy seems effective for all seven languages tested. Moreover, because learners using this segmentation strategy are looking for useful units, which can be realized in different ways across languages, they can identify foundational aspects of a language that are both smaller and larger than orthographic words.

References

Anderson, J. (1990). *The adaptive character of thought*. Hillsdale, NJ: Erlbaum.

Baldwin, D. A. (1993). Early referential understanding: Infants' ability to recognize referential acts for what they are. *Developmental Psychology*, *29*(5), 832.

Bergelson, E., & Swingley, D. (2012). At 6–9 months, human infants know the meanings of many common nouns. *Proceedings of the National Academy of Sciences*, *109*(9), 3253–3258.

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.

Bonawitz, E., Denison, S., Chen, A., Gopnik, A., & Griffiths, T. (2011). A simple sequential algorithm for approximating bayesian inference. In *Proceedings of the 33rd annual conference of the cognitive science society.*

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.

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

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.

Carey, S. (1978). The child as word learner. In J. Bresnan, G. Miller, & M. Halle (Eds.), *Linguistic theory and psychological reality* (pp. 264–293). Cambridge, MA: MIT Press.

Christiansen, M. H., Allen, J., & Seidenberg, M. S. (1998). Learning to segment speech using multiple cues: A connectionist model. *Language and cognitive processes*, *13*(2-3), 221–268.

Cole, R., & Jakimik, J. (1980). Perception and production of fluent speech. In R. Cole (Ed.), (pp. 133–163). Hillsdale, NJ: Erlbaum.

Cornell, E. H., & Bergstrom, L. I. (1983). Serial-position effects in infants' recognition memory. *Memory & Cognition*, *11*(5), 494–499.

Davis, S. J., Newport, E. L., & Aslin, R. N. (2011). Probability-matching in 10-month-old infants. *Proceedings of the 33rd Cognitive Science Society*, 3011–3015.

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

Doyle, G., & Levy, R. (2013). Combining multiple information types in bayesian word segmentation. In *Hlt-naacl* (pp. 117–126).

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.

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

Fernald, A., & Morikawa, H. (1993). Common themes and cultural variations in japanese and american mothers' speech to infants. *Child development*, *64*(3), 637–656.

Fourtassi, A., Börschinger, B., Johnson, M., & Dupoux, E. (2013). Whyisenglishsoeasytosegment. In *Cognitive Modeling and Computational Linguistics 2013* (pp. 1–10).

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

Galliers, J., & Jones, K. S. (1993). *Evaluating natural language processing systems* (Tech. Rep. No. 291). Computer Laboratory, University of Cambridge.

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.

Gervain, J., & Erra, R. G. (2012). The statistical signature of morphosyntax: A study of Hungarian and Italian infant-directed speech. *Cognition*, *125*(2), 263–287.

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

Goldwater, S., Griffiths, T., & Johnson, M. (2011). Producing power-law distributions and damping word frequencies with two-stage language models. *Journal of Machine Learning Research*, *12*, 2335–2382.

Gulya, M., Rovee-Collier, C., Galluccio, L., & Wilk, A. (1998). Memory processing of a serial list by young infants. *Psychological Science*, *9*(4), 303–307.

Hohne, E., & Jusczyk, P. (1994). Two-month-old infants' sensitivity to allophonic differences. *Perception & Psychophysics*, *56*(6), 613–623.

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). Unsupervised word segmentation for sesotho using adaptor grammars. In *Proceedings of the tenth meeting of the acl special interest group on computational morphology and phonology* (pp. 20–27).

Johnson, M., & Demuth, K. (2010). Unsupervised phonemic chinese word segmentation using adap-

tor grammars. In *Proceedings of the 23rd international conference on computational linguistics* (pp. 528–536).

Johnson, M., Demuth, K., Jones, B., & Black, M. J. (2010). Synergies in learning words and their referents. In *Advances in neural information processing systems* (pp. 1018–1026).

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

Jusczyk, P., Cutler, A., & Redanz, N. (1993). Infants' preference for the predominant stress pattern of english words. *Child Development*, *64*(3), 675–687.

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

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

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.

Kam, C. H., & Newport, E. (2005). Regularizing unpredictable variation: The roles of adult and child learners in language formation and change. *Language learning and development*, *1*(2), 151–195.

Kam, C. L. H., & Newport, E. L. (2009). Getting it right by getting it wrong: When learners change languages. *Cognitive psychology*, *59*(1), 30–66.

Karins, K., MacIntyre, R., Brandmair, M., Lauscher, S., & McLemore, C. (1997). *CALLHOME German Lexicon*. Linguistic Data Consortium.

Kingsbury, P., Strassel, S., McLemore, C., & MacIntyre, R. (1997). *CALLHOME American English Lexicon (PRONLEX)*. Linguistic Data Consortium.

Kolodny, O., Lotem, A., & Edelman, S. (2015). Learning a generative probabilistic grammar of experience: A process-level model of language acquisition. *Cognitive Science*, *39*, 227–267.

Köpcke, K.-M. (1998). The acquisition of plural marking in english and german revisited: schemata versus rules. *Journal of child language*, *25*(02), 293–319.

Korman, M. (1984). Adaptive aspects of maternal vocalizations in differing contexts at ten weeks. *First language*, *5*, 44–45.

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.

Lignos, C. (2012). Infant word segmentation: An incremental, integrated model. In *Proceedings of the 30th west coast conference on formal linguistics* (pp. 237–247).

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.

Markman, E. (1989). *Categorization and naming in children: Problems of induction*. Cambridge, MA: MIT Press.

Markman, E., & Wachtel, G. (1988). Children's use of mutual exclusivity to constrain the meanings of words. *Cognitive Psychology*, *20*, 121–157.

Markman, E., Wasow, J., & Hansen, M. (2003). Use of the mutual exclusivity assumption by young word learners. *Cognitive Psychology*, *47*, 241–275.

Markson, L., & Bloom, P. (1997). Evidence against a dedicated system for word learning in children. *Nature*, *385*, 813–815.

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* (pp. 319–326).

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

Pearl, L. (2014). Evaluating learning strategy components: Being fair. *Language*, *90*(3), e107–e114.

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)

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

Phillips, L. (2015). *The role of empirical evidence in modeling speech segmentation* (Unpublished doctoral dissertation). University of California, Irvine.

Phillips, L., & Pearl, L. (2012). 'Less is More' in Bayesian word segmentation: When cognitively plausible leaners outperform the ideal. In *Proceedings of the 34th Annual Conference of the Cognitive Science Society* (pp. 863–868).

Phillips, L., & Pearl, L. (2014a). Bayesian inference as a cross-linguistic word segmentation strategy: Always learning useful things. In *Proceedings of the Computational and Cognitive Models of Language Acquisition and Language Processing Workshop.*

Phillips, L., & Pearl, L. (2014b). Bayesian inference as a viable cross-linguistic word segmentation strategy: It's all about what's useful. In *Proceedings of the 36th annual conference of the cognitive science society* (p. 2775-2780). Quebec City, CA: Cognitive Science Society.

Phillips, L., & Pearl, L. (2015a). Utility-based evaluation metrics for models of language acquisition: A look at speech segmentation. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics 2015.* NAACL.

Phillips, L., & Pearl, L. (2015b). The utility of cognitive plausibility in language acquisition modeling: Evidence from word segmentation. *Cognitive Science*, *39*(8), 1824–1854.

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.

Rollins, P. (2003). Caregiver contingent comments and subsequent vocabulary comprehension. *Applied Psycholinguistics*, *24*, 221–234.

Rose, S. A., Feldman, J. F., & Jankowski, J. J. (2001). Visual short-term memory in the first year of life: Capacity and recency effects. *Developmental Psychology*, *37*(4), 539–549.

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.

Smith, L., & Yu, C. (2008). Infants rapidly learn word-referent mappings via cross-situational statistics. *Cognition*, *106*(3), 1558–1568.

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.

Teinonen, T., Fellman, V., Näätänen, R., Alku, P., & Huotilainen, M. (2009). Statistical language learning in neonates revealed by event-related brain potentials. *BMC neuroscience*, *10*(1), 21.

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

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.

Tincoff, R., & Jusczyk, P. W. (1999). Some beginnings of word comprehension in 6-month-olds. *Psychological Science*, *10*(2), 172–175.

Tincoff, R., & Jusczyk, P. W. (2012). Six-month-olds comprehend words that refer to parts of the body. *Infancy*, *17*(4), 432–444.

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)

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.

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

Xu, F. (2002). The role of language in acquiring object kind concepts in infancy. *Cognition*, *85*(3), 223–250.

Yu, C., & Smith, L. B. (2007). Rapid word learning under uncertainty via cross-situational statistics. *Psychological Science*, *18*(5), 414–420.

Yu, C., & Smith, L. B. (2011). What you learn is what you see: using eye movements to study infant cross-situational word learning. *Developmental Science*, *14*(2), 165–180.