Evaluating language acquisition strategies: A cross-linguistic look at early segmentation

L. Phillips[a,*], L. Pearl[a]

[a]*Department of Cognitive Sciences, University of California, Irvine, 3151 Social Sciences Plaza, Irvine, CA 92697*

Abstract

Language acquisition strategies are typically intended to work for any language a child might need to learn. Yet, these strategies are rarely evaluated on multiple languages. Demonstrating cross-linguistic success is especially important for early acquisition strategies, when children have minimal knowledge of their native language to aid learning. We focus on the early stages of speech segmentation as a case study, evaluating two approaches previously shown to be successful for English: a Bayesian strategy and a subtractive segmentation strategy. These are tested against child-directed speech from a set of seven languages differing in the richness of their morphological systems and how inherently ambiguous they are for segmentation. We find that both strategies can succeed cross-linguistically, though the subtractive segmentation strategy requires a more nuanced evaluation metric involving "reasonable errors" to show this successful performance. This suggests the Bayesian strategy is more robust, though both strategies may be viable universal early segmentation strategies. Because linguistic structure affects error type and error quantity for each strategy, we find that the strategies also generate different empirically testable predictions about infant cross-linguistic segmentation behavior.

Keywords: Bayesian learning, computational modeling, cognitive plausibility, cross-linguistic, language acquisition, speech segmentation, statistical learning

*Corresponding author

Email addresses: lawphill@uci.edu (L. Phillips), lpearl@uci.edu (L. Pearl)

1. Evaluating language acquisition strategies

An important goal of language acquisition research is to understand the strategies children use to learn language as quickly and as well as they do. Computational modeling is a powerful tool for deciding whether a particular strategy can actually work (Pearl, 2014; Kol et al., 2014), and has been used to investigate acquisition tasks as diverse as identifying meaningful sound distinctions (Feldman et al., 2009; Dillon et al., 2013), segmenting words in fluent speech (Brent, 1999; Goldwater et al., 2009), discovering syntactic categories (Wang and Mintz, 2008), learning the mapping between word form and meaning (Frank et al., 2009), and identifying constraints on syntactic structure (Perfors et al., 2011b; Pearl and Sprouse, 2013). A computational model requires that both the acquisition task and the learning strategy be made explicit, and then provides a concrete testing ground for that learning strategy on that particular acquisition task.

One property of the earliest language learning strategies is that they should be *universal*, in the sense that children can successfully use these strategies on any language they might encounter – and importantly, before they have learned much else about their native language. Such strategies have been called *language-independent* (e.g., Gambell and Yang, 2006; Goldwater et al., 2009; Pearl et al., 2011; Phillips and Pearl, in press), since they do not require derived language-specific knowledge in order to be implemented. This contrasts with *language-dependent* strategies, which are often more reliable but which require children to already know something about the structure of their language. So, language-independent strategies are generally proposed for tasks occurring when children first begin acquisition of their native language, as children at that stage do not yet have (much) language-specific knowledge.

For example, consider the task of carving up the fluent speech stream into useful units, typically called *segmentation*. A segmentation strategy that tracks the probabilities between

syllables would be language-independent, because the way these probabilities are calculated and used to segment an utterance does not vary across languages. Infants typically prefer this type of language-independent segmentation strategy around seven months (Thiessen and Saffran, 2003). In contrast, a segmentation strategy that utilizes language-specific stress patterns is necessarily language-dependent, since the way the strategy is implemented depends on how the language uses stress. For example, English words tend to begin with a stressed syllable, and so a stressed syllable would indicate a word beginning. In contrast, French words tend to end with stressed syllables, and so a stressed syllable would indicate a word ending. Infants typically prefer this type of language-dependent segmentation strategy around nine months (Thiessen and Saffran, 2003). Importantly, there must be some language-independent strategy, such as statistical learning, which yielded the appropriate stress information that allows infants to use the stress-based, language-dependent strategy by nine months.

Notably, for any strategy intended to be universal, we must evaluate whether it actually works across the vast spread of human languages. Because languages can vary quite drastically on what they consider a word, evaluating whether a segmentation strategy has succeeded can be difficult. This variation suggests that what constitutes "useful" segmentation output may vary from language to language. Crucially, the output of early learning strategies may *not* be equivalent to the adult linguistic representation, since language acquisition typically occurs in stages. That is, the result of the early learning strategies infants use is unlikely to be adult-level knowledge.

It is with these issues in mind that we investigate the early stages of speech segmentation. We focus on two segmentation strategies, both of which have been successful in English and are intended as universal early learning strategies: (i) a Bayesian segmentation strategy (Goldwater et al., 2009; Pearl et al., 2011; Phillips and Pearl, 2015, in press), and (ii) a subtractive segmentation strategy (Lignos and Yang, 2010; Lignos, 2011, 2012). We first review relevant

aspects and empirical data about the speech segmentation task, including the developmental trajectory, the output of the process, and previous approaches to cross-linguistic evaluation of segmentation strategies. We then briefly describe the relevant details of the segmentation strategies we evaluate here, including the idealized and more cognitively plausible implementations of the Bayesian strategy. Following this, we review the linguistic properties of seven languages that we evaluate these strategies on (English, German, Spanish, Italian, Farsi, Hungarian, and Japanese), as well as the details of the child-directed speech corpora we use as input. These languages vary both in how inherently ambiguous they are to segment as well as in the richness of their morphological systems, among many other aspects that may impact segmentation. We subsequently discuss components of the input representation and output evaluation that are particularly relevant for assessing cross-linguistic segmentation success.

Our results suggest that the Bayesian segmentation strategy does well across this test set of seven languages regardless of how the output is evaluated, while the subtractive segmentation strategy does well across all languages only when a more generous evaluation metric is used. To our knowledge, these are the only language-independent strategies that have been explicitly evaluated on such a wide range of languages and found to succeed. Thus, both strategies gain some support as viable universal early segmentation strategies. We discuss the factors in each language that impact segmentation performance, and how the segmentation strategies are able to deal with each of them. Notably, both strategies tend to produce a significant number of "reasonable errors", particularly in languages other than English. This reinforces the methodological point made by Phillips and Pearl (in press) about what early segmentation success could look like, especially when considering that the useful units a child could segment out of an utterance may differ across the world's languages.

2. Speech segmentation

Speech segmentation, also known as word segmentation, is the process of separating fluent speech into meaningful units, usually thought of as words (e.g., /ðəpɛŋgwɪnzhuwɑɹɹilikjutɑ ɹswɪmɪŋ/ broken into /ðə pɛŋgwɪnz huw ɑɪ ɹili kjut ɑɪ swɪmɪŋ/ = *the penguins, who are really cute, are swimming*). Notably, the speech stream largely lacks consistent pauses at word boundaries (Cole and Jakimik, 1980), though child-directed speech tends to have pauses at clause boundaries (e.g., after *cute* above) which may aid segmentation (Broen, 1972; Hirsh-Pasek et al., 1987). Speech segmentation is a fundamental acquisition task because so many other tasks depend on segmented units, including learning lexical meaning, grammatical categories, morphology, and syntax.

2.1. Cues to segmentation

Infants are known to rely on a variety of cues when segmenting. Many of these are language-dependent and are leveraged between the ages of seven and a half to nine months, such as metrical stress patterns (7.5 months: Jusczyk et al., 1999c; 9 months: Morgan and Saffran, 1995), coarticulation effects (8 months: Johnson and Jusczyk, 2001), phonotactic cues (9 months: Mattys et al., 1999), and allophonic variation (9 months: Jusczyk et al., 1999b). Of course, to use these language-dependent cues, infants must have already learned something about their native language's words, which suggests that segmentation using language-independent cues occurs before seven and a half months.

We know that six-month-olds are able to segment new words by taking advantage of highly familiar, adjacent words, such as *mommy* or their own names (Bortfeld et al., 2005). A plausible developmental trajectory is that infants use language-independent cues between approximately six and seven and a half months in order to discover the language-dependent cues they rely on afterwards. Because very young infants are capable of segmenting words based on differ-

ent probabilistic relationships between syllables (Saffran et al., 1996; Aslin et al., 1998; Johnson and Jusczyk, 2001; Thiessen and Saffran, 2003; Pelucchi et al., 2009), a candidate for a general-purpose language-independent cue is some measure of the association between salient perceptual units like syllables.

2.2. Differences across languages

One challenge for any segmentation strategy is capturing the diversity of the useful units that could be segmented from fluent speech. While the target of segmentation is traditionally thought of as words, morphemes are actually the smallest meaningful units of a language, and languages vary both in how they relate morphemes to words and how they relate words to phrases. Isolating languages such as Mandarin, Vietnamese, and English represent one end of the spectrum, where a sentence like *I like the penguin* is made up of four morphemes that are each their own word. On the other side, polysynthetic languages like Greenlandic, Nuhuatl, and Siberian Yupik can have a single word which is equivalent to an entire sentence in another language (e.g., the Siberian Yupik word *angya-ghlla-ng-yug-tuq* can be translated into English as *He wants to acquire a big boat* (Comrie, 1989)).

What then is the appropriate target of early segmentation? If a learning strategy generates morphemes rather than words, this may be quite useful to the child anyway (especially for polysynthetic languages), since knowledge about morphemes is necessary to make full sentences. If a learning strategy generates morpheme sequences that correspond to syntactic frames, such as *I like the* in English, this could also potentially be useful (especially for isolating languages with common collocations).

One possible solution is to consider whether the segmentation strategy has produced output that is "reasonable" or "good enough". Since the language-dependent cues infants rely on after seven and a half months are likely derived from their initial language-independent segmentation output, we suggest that a language-independent strategy should generate a set of units that is

reliable enough to infer the language-specific instantiations of cues like metrical stress, coarticulation effects, phonotactics, and allophonic variation. If the strategy can do so, this is "good enough" for the initial segmentation occurring between six and seven and half months.

2.3. Universal segmentation strategy properties

While we propose that the ability to generate "good enough" output is one desirable property for a universal language-independent strategy, there are others that are also important. We believe that a universal language-independent strategy should satisfy all of the following:

1. *realistic input*: The strategy should be successful when operating over the input that infants use.
2. *usable:* The strategy should be deployable by infants, who have cognitive limitations on their inference process.
3. *useful output:* The strategy should generate output which is good enough for later learning processes.
4. *cross-linguistic success:* The strategy should generate useful output across all languages.

In terms of realistic input, essentially all models of segmentation strategies attempt to learn from realistic child-directed speech, although there is considerable debate about the unit of representations that infants perceive (for an overview of the literature, see Jusczyk (1997) and Phillips and Pearl (in press)).

Usable also seems like a necessary property for a viable segmentation strategy, though often the first step for investigating a strategy is implementing an ideal learner model without cognitive limitations (e.g., Johnson, 2008b; Frank et al., 2009; Goldwater et al., 2009; Dillon et al., 2013; Feldman et al., 2013). This makes a strategy's deficits apparent even before a more realistic inference process is used. In particular, if the learning strategy *already* struggles (or fails) when inference is optimal, there is likely to be a bigger problem once inference is more realistic

(Pearl, 2011). Still, we believe the ultimate goal is for a strategy to succeed when inference is constrained, and several models have evaluated strategies using an incremental learning process (Brent, 1999; Venkataraman, 2001; Blanchard et al., 2010) and implemented constrained, cognitively-inspired inference algorithms (Pearl et al., 2011; Lignos, 2011).

Interestingly, the utility of segmentation output has only recently been explicitly recognized as a reasonable metric for segmentation success as compared to matching the adult representation (Phillips and Pearl, 2015, in press) . This is most likely because it can be difficult to evaluate what "useful' means in a principled way, while a comparison against adult segmentation (typically, orthographic text where a word is delineated by spaces) is easy to reach consensus about. Still, given that universal language-independent strategies are likely being used between six and seven and a half months, achieving a useful output seems a more reasonable metric than adult words.

Cross-linguistic success has also had significant attention devoted to it (e.g., Batchelder, 2002; Fleck, 2008; Johnson, 2008a; Blanchard et al., 2010; Fourtassi et al., 2013). Evaluation, however, has only occurred over a small number of languages at a time for any given segmentation strategy (typically two or three, where one of the test languages is English[1]). We review these briefly below. The main point, however, is that it is unclear that the *cross-linguistic* property truly holds for any currently proposed segmentation strategy.

[1]English's popularity is likely due to the previous availability of English child-directed speech corpora reasonably close to the age range appropriate for word segmentation, such as the Bernstein-Ratner corpus (Bernstein-Ratner, 1984) and the UCI Brent Syllables corpus (Phillips and Pearl, in press), compared with the relative scarcity of such resources in other languages. Notably, the derived corpora section of CHILDES (*http://childes.psy.cmu.edu/derived/*) is becoming a much better resource for non-English data that can be utilized to evaluate segmentation strategies, and currently includes derived corpora with phonemic representations of Sesotho, Hungarian, Italian, and Polish.

2.4. Cross-linguistic evaluation for segmentation strategies

2.4.1. Previously evaluated segmentation strategies

Several segmentation models, such as WordEnds (Fleck, 2008) and PHOCUS (Blanchard et al., 2010), use a phonotactic strategy, leveraging the transitional probability of phonemes to identify where word boundaries are. The idea behind these strategies is that there are common phoneme sequences that begin and/or end words in many languages (e.g., [pl] begins words in English while [lp] does not). While both WordEnds and PHOCUS seem to perform well enough on English data from the commonly used Bernstein-Ratner corpus (Bernstein-Ratner, 1984), their performance suffers noticeably on Spanish and Arabic (WordEnds) or Sesotho (PHOCUS). However, it should be noted that only the Sesotho corpus was composed of child-directed speech, while the Spanish and Arabic corpora were derived from adult-directed telephone conversations. This could well impact an early segmentation strategy's success, as adult-directed speech typically has more words per utterances and more lexical items than child-directed speech. More generally, phonotactic strategies will likely not be very useful for languages with very simple phonotactics, such as languages which only allow for CV syllable structure (i.e., consonants are never allowed after a vowel within a syllable). These include languages such as Hawaiian, Malagasy, Maori, Sesotho, Swahili, Tahitian, and Yoruba. For these languages, there are no sound sequences such as /lp/ or /pl/ that might indicate a possible word boundary.

Several other statistical segmentation strategies are implemented with the explicit goal of inferring a lexicon of word forms that explains the observed input data, with learners segmenting utterances based on that inferred lexicon (MDBP-1: Brent, 1999; n-gram learners: Venkataraman, 2001; BootLex: Batchelder, 2002; Adaptor grammars: Johnson et al., 2007; Bayesian: Goldwater et al., 2009). To infer a lexicon, a learner using this kind of strategy typically leverages statistical regularities and underlying assumptions about how the observed language data are generated. For English, many of these strategies perform well when evaluated against child-

directed English data (MDBP-1, n-gram: Boruta et al., 2011; Adaptor grammars: Johnson, 2008b; Bayesian: Goldwater et al., 2009). However, several have significantly poorer performance when evaluated against languages like French (MDBP-1, n-gram: Boruta et al., 2011), Japanese (MDBP-1, n-gram: Boruta et al., 2011), Spanish (Bayesian: Fleck (2008)), and Arabic (Bayesian: Fleck (2008)). As before, part of the issue may be that some of the language data were adult-directed rather than child-directed speech samples, as with the Spanish and Arabic data. Another issue concerns the input representation, which we discuss in more detail in section 5. In particular, all these evaluations assume the input is a stream of phonemes, which may affect the difficulty of the segmentation task. Nonetheless, the adaptor grammar approach fares well when evaluated on child-directed Sesotho (Johnson, 2008a), Chinese (Johnson and Demuth, 2010), and Japanese (Fourtassi et al., 2013), though its good performance is predicated on choosing the correct generative language model for each language.

2.4.2. Evaluation issues

One obvious concern with English as the primary evaluation language is that a given strategy may be successful on English due to factors that will not hold for all the world's languages (as some of the previous evaluations demonstrate). For instance, there may be structural aspects that are beneficial for English, such as inherently less segmentation ambiguity (Fourtassi et al., 2013), a lack of complex morphological structure, or a high ratio of monosyllabic words. Also, the orthographic segmentation typically used as the target segmentation may be biased by a language's orthographic tradition (e.g., treating compound words as separate words: *truck driver, ice cream*, and *book lover* in English). This can affect the difficulty of segmentation. So, the success or failure of a strategy may be caused by the particular idiosyncrasies of the language being evaluated.

More generally, with results from only a few languages, any number of possible explana-

tions may fit the observed segmentation strategy behavior. That is, if we think of the evaluation on each language as an individual data point, evaluating only two or three languages means that there may be many different explanations which match those two or three data points. For this reason, testing a variety of languages seems wise, with the hope that the language set will provide a better range of the linguistic aspects impacting the segmentation task. Then, spurious explanations which might have seemed reasonable when tested against only a few languages are more likely to be found inadequate. More importantly, if a segmentation strategy succeeds on a variety of languages, we can feel more confident that the strategy is robust to the linguistic variation a child may encounter in whatever language she is trying to segment, and so has the desired cross-linguistic property. For this reason, one important methodological contribution we make is a collection of derived child-directed speech corpora in several additional languages that are in a format segmentation models can utilize.[2]

With this general evaluation issue in mind, we focus on cross-linguistically evaluating two segmentation strategies that have previously been very successful for child-directed English data, satisfying *realistic input* (Bayesian: Goldwater et al., 2009; Phillips and Pearl, in press; Subtractive: Lignos, 2011, 2012) and *usable inference* (Bayesian: Pearl et al., 2011; Phillips and Pearl, 2012, in press; Subtractive: Lignos, 2011, 2012). The Bayesian strategy additionally is known to satisfy *useful output* for English (Phillips and Pearl, 2014, 2015, in press), and the subtractive segmentation strategy does so as well, though less robustly (Phillips and Pearl, 2015). We will demonstrate that both can also satisfy *cross-linguistic success*, based on our test set of seven languages, and that both additionally yield *useful output* of various kinds across languages. This makes these the only strategies to our knowledge that have been explicitly found to satisfy each of the four properties of a viable universal early segmentation strategy.

[2] Available at *http://github.com/lawphill/PhillipsPearl_Corpora/* and to be released to CHILDES.

3. Segmentation strategies

3.1. Bayesian segmentation

The Bayesian segmentation strategy we investigate is one version of a statistical learning strategy that does not rely on language-dependent information. This makes it a good candidate for the early stages of word segmentation occurring before seven and a half months, when an infant does not yet know (m)any words of the language.

One benefit of Bayesian learning strategies is that they explicitly distinguish between the learner's pre-existing beliefs (the *prior*: $P(h)$) and how the learner evaluates incoming data (the *likelihood*: $P(d|h)$). This information is combined using Bayes' theorem (1) to generate the updated beliefs of the learner (the posterior: $P(h|d)$). Bayesian models take advantage of this distinction in order to make a trade-off between fit to the data and knowledge generalizability (Perfors et al., 2011a).

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

3.1.1. The underlying model

The Bayesian segmentation strategy was originally described by Goldwater et al. (2009) (**GGJ** henceforth), and can be implemented by a learner using a generative model of how observed utterances are produced. In particular, the observed utterances are perceived as sequences of word tokens drawn from an underlying lexicon of word forms, and the Bayesian learner infers this lexicon of word forms. The generative model explicitly encodes how words in observed utterances are produced. Given the limited knowledge of language structure which infants likely possess at the relevant age, GGJ described two simple generative models.

The first model assumes independence between words (a *unigram* assumption) – the learner effectively believes word tokens are randomly selected with no relation to each other. To encode

this assumption in the model, GGJ use a Dirichlet Process (Ferguson, 1973), which supposes that the observed sequence of words $w_1 ... w_n$ is generated sequentially using a probabilistic generative process. In the unigram case, the identity of the *i*th word is chosen according to (2-3), where the probability of the current word is a function of how often it has occurred previously.

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

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

$n_{i-1}(w_i)$ is the number of times word w_i appears in the previous $i - 1$ words, α is a free parameter of the model which encodes how likely a novel word is to be generated, and P_0 is a base distribution (3) specifying the probability that a novel word will consist of particular units (e.g., phonemes or syllables) $x_1 ... x_m$.

P_0 can be interpreted as a parsimony bias, giving the model a preference for shorter words. This is because the more units that comprise a word, the smaller the probability of that word, and so shorter words are more probable. α can be interpreted as controlling the bias for the number of unique lexical items in the corpus, since α controls the probability of creating a new word in the lexicon. For example, when α is small, the learner is less likely to hypothesize new words to explain the observable corpus data, and so prefers fewer unique items in the lexicon. We note that the model relies on the perceived frequency of lexical items to make its decisions – that is, it does not know the true word frequencies (this is what it is trying to learn in the first place), but it estimates these based on the number of times it believes the word has appeared previously.

The second generative model makes a slightly more sophisticated assumption about the

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

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

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

$n_{i-1}(w', w_i)$ is the number of times the bigram (w', w_i) has occurred in the first $i-1$ words, $n_{i-2}(w')$ is the number of times the word w' occurs in the first $i-2$ words, $b_{i-1}(w_i)$ is the number of bigram types which contain w_i as the second word, b is the total number of bigram types previously encountered, P_0 is defined as in (3), and β and γ are free model parameters. Both the β and γ parameters, similar to the α parameter in (2), control the bias towards fewer unique bigrams (β) and towards fewer unique lexical items as the second word in a bigram (γ). Like the unigram model, the bigram version tracks the perceived frequency of lexical items, as well as the frequency of bigram pairs. In addition, it relies on the transitional probabilities between words that comprise bigrams, due to (4). In particular, a word w_i which has more frequently followed w' (i.e., has a high transitional probability) is more likely to be generated in bigrams that begin with w'.

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

3.1.2. Bayesian inference

3.1.2.1. Idealized inference. Learners implementing an idealized inference process are often used to determine if a model's learning assumptions are useful for a child (Johnson, 2008b; Frank et al., 2009; Goldwater et al., 2009; Dillon et al., 2013; Feldman et al., 2013). The basic intuition is that useful learning assumptions will lead to to the correct acquisition outcome when inference is optimal. If that is shown to be true, then we can proceed to less-than-optimal inference, such as the process likely occurring in infants which is constrained by their cognitive limitations. However, if a learning assumption is not useful even when inference is unhindered, then it is unlikely to be useful when realistic inference occurs. So, idealized inference can be an informative first step for testing the learning assumptions encoded in a model.

GGJ originally used ideal learners to evaluate the usefulness of the Bayesian segmentation strategy on English. In general, ideal learners use an inference algorithm guaranteed to converge on the best hypothesis that balances the learner's prior beliefs (as encoded by the model) with the likelihood of the observed data. The particular inference algorithm used for GGJ's ideal learners was Gibbs sampling (Geman and Geman, 1984), which is a Markov Chain Monte Carlo (MCMC) algorithm.

MCMC algorithms provide a way to search the hypothesis space defined by a model and identify the best hypothesis, given the observable data. This kind of algorithm operates by first guessing a value for every latent variable in the model (for the segmentation strategy, this corresponds to whether a word boundary exists between any two syllables). Then, one at a time, each variable's value is chosen anew, conditioned on the current values of all the other variables, which is called *sampling* a boundary. So, for this segmentation strategy, each potential word boundary is sampled one at a time, based on the current values of all the other potential word boundaries. This sampling process is repeated for all variables (here, all potential word boundaries), and a complete cycle through all variables (here, all potential word boundaries) is

a single iteration of the algorithm. Many iterations are necessary to identify the best hypothesis, with GGJ using 20,000 iterations for their ideal learners.

Clearly, this MCMC-based inference process is unlikely to be representative of the way humans implement inference, since humans are unlikely to remember a large batch of data with the precise detail required for this kind of iterative learning process. But the core question of whether the strategy is useful can be answered effectively – with the learning assumptions of this model, is good segmentation possible? For English, the answer seems to be yes (Goldwater et al., 2009; Phillips and Pearl, 2015, in press). Because this ideal learner processes the input in a batch and finds what it considers the optimal segmentation, we refer to it as the **BatchOpt** learner.

3.1.2.2. Usable: Constrained inference. To satisfy the *usable* criterion, we need to implement learners who have constraints on their inference, due to limitations on their memory and processing abilities. We consider the algorithms investigated by Pearl et al. (2011) that incorporate several different cognitively-inspired constraints which potentially better match the inference process in infants.

The first constrained algorithm simply makes the inference algorithm incremental (sometimes called *online*), so that data are processed as soon as they are encountered, rather than being saved for batch processing at some later point. As this learner processes each utterance, it attempts to identify what it considers the optimal segmentation, which is the one with the highest probability given the data already seen. This learner was implemented with a dynamic programming technique called the Viterbi algorithm[3] which allows quick identification of the hypothesis that is optimal given current knowledge. Because this learner uses an online process to identify the locally optimal segmentation of an utterance, we refer to it as the **OnlineOpt**

[3]We note that the unigram version of this learner is essentially equivalent to the incremental learner of Brent (1999) and was referred to by Pearl et al. (2011) as the Dynamic Programming with Maximization (DPM) learner.

learner.

The second constrained inference algorithm is similar to the OnlineOpt learner in processing the data as they are encountered, but relaxes the requirement that the best segmentation hypothesis be chosen each time. Instead, a dynamic programming technique called the Forward algorithm is used to quickly calculate segmentation hypothesis probabilities, and a single hypothesis is then sampled based on the calculated probabilities.[4] For example, if there are two possible segmentations A and B, and the calculated probability of A is 0.75 while the calculated probability of B is 0.25, this learner will choose segmentation A 75% of the time and segmentation B 25% of the time. In essence, the learner is allowed to explore hypotheses that are less probable given current knowledge, and this is an advantage since local decisions that are sub-optimal can lead to hypotheses that are optimal once more information is available. Notably, hypothesis sampling (as opposed to always choosing the highest probability hypothesis) is consistent with evidence about how young children make decisions (Denison et al., 2013), and so could be viewed as more cognitively plausible.[5] Because this learner uses an online process to identify a segmentation of an utterance that may be locally sub-optimal, we refer to it as the **OnlineSubOpt** learner.

The third constrained algorithm maintains the online processing constraint while also implementing a recency effect, which can be viewed as a kind of short-term memory limitation. In particular, this learner focuses its processing resources on more recent data, rather than giving all data equal attention. It is actually most similar to the BatchOpt learner, as it uses a version of Gibbs sampling to converge on a segmentation. However, it uses Decayed Markov Chain Monte

[4]We note that this was referred to as the Dynamic Programming with Sampling (DPS) learner in Pearl et al. (2011).

[5]This inference process is similar to a single-particle particle filter, where a probability distribution is represented as a single point estimate. Interestingly, Sanborn et al. (2010) demonstrate that a single-particle particle-filter approximates human category judgment patterns quite reasonably, though the underlying model is different from the segmentation model here.

Carlo sampling (Marthi et al., 2002)[6] so that it is biased to sample more recent potential word boundaries, rather than all potential word boundaries equally. In particular, for every utterance, this learner samples some number s of previous potential word boundaries. We follow Phillips and Pearl (in press) and set s to 20,000 for the simulations reported below; Phillips and Pearl (in press) note that this amounts to 74% less sampling than the equivalent BatchOpt learner using regular Gibbs sampling and so represents a significant processing reduction. The probability of sampling a potential boundary b is proportional to the decay function b_a^{-d}, where b_a is the number of potential boundary locations between b and the end of the current utterance ("how many **b**oundaries **a**way from the end") and d is the decay rate. So, the farther away a potential boundary is from the end of the current utterance, the less likely it is to be sampled. Larger values of d indicate a stricter memory constraint. Following Phillips and Pearl (in press), we use a set, non-optimized d of 1.5, which implements a heavy memory constraint: 86.3% of the potential boundary samples occur in the current utterance, 11.8% occur in the previous utterance, and only 4.6% occur in utterances prior to that (predominantly the next-most-previous utterance). Intuitively, this can be interpreted as the learner having the current utterance in memory, and some decaying version of previously heard utterances. So, a decision made about the segmentation of the current utterance can have some limited effect on previously heard utterances, depending on how recently heard they were.[7] Because of this imposed memory limitation, we refer to this third online learner as the **OnlineMem** learner.

3.1.2.3. Learner summary. Table 1 summarizes the differences between the different Bayesian learners we investigate: one ideal learner to test whether the assumptions of the learning model

[6]Which is why this learner was referred to as the DMCMC learner by Pearl et al. (2011).

[7]We note that the other online learners can also be interpreted as having some kind of recency effect because they are constrained to process *only* the most recent utterance. Notably, however, the potential boundaries in those utterances are equally likely to be examined, while this third learner is biased to prefer potential boundaries near the end of the current utterance.

are useful and three constrained learners to test whether the assumptions of the learning model are usable. Of the three constrained learners, we suggest that the OnlineMem learner is the most cognitively plausible since it incorporates both online processing and a type of short-term memory. So, we will be particularly interested in the performance of the modeled learner using this inference procedure, since it is the one we believe is most realistic of the ones investigated and therefore most informative for understanding infant word segmentation. Nonetheless, the OnlineSubOpt learner also incorporates online processing and probabilistic sampling, and so provides a second constrained inference checkpoint for the strategy's usability.

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

Table 1: Summary of modeled learners used for Bayesian segmentation, including the relevant parameters and learning assumptions encoded by each. Parameters were set individually by language after a non-exhaustive search that maximized the word token F-score performance of the BatchOpt learner. All learners set $\alpha = 1$ and all unigram learners set $\beta = 1$. Bigram learners had the following parameter values: English, $\beta = 1, \gamma = 90$; German, $\beta = 1, \gamma = 100$; Spanish, $\beta = 200, \gamma = 50$; Italian, $\beta = 20, \gamma = 200$; Farsi, $\beta = 200, \gamma = 500$; Hungarian, $\beta = 300, \gamma = 500$; Japanese, $\beta = 300, \gamma = 100$.

3.2. Subtractive segmentation

Heuristic segmentation strategies like those investigated by Yang and Lignos (Gambell and Yang, 2006; Lignos and Yang, 2010; Lignos, 2011, 2012) typically rely on some amount of prior linguistic knowledge. For example, these strategies often rely on bootstrapping from previously identified words (similar to the six-month-olds in Bortfeld et al. (2005)) and/or knowledge that words contain a single primary stress. This has led to considerable success for English child-directed input data, with better segmentation performance than other probabilistic segmentation

strategies.

The particular subtractive segmentation learner we investigate is the most successful one described by Lignos (2011) which does not utilize stress information – this strategy thus has access to the same probabilistic cues as the Bayesian strategy. A learner using this strategy processes the corpus one utterance at a time, making this an online strategy. The learner begins by assuming that every utterance is a single word. Then, as the learner adds vocabulary to its lexicon, it can segment out these known words when possible. Thus, it can capitalize on the occurrence of short utterances containing a single word (e.g., *nice!*) – since these actually *are* words, they appear in longer utterances and can be "subtracted" out of the longer utterances.

When there are multiple possible segmentations for an utterance, the Lignos learner variant we implement uses beam search to sort through different segmentation options. In particular, for each potential word boundary that is ambiguous (e.g., *bunny ? in*), the learner considers two possible segmentations: one with the boundary present (*bunny in*) and one with the boundary absent (*bunnyin*). The segmentation selected is the one with a higher score, where a segmentation's score is the geometric mean of the score of each word in the potential segmentation (*bunny* and *in* vs. *bunnyin*). A word's score is determined by two factors: (i) its frequency in previously inferred segmentations, and (ii) how often it has been part of a potential segmentation that was previously rejected.

One key advantage of beam search is that it gives the learner a way to avoid making decisions that are solely based on what seems locally optimal, given the word form's frequency. This is very important, given that the early inferred lexicon is rife with undersegmentation errors due to how the learner hypothesizes words in the first place (based on entire utterances). For example, suppose that an undersegmentation error *bunnyin* has entered the hypothesized lexicon at an early stage of learning and the learner is considering the potential word boundary between *bunny* and *in* in another utterance. If the learner cared only for high frequency word

forms, the optimal choice would be to not segment *bunny* out, since *bunnyin* has a higher inferred frequency than *bunny* at this stage of learning. A greedy learner focused exclusively on locally optimal choices would thus persist in the *bunnyin* undersegmentation error.

However, the learner using beam search considers not only the frequency of the word to be segmented out, but also the frequency of all words involved in the segmentation (here: both *bunny* and *in*). In addition, by considering how often a word form has been part of a rejected segmentation, it can downweight the effect of word form frequency. Both of these allow it to capitalize on the frequency and prior success of *in*, even if *bunny*'s score is relatively low. In this way, the undersegmentation error *bunnyin* can be recovered from fairly quickly.

4. Languages

The language data we use to evaluate the two segmentation strategies are from a variety of languages that have child-directed speech data available through the CHILDES database (MacWhinney, 2000). Notably, the task of collecting an appropriate cross-linguistic range of input data is non-trivial, even with this vast database. This is due primarily to two factors. First, even with its broad collection of child-directed speech input, the CHILDES database does not have a large number of corpora directed at children under one year of age, which is when early segmentation strategies would be in use. Second, even when age-appropriate data are available, they are often available as orthographic transcripts, rather than the representation that would be perceived by infants of this age (e.g., phonetic representations of syllables – see discussion in section 5).

So, in order to use such data as input to a segmentation model, researchers must convert the orthographic representation into a more appropriate input representation (e.g., "what a pretty kitty" converted to a stream of syllables like /wʌt ə pɪɪ ɾi kɪ ɾi/). This is relatively straightforward for words appearing in syllabified pronunciation dictionaries such as the MRC or Callhome

databases (Wilson, 1988; Kingsbury et al., 1997), but becomes more difficult for words that are not, such as the "motherese" words often found in child-directed speech (e.g., *fishies, binkies, dada's*). For these words, the knowledge of a native speaker trained in linguistic representation is often required. For languages where no syllabified pronunciation dictionary is available, a fluent speaker may be the only resource.

Fortunately, some derived corpora already exist in CHILDES that provide the phonemic transcription or the syllabified phonemic transcription of child-directed speech corpora in some languages (e.g., phonemic: Sesotho, Polish; syllabified phonemic: Hungarian, Italian). We add to these derived corpora by creating syllabified phonemic transcriptions of several additional languages: Japanese, Farsi, Spanish, and German. These transcriptions were created by relying on pronunciation dictionaries and linguistically trained fluent speakers. For syllabification, we primarily used adult syllabification judgments, but when these were unavailable, syllabification was done using the maximum-onset principle (Selkirk, 1981). This principle states that the beginning of any syllable should contain as many sounds as possible so long as phonotactic principles in the language are not violated. For instance, in the English word *onset*, the maximum-onset principle would suggest a syllabification of *on set* over *o nset* or *ons et*. The largest onset *ns* can be ruled out because it violates English phonotactic rules. This leaves either *s* or $\emptyset$ as possible onsets. The maximum onset rule states that the larger onset should be preferred (*s*), which produces the English syllabification *on set*. Although adult judgments sometimes violate the maximum-onset principle, it is a simple guideline which is able to accurately describe a great deal of syllabification across the world's languages.

In choosing our languages, we have attempted to use corpora which are (i) speech directed at children of the appropriate age range, (ii) from languages that differ in their linguistic structure as well as their historical background, and (iii) already phonemically transcribed. A brief summary of the corpora selected is given in Table 2. While many of these languages are Indo-

European (Germanic: English, German; Romance: Spanish, Italian; Indo-Iranian: Farsi), two are from other language families: Hungarian is Uralic while Japanese is Japonic. While there are certain structural similarities, especially among the Indo-European sub-branches, these languages differ in a number of important ways, including their syntax, morphology, and phonotactics.

Language	Corpora	Age range	# Utt	# Syl types	Syl TT-ratio	Syls/Utt	B Prob
English	Brent	0;6-0;9	28391	2330	30.03	4.16	0.763
German	Caroline	0;10-4;3	9378	1683	14.43	6.30	0.745
Spanish	JacksonThal	0;10-1;8	16924	524	36.66	4.80	0.513
Italian	Gervain	1;0-3;4	10473	1158	11.32	8.78	0.499
Farsi	Family, Samadi	1;8-5;2	31657	2008	10.92	6.98	0.438
Hungarian	Gervain	1;11-2;11	15208	3029	7.26	6.30	0.512
Japanese	Noji, Miyata, Ishii	0;2-1;8	12246	526	12.15	4.22	0.443

Table 2: Summary of child-directed language corpora and different statistics, 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 ratio of syllable tokens to syllable types (Syl TT-ratio), the average number of syllables per utterance (Syls/Utt), and the probability of a word boundary appearing between syllables (B Prob).

For the purposes of the segmentation strategy evaluation, the morphology variation is salient. In particular, if a language has regular morphology, an infant might reasonably segment morphemes as useful units, rather than words. This would not match the adult orthographic word representation, which would only count complete words as correct. Of the languages here, both Hungarian and Japanese are more agglutinative and so have richer morphology available, while the others are more fusional and so have less regular morphology available.

Additionally, the child-directed speech in these languages differs by how short the words of the language are, as noted by how often a word boundary appears between syllables (B Prob). Languages with more monosyllabic word tokens, such as English and German, have a higher probability of a boundary appearing after a syllable; languages with fewer monosyllabic word tokens, such as Farsi and Japanese, have a lower probability of a boundary appearing after a syllable. This notably affects modeled learners that have a bias towards oversegmentation

(i.e., positing word boundaries where they aren't any, such as splitting *flying* into *fly* and *ing* in English). If a language tends to have shorter words anyway, this error may not be as damaging as it is for a language where the words are often multiple syllables. We return to this aspect in section 7.2, as many of the modeled learners do possess an oversegmentation bias.

5. Realistic input representation

Choosing an appropriate input representation for the modeled learner is a non-trivial decision as the developmental literature has long been concerned with identifying the basic unit of speech perception for infants. While no study to date offers conclusive evidence, there is a general trend suggesting that larger units, such as syllables, are more easily perceived than phonetic segments by infants younger than six months (Jusczyk and Derrah, 1987; Bertonicini et al., 1988; Bijeljac-Babic et al., 1993; Eimas, 1999); in contrast, aspects about phonetic segments seem to be primarily learned after this point (Polka and Werker, 1994; Pegg and Werker, 1997; Jusczyk et al., 1999b; Maye et al., 2008; Jusczyk et al., 1999a). Additionally, although infants are known to track probabilities between syllables before eight months (Saffran et al., 1996; Thiessen and Saffran, 2003; Pelucchi et al., 2009), they do not appear to track probabilities between phonemes (and so learn phonotactics) until after this age (Mattys et al., 1999). Given this experimental evidence, we assume that modeled learners should perceive the input as a stream of syllables.[8] Practically speaking, this means each syllable is treated as an atomic unit, with no recognizable internal structure.

Interestingly, this kind of syllable-based learner can partially address one common concern about phoneme-based learners, which is the lack of allophonic variation in child-directed speech corpora. For instance, both *cat* and *can* are phonemically transcribed as using the vowel /æ/

[8]However, we note that it is possible phonetic learning also affects segmentation (Elsner et al., 2012), as the two processes may coincide. For a more thorough discussion of experimental work on the unit of representation in infants, see Phillips and Pearl (in press).

(/kæt/, /kæn/), even though the vowel in *can* is nasalized, among other phonetic differences ([k^{h}æt$^{\urcorner}$], [k^{h}æ̃n]). Because phonemically-transcribed corpora treat both allophones of /æ/ as identical, a phoneme-based model implicitly assumes that learners have already acquired the full set of allophonic rules in their native language, such as the one that maps [æ̃] to /æ/. Of course, for six or seven-month-old infants, this is unlikely to be true for all sounds, given that they are still perceiving phonetic contrasts that do not appear in their language (Werker and Lalonde, 1988; Werker and Tees, 1984; Pegg and Werker, 1997). In contrast, because each syllable for a syllable-based learner is treated as an atomic unit, allophonic variations within syllables (e.g., between [k^{h}æt$^{\urcorner}$] and [k^{h}æ̃n]) are implicitly captured because they are aready treated as completely distinct units.

However, a syllable-based learner is still susceptible to two concerns that other modeled learners also face, related to phonetic variation. First, it cannot solve the problem of variation that occurs across syllable boundaries (e.g., *did you* produced as *didjou* [dɪdʒu], though it is phonemically transcribed as [dɪdju]). This issue is unlikely to be resolved by any modeled learner until phonetic transcriptions are available for child-directed speech data so that this segmental variation is accurately represented in the learner's input. Second, a syllable-based learner cannot solve the problem of free variation, where speakers may use one of many possible sounds without changing the word's meaning (e.g. *aunt* produced either as [ant], [ænt], or [ɔnt]). From orthographic transcripts alone, it is unknown what form the speaker actually produced. This second concern will also hold for any modeled learner until phonetic transcriptions become available.

6. Output evaluation

A common evaluation metric used by many segmentation studies is comparison to the adult orthographic representation, i.e., the way the words would appear when transcribed in the lan-

guage by an adult speaker. Word token precision and recall are typically used to assess how many of the words identified by the modeled learner were true words (precision, shown in (6)) and how many of the words that should have been identified were identified by the modeled learner (recall, shown in (7)).

$$Precision = \frac{\#\ identified\ true\ word\ tokens}{\#\ all\ identified\ word\ tokens} \tag{6}$$

$$Recall = \frac{\#\ identified\ true\ word\ tokens}{\#\ all\ true\ word\ tokens} \tag{7}$$

These two scores, which range between 0 and 1, are typically combined into a single summary statistic via the harmonic mean, referred to as the F-score (8).

$$F = 2 * \frac{precision * recall}{precision + recall} \tag{8}$$

This evaluation metric has the benefit of being simple to calculate, assuming an orthographic transcript of the data is available, and so is a convenient way to compare the results of modeled segmentation strategies across studies. Still, a known disadvantage is that the target segmentation is assumed to be the adult orthographic segmentation, which is unlikely to be true for six- to seven-month-olds using early segmentation strategies. For example, anecdotal evidence suggests segmentation errors are present even in much older children – two-year-old children produce systematic "chunking" errors such as segmenting *is that a* as *isthata* (Peters, 1983). So, the output of an infant's segmentation strategy likely won't match adult segmentation.[9] Given this, we consider additional metrics with an eye towards what may constitute *useful output* – in particular, "reasonable" errors.

[9]There are additional issues with using the orthographic segmentation, rather than a phonological segmentation. See Blanchard et al. (2010) for discussion of this point.

We suggest that the three reasonable errors Phillips and Pearl (in press) describe are useful output since these errors could produce units that are useful, even if these units are not the words an adult would segment from a given utterance:

1. Oversegmentations that result in real words (e.g., *alright* /ɑl ɹajt/ segmented as *all* /ɑl/ and *right* /ɹajt/)
2. Oversegmentations that result in productive morphology (e.g., segmenting off *–ing* /iŋ/)
3. Undersegmentations that produce function word collocations (e.g., segmenting *is that a* as *isthata*)

The first reasonable error, while an oversegmentation, can nonetheless be beneficial if the segmented words are frequent. For example, a child who oversegments *alright* into *all* and *right* increases her belief that *all* and *right* are words in her language. This ensures she is more likely to correctly segment these words when they actually appear in her input. The second reasonable error, also an oversegmentation, can identify units that will be mapped to meaning later on and can help identify the grammatical category of a word, even though the language represents these units as affixes. For example, the morpheme *–ing* in English indicates imperfective aspect and identifies the word it attaches to as a verb. Both of these reasonable oversegmentations contrast with oversegmentations that yield no useful units, such as segmenting *doggie* (dɑgi/) into *do* (/dɑ/) and *ggie* (/gi/).

The third reasonable error is similar to the second in that its utility is in scaffolding future acquisition processes. Specifically, function word collocations may serve as helpful syntactic or semantic frames that indicate the category of a word that follows them (e.g., *isthata* would precede a noun/countable object like *doggie*). In addition, this undersegmentation error corresponds qualitatively to observed segmentation behavior in toddlers, who identify chunks rather than individual words (Brown, 1973; Peters, 1983).

With this in mind, we evaluate the modeled learners who use the Bayesian segmentation strategy and the subtractive segmentation strategy on both (i) the standard token F-score metric that compares the output against adult segmentation, and (ii) the patterns of reasonable errors generated. The second evaluation metric is particularly helpful for our cross-linguistic analysis, since the units that are useful in a given language may be larger or smaller than an orthographic word in that language.

7. Results and discussion

We analyze our data using train/test cross-validation in order to assess how well learners generalize their inferred segmentation knowledge to new data. In particular, each learner uses the training set to learn what lexicon items tend to appear and how often they tend to appear, and then applies this knowledge as it continues to learn and segment the test set. Five training and test sets were constructed for cross-validation of each language to ensure that any vagaries of a particular data set were averaged out. A given training set consists of 90% of the corpus and the test set consists of the remaining 10%. Each training-test set pair was a random split of the corpora described in section 4. All results presented here are averaged over the results of the five data sets.

7.1. Comparison to adult segmentation: Token F-score

Table 3 presents the token F-score evaluations across the seven languages for all modeled learners, including those using the Bayesian strategy, the subtractive segmentation strategy, and a baseline comparison segmentation strategy. The Bayesian learners include both ideal (BatchOpt) and online (OnlineOpt, OnlineSubOpt, OnlineMem) learners, with either a unigram or bigram generative assumption, while the subtractive segmentation learner follows the strategy described in Lignos (2011). The comparison learner is a syllable-based learner used by Lignos

(2012) that randomly guesses whether a word boundary is present at each potential boundary between syllables, given the true probability of a boundary existing (as estimated from the adult orthographic segmentation). So, for example, if a language's child-directed data tends to be monosyllabic, like English, it may have a probability of 0.75 of a boundary existing (i.e., three out of every four syllables has a word boundary after it). The random guesser would then insert a word boundary after each syllable with probability 0.75, with three out of every four syllables getting a boundary on average. We note that this learner is not intended as a plausible segmentation strategy itself, since it knows the true boundary probability before seeing the language data, but rather serves as a best case scenario for random guessing.

		Eng	**Ger**	**Spa**	**Ita**	**Far**	**Hun**	**Jpn**
Unigram	BatchOpt	0.531	0.603	0.550	0.619	0.666	0.599	0.632
	OnlineOpt	0.588	0.507	0.559	0.599	0.678	0.529	0.622
	OnlineSubOpt	0.637	0.631	0.540	0.602	0.659	0.545	0.613
	OnlineMem	0.551	0.603	0.561	0.586	0.596	0.545	0.637
Bigram	BatchOpt	0.771	0.731	**0.648**	**0.713**	**0.696**	**0.662**	**0.665**
	OnlineOpt	0.751	0.750	0.611	0.671	0.698	0.620	0.642
	OnlineSubOpt	0.778	0.760	0.528	0.613	0.553	0.511	0.519
	OnlineMem	0.863	0.826	0.602	0.609	0.625	0.595	0.633
	Subtractive Seg	**0.879**	**0.843**	0.445	0.423	0.351	0.502	0.321
Baseline	Random Guess	0.559	0.500	0.278	0.233	0.209	0.274	0.257

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.

When we examine the word token F-scores for the Bayesian learners, we see that performance remains relatively high across languages (Unigram: 0.507–0.666, Bigram: 0.528–0.863), though there is considerable variability in exact score. Importantly, these scores are comparable to some previous segmentation strategy evaluations (Goldwater et al., 2009; Johnson, 2008b; Pearl et al., 2011; Fourtassi et al., 2013; Phillips and Pearl, in press), and better than others (Batchelder, 2002; Fleck, 2008; Blanchard et al., 2010; Boruta et al., 2011). This represents clear support for the cross-linguistic property for the Bayesian segmentation strategy. One

of the most obvious differences between the current strategies and many of those investigated previously is the use of a syllable-based input representation. Because there are fewer syllables than phonemes in any given utterance (e.g., *what a nice kitty*, [wʌt ə naɪs kʰɪri] = 12 phonemes but 5 syllables), there are fewer potential boundaries. So, the segmentation task is easier in this sense.[10] This highlights the importance of modeling assumptions – the input representation appears to have a significant impact on the success of the segmentation strategy.

Although the modeling problem may be easier in one sense, the baseline suggests it is far from trivial. There is significant variability in the random guesser's performance, ranging from terrible (Farsi: 0.209, Italian: 0.233) to comparable to some of the poorer performing Bayesian learners (English: 0.559, German: 0.500). Interestingly, the subtractive segmentation strategy also shows this wide variability, ranging from fairly terrible (Japanese: 0.321) to outstanding (English: 0.879, German: 0.843). Notably, the languages that are easier for the random guesser are the ones the Lignos strategy excels at (English, German). This contrasts with the relative constancy of the Bayesian strategy across languages, though the highest segmentation scores are also achieved on these same two languages (English OnlineMem bigram: 0.863, German OnlineMem bigram: 0.826).

Looking within the Bayesian learners, the ideal learner (BatchOpt) generally fares best across languages, whether the learner uses the unigram or bigram assumption. As this learner is capable of optimal inference and batch processing, and so does not have the cognitive constraints the other learners do, this is perhaps unsurprising. For this learner type, the bigram assumption is always helpful – that is, across these seven different languages, it is helpful to believe that the current word predicts the next word, rather than viewing words as independent, if

[10]However, we do note that a statistical learning strategy needs to track more individual units, since there are more syllable types than there are phoneme types. So, for strategies which depend on tracking statistics between types, the syllable-based segmentation task is also harder to some degree.

children were capable of making the optimal inference. However, once cognitive limitations on inference are incorporated, there is some variation for the bigram assumption's utility. While it remains helpful no matter what the cognitive limitations in some languages (English, German, and Italian), others show it having little positive impact (OnlineMem: Japanese) or even a potentially detrimental effect (OnlineSubOpt: Spanish, Farsi, Hungarian, Japanese). This may have to do with the particular error patterns each constrained learner is subject to, which we explore in more detail in section 7.2.

Notably, however, there are languages where a constrained learner fares better than the ideal learner: English and German. For English, all three of the constrained unigram learners fare better than the ideal unigram learner, while the OnlineMem bigram learner fares better than the ideal bigram learner. For German, the OnlineSubOpt unigram learner fares better than the ideal unigram learner, while all three constrained bigram learners fare better than ideal bigram learner. Recall from above that English and German are also notable for being the only two languages where the random guesser baseline fares well, the ones where the subtractive segmentation learner does exceptionally well, and the ones where the Bayesian bigram learners fare the best.

What might be so different about English and German? Fourtassi et al. (2013) have 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., /ɑlɹajt/ segmented as *alright* and *all right* in English). The degree to which this happens varies by language, with the idea that languages with high inherent ambiguity would be harder to correctly segment. If this is true, we might expect that low inherent segmentation ambiguity correlates to high performance by statistical segmentation strategies. With this in mind, perhaps English and German have lower inherent segmentation ambiguity than the other languages.

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

entropy (NSE) metric:

$$NSE = -\sum_{i} P_i \log_2(P_i)/(N-1) \tag{9}$$

where P_i represents the probability of a possible segmentation *i* of an utterance and N represents the length of that utterance in terms of potential word boundaries (so this is determined by the number of syllables for our learners). To calculate the probability of an utterance, we use the unigram or bigram generative model equations described in section 3.1, 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 /ɑl ɹajt ðɛn/, we use the unigram and bigram model equations to generate the probability of every segmentation comprised of true English words (P_i above). In this case, two segmentations are possible: *alright then* and *all right then*. The probabilities for each segmentation are then used in (9) 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 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 BatchOpt learners for comparison.

From Table 4, we see that while German does appear to have the best NSE for both unigram and bigram learners, English does not (ranking fourth overall for both a unigram and bigram learner). So, the high segmentation performance on both German and English cannot simply be due to both having lower inherent segmentation ambiguity – only German does.

More generally, it becomes clear by looking at all seven languages that low NSE does not always lead to higher token F-scores. If it was, we would expect to find a significant negative

Unigram	NSE	F	Bigram	NSE	F
German	0.000257	0.603	**German**	0.000502	0.730
Italian	0.000348	0.619	**Italian**	0.000604	0.713
Hungarian	0.000424	0.599	**Hungarian**	0.000694	0.662
English	0.000424	0.531	**English**	0.000907	0.771
Farsi	0.000602	0.666	**Spanish**	0.00103	0.648
Japanese	0.00126	0.550	**Farsi**	0.00111	0.696
Spanish	0.00128	0.632	**Japanese**	0.00239	0.665

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

correlation between NSE score and token F-score – but this does not happen (unigram: r = -0.084, p = 0.86; bigram, r = -0.341, p = 0.45). When we examine individual languages in Table 4, this lack of correlation is apparent. The unigram Farsi NSE score is ranked fifth lowest, but in fact has the highest F-score, while the unigram Spanish NSE score is actually the worst, though it has the second best F-score. When we turn to the bigram 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 cannot be the sole factor determining segmentation performance, though it may still play a non-trivial role.

Interestingly, something which does seem to correlate with the remarkable "easiness" of English and German segmentation is the probability of a word boundary between any two syllables (recall Table 2). Both English and German tend towards monosyllabic words in our corpora, and word boundaries appear after a syllable at a much higher rate than 50% (Eng: 76.3%, Ger: 74.5%); in contrast, the other languages have a much lower boundary probability (43.8%–51.3%). This effectively means English and German are easier to randomly guess word boundaries for if the learner knows to expect many boundaries from these languages.

To understand why this is, consider the decision the learner is making for every potential boundary: Is this a true boundary or not? If the language's boundary probability is significantly

different from 50%, the random guessing learner with knowledge of this boundary probability is biased to guess in whichever direction the difference is (i.e., guess a boundary is there if the boundary probability is higher than 50% or guess a boundary isn't there if the boundary probability is lower than 50%). The farther away from 50% the boundary probability is, the more this guessing strategy will pay off. For instance, approximately three out of every four potential boundaries will be true boundaries in English (boundary probability = 76.3%) and a random guesser with this boundary probability knowledge will guess boundaries at three out every four potential boundaries. Because more true boundaries will exist in English and this random guesser will be inclined towards placing word boundaries, the potential for overlap is higher. In contrast, in languages like Italian that have a true boundary probability near 50% (Italian = 49.9%), the random guesser is placing a boundary at one of every two possible locations, but true boundaries only exist at one of every two possible locations – so the potential for overlap between the randomly hypothesized boundaries and the true boundaries is lower. This is demonstrated explicitly in Figure 1, which shows the percentage of possible boundaries where the informed random guesser matches the boundaries in the true corpus.

To generate this figure, the informed random guesser was given a corpus with 1,000,000 potential boundaries, where true boundaries randomly appeared with a given boundary probability (e.g., 0.763). The informed guesser then guessed that a boundary existed at each potential boundary location with that given boundary probability. The line indicates how often the random guesser was correct. Each cross indicates one of the seven languages used for evaluation – all seven align more or less as expected. Languages with boundary probabilities farther away from 50% are easier to randomly guess boundaries for if the boundary probability is known. So, the lower bound on informed random guessing segmentation performance for a language is tied to its boundary probability. Languages with boundary probabilities farther from 50% (like English and German) are inherently easier in this sense.

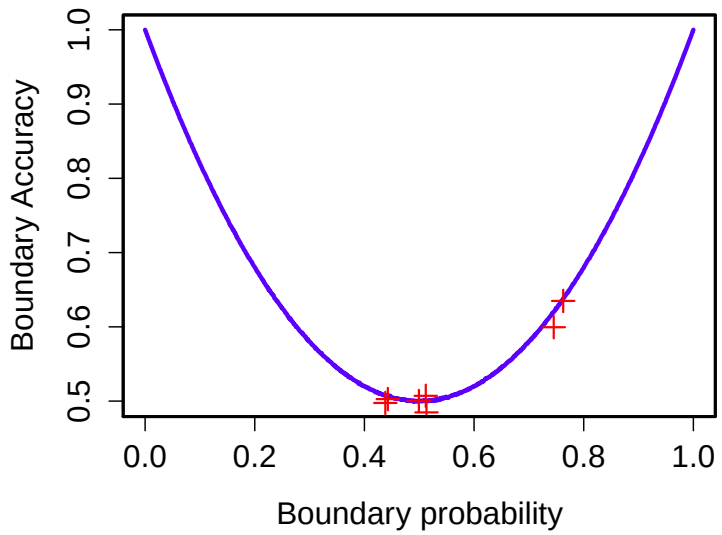

Figure 1: Expected boundary accuracy for an informed random guesser (a random oracle) given a corpus of 1,000,000 potential boundaries, with true boundaries appearing randomly with the given boundary probability. Crosses represent each of the seven languages used for evaluation.

Of course, the modeled learners do not yet know the boundary probability for the language they are trying to segment – but it is in fact one of the things they implicitly learn from their inferred lexicon of word forms. In particular, if they recognize the words comprising the utterances they encounter, they also implicitly know how often a word boundary appears in those utterances, and this is the boundary probability. Learners may then benefit from this implicit boundary probability knowledge when guessing about future utterances whose segmentation is uncertain. For example, an English learner that has many monosyllabic words in its lexicon (and thus a higher boundary probability for utterances more generally) will be more inclined to segment out those monosyllabic words – which is equivalent to guessing that more word boundaries exist, rather than fewer. In English, this is a more successful strategy for the reasons mentioned above. This relates more generally to the utility of a bias for oversegmentation, which we discuss below in 7.2.

7.2. Reasonable error patterns

To better analyze the errors that each learner produces, we searched for errors which appeared to be "reasonable", as described in section 6. In particular, the three types of errors we treated as reasonable were (i) errors that produced real words, (ii) errors that produced common morphemes, and (iii) errors that produced function word collocations. Sample reasonable errors of each type are shown below in Table 5, coming from different languages.

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

Table 5: Examples of reasonable errors (with English glosses) made by learners in different languages. *True* words refer to the segmentation in the original corpus, while *Learner* output represents the segmentation generated by a modeled learner.

Because two of the reasonable error types are oversegmentations (real words, productive morphology) while one is an undersegmentation (function word collocations), the frequency with which different learners generate these errors is related to the kind of segmentation bias the learner has. In particular, learners that tend to oversegment yield more real words and productive morphology, while learners that tend to undersegment yield more function word collocations. We step through the specific error patterns for each reasonable error type below, but note that all our results follow this general trend. Table 6 shows the percentage of oversegmentation errors for each learner across all languages. Because all constrained Bayesian learners had

qualitatively similar behavior, we include only the OnlineMem learner for brevity, as it is the one we believe is most cognitively plausible of the three constrained learners.

		Overseg Errors (%)						
		Eng	Ger	Spa	Ita	Far	Hun	Jpn
Unigram	BatchOpt	1.65	9.06	8.74	39.92	47.67	45.27	39.00
	OnlineMem	8.96	15.91	25.82	53.83	67.97	55.26	53.51
Bigram	BatchOpt	13.83	26.02	32.95	73.09	59.82	57.98	58.38
	OnlineMem	44.82	60.62	72.78	89.85	93.35	82.66	79.86
	Subtractive Seg	90.72	85.35	96.83	99.22	99.00	96.60	99.43
Baseline	Random Guess	53.16	58.73	58.43	58.06	60.45	56.54	54.72

Table 6: Percentage of errors which resulted in an oversegmentation as compared to adult orthographic segmentation.

We first observe that the Bayesian bigram learners have a stronger oversegmentation bias than their unigram counterparts across all languages. Intuitively, this follows from the unigram model's inability to account for frequent word sequences (e.g., *that's a*) as separate words – because these items appear so frequently together, it can only assume that they are a single word. So, the unigram learner is more prone to undersegmentation of collocations (e.g., *that'sa*). In addition, certain languages cause a stronger oversegmentation bias than others for all Bayesian learners: Spanish, Farsi, Hungarian, and Japanese. So, we may expect more of the reasonable oversegmentation errors in these languages.

The subtractive segmentation learner has a massive oversegmentation bias across all languages (85.35-99.43% of all errors are oversegmentations). We therefore expect this learner to generate real word and productive morphology errors, and very few function word collocations (if any). The random guesser comparison learner also has a consistent oversegmentation bias across all languages (53.16-60.45%), suggesting that it may find more real word and morphology errors as well.

7.2.1. Oversegmentations that are real words

Table 7 shows the percentage of oversegmentation errors that yielded at least one real word (e.g., *alright* oversegmented as *all right*). Because low frequency words are not very useful errors, we only count an oversegmentation as a "real word" error if the segmented real word occurred at least five times in the original input.

		Real Word Errors (%)						
		Eng	**Ger**	**Spa**	**Ita**	**Far**	**Hun**	**Jpn**
Unigram	BatchOpt	0.99	3.25	2.81	23.65	20.09	11.88	17.65
	OnlineMem	3.41	4.52	6.31	27.62	25.45	14.91	21.40
Bigram	BatchOpt	5.76	7.89	11.17	38.32	24.63	17.80	26.71
	OnlineMem	29.58	17.63	15.09	57.02	41.46	27.72	34.60
	Subtractive Seg	49.71	25.28	29.32	61.98	42.44	37.99	44.99
Baseline	Random Guess	16.23	7.45	12.81	15.78	11.93	9.14	12.31

Table 7: Percentage of model errors which produced at least one true word in the corpus, excluding true words occurring fewer than five times.

The first observation is that all learners produce this kind of error, though the frequency varies significantly by learner and by language. Turning first to the Bayesian learners, we see that bigram learners generate more of these errors than their unigram counterparts, despite making fewer errors overall. This follows intuitively from the stronger oversegmentation bias the bigram learners have. There is also significant variation across languages for the Bayesian learners, with German and Spanish (and occasionally English) producing real word errors at a much lower rate than the other languages. This may be because oversegmentation errors in these languages result more often in productive morphology errors (see Table 9), rather than real word errors.

Turning to the subtractive segmentation learner, we see that it finds many more real word errors than the Bayesian learners. This can be attributed to this learner's strong oversegmentation tendency. In contrast, the random guesser comparison learner tends to have fewer of these errors (in fact, fewer than the Bayesian learners for several languages), suggesting that its strong

oversegmentation bias doesn't tend to generate this kind of reasonable error.

So how do we interpret this? In essence, the relatively high rate of real word errors in many learners across most languages suggests that some "poor" segmentation performance may really be the result of a segmentation strategy identifying frequently occurring short (often monosyllabic) words. A more nuanced evaluation metric that allows such reasonable errors to count as correct can diminish the performance differences across languages, as Table 8 shows. To calculate the effect of these reasonable errors on the word token F-score, we created a modified target segmentation. This new target segmentation was adjusted to match the reasonable errors which the model output contained. For instance, if the true corpus contained *before* [bəfɔɹ], but the model produced the segmentation *buh four* [bə fɔɹ], the true corpus would be changed to the alternate segmentation. Because [bə] is not an English word, it would still be counted as an incorrect word token even though [fɔɹ] would be counted as correct. Otherwise, token F-score was calculated as normal, but over the modified target segmentation.

We find that the subtractive segmentation strategy benefits enormously, with its lowest F-score now at a respectable 0.550 (Japanese) and its highest F-score outpacing all other strategies (English: 0.917), with an average F-score of 0.690. Still, it is clearly biased to perform better on English and German, while the Bayesian learners now do quite well on Italian as well, with the bigram OnlineMem learner's lowest F-score is 0.703 (Italian) and an average F-score of 0.755. The random guesser comparison learner, in contrast, retains its English and German bias, and generally fares much more poorly on average (0.410) because it has few of these reasonable errors.

7.2.2. Oversegmentations that are productive morphology

For each language, we generated a list of common, productive morphological prefixes and suffixes, both derivational and inflectional. We excluded sub-syllabic morphology because all

		Word Token F-Score, Adjusted for Real Word Errors						
		Eng	**Ger**	**Spa**	**Ita**	**Far**	**Hun**	**Jpn**
Unigram	BatchOpt	0.535	0.615	0.562	0.655	0.702	0.639	0.671
	OnlineMem	0.563	0.617	0.621	0.626	0.650	0.603	0.683
Bigram	BatchOpt	0.782	0.750	**0.730**	**0.748**	**0.741**	**0.711**	**0.728**
	OnlineMem	0.888	0.850	0.707	0.703	0.715	0.708	0.714
	Subtractive Seg	**0.917**	**0.870**	0.674	0.586	0.572	0.661	0.550
Baseline:	Random Guess	0.633	0.537	0.401	0.293	0.292	0.326	0.385

Table 8: Word token F-scores measured against a modified corpus that treats real word errors as correct. The best score for each language is in bold.

learners perceived the input as a sequence of syllables, and so could not possibly segment off anything smaller than a syllable.[11] For an error to be counted as a productive morphology segmentation, the segmented affix had to occur in the appropriate location – i.e., before the word if it was a prefix, such as *re-* in English, or after the word if it was a suffix, such as *-ness* in English. Table 9 contains the percentage of total errors which produced at least one true prefix or suffix of this kind.

		Productive Morphology Errors (%)						
		Eng	**Ger**	**Spa**	**Ita**	**Far**	**Hun**	**Jpn**
Unigram	BatchOpt	0.20	2.68	2.75	3.33	5.00	2.50	9.11
	OnlineMem	0.57	4.57	7.45	4.78	7.45	3.44	10.45
Bigram	BatchOpt	1.01	7.67	10.43	6.30	8.40	3.32	10.34
	OnlineMem	2.62	24.91	20.35	6.70	2.96	4.56	16.88
	Subtractive Seg	5.64	21.94	30.06	10.07	19.14	7.31	30.82
Baseline	Random Guess	2.87	11.52	11.28	2.83	5.14	2.23	9.72

Table 9: Percentage of errors which produced at least one true prefix or suffix in the appropriate location.

As with the real word errors, we see cross-linguistic differences with respect to this error type's frequency. For the Bayesian learners, Japanese tends to produce the most across all learners, though German and Spanish cause the bigram learners to generate a large number as well, since many of the oversegmentation errors for these languages are productive morphology

[11] This therefore ruled out much of the common inflectional morphology of Indo-European languages.

errors rather than real word errors. In contrast, Hungarian, Italian, Farsi, and English generally yield fewer of these errors due to their relative lack of productive syllabic morphology.[12]

For the subtractive segmentation learner, we see this same qualitative trend, with Japanese, German, Farsi, and Spanish producing a larger number, while Hungarian and English produce the least. The random guesser comparison learner also shows this same trend, though German and Spanish generate more productive morphology errors than Japanese for this learner.

The persistence of this cross-linguistic trend suggests that this error type is usefully capturing some of the cross-linguistic variation. If these error types are counted as reasonable output (as shown in Table 10), performance for Spanish increases most across all learners (by 0.08 on average), while Japanese and Farsi also benefit to a large degree (by 0.05 and 0.04, respectively). This brings these languages' performance much closer to the high token F-scores found originally for English and German.

		Word Token F-Score, Adjusted for Morph Errors						
		Eng	**Ger**	**Spa**	**Ita**	**Far**	**Hun**	**Jpn**
Unigram	BatchOpt	0.532	0.609	0.554	0.628	0.676	0.611	0.652
	OnlineMem	0.552	0.610	0.607	0.591	0.617	0.561	0.659
Bigram	BatchOpt	0.771	0.744	**0.715**	**0.712**	**0.715**	**0.672**	**0.685**
	OnlineMem	0.862	**0.843**	0.681	0.618	0.672	0.634	0.662
	Subtractive Seg	**0.869**	0.838	0.638	0.461	0.484	0.520	0.502
Baseline	Random Guess	0.593	0.533	0.365	0.255	0.271	0.279	0.365

Table 10: Word token F-scores evaluated against a modified corpus that treats productive morphology reasonable errors as correct. The best score for each language is in bold.

[12]This may initially seem surprising for English, which does have productive syllabic affixes like *-ing*. However, recall that the learners perceive the input as a stream of syllables with no awareness of sub-syllabic information. So, *reading* becomes *rea ding* and *running* becomes *ru nning*. The *-ing* that *ding* and *nning* have in common is invisible to these learners. Similar issues occur for the other languages mentioned.

7.2.3. Undersegmentations that are function word sequences

The third error type is an undersegmentation error yielding a sequence of function words, such as *that'sa*. A set of function words was derived by hand for each language, drawing on native speaker knowledge and available dictionaries. As with the other error types, there are significant cross-linguistic differences, shown in Table 11.

		Function Word Errors (%)						
		Eng	**Ger**	**Spa**	**Ita**	**Far**	**Hun**	**Jpn**
Unigram	BatchOpt	8.82	27.17	8.94	6.35	4.34	2.31	6.85
	OnlineMem	10.20	26.71	7.70	5.80	3.05	2.33	7.69
Bigram	BatchOpt	15.71	28.19	6.42	5.81	3.65	2.85	5.16
	OnlineMem	9.92	10.79	2.30	1.09	0.19	1.73	2.60
	Subtractive Seg	0.47	1.91	0.33	0.00	0.00	0.00	0.00
Baseline	Random Guess	4.07	9.40	4.11	3.45	0.97	0.91	1.68

Table 11: Percentage of errors composed of a sequence of function words.

For the Bayesian learners, German and English yield far more than any other language, perhaps due to the frequency of monosyllabic function words. In particular, 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 Bayesian unigram learners are especially subject to generating this error, due to their tendency towards undersegmentation. In contrast, the OnlineMem bigram learner and the subtractive segmentation learner have very strong oversegmentation biases, and so generate this error much less often. The random guesser generates more of these errors, due to having a weaker oversegmentation bias than the subtractive segmentation learner.

When this reasonable error is also viewed as acceptable segmentation output, the segmentation performance of all Bayesian learners benefits, as shown in Table 12. German and English benefit the most, as they generate the most function word errors, though all Bayesian learners benefit some, regardless of language. In contrast, the subtractive segmentation learner doesn't

benefit much, due to the infrequency of this error type in its output (which is attributable to its powerful oversegmentation bias). The random guesser does benefit some, though both it and the subtractive segmentation learner only yield acceptable performance in German and English. This suggests that this error type captures another aspect of German and English that sets them apart from the other languages examined here: their function words are prone to undersegmentation.

		Word Token F-Score, Adjusted for Func Errors						
		Eng	**Ger**	**Spa**	**Ita**	**Far**	**Hun**	**Jpn**
Unigram	BatchOpt	0.564	0.695	0.581	0.641	0.678	0.615	0.652
	OnlineMem	0.587	0.692	0.617	0.597	0.605	0.564	0.659
Bigram	BatchOpt	0.802	0.800	**0.693**	**0.714**	**0.705**	**0.675**	**0.679**
	OnlineMem	0.871	**0.843**	0.641	0.607	0.625	0.632	0.641
	Subtractive Seg	**0.875**	0.839	0.465	0.427	0.362	0.502	0.380
Baseline	Random Guess	0.607	0.545	0.315	0.259	0.233	0.278	0.341

Table 12: Word token F-scores evaluated against a modified corpus which treats function word collocations as correct. The best score for each language is in bold.

7.2.4. All reasonable errors included

Table 13 shows the results of considering all three reasonable error types as acceptable segmentation.

		Eng	**Ger**	**Spa**	**Ita**	**Far**	**Hun**	**Jpn**
Unigram	BatchOpt	0.568	0.711	0.592	0.680	0.721	0.648	0.711
	OnlineMem	0.600	0.713	0.646	0.651	0.755	0.614	0.727
Bigram	BatchOpt	0.815	0.829	**0.747**	**0.768**	**0.764**	**0.720**	**0.763**
	OnlineMem	0.901	**0.884**	0.715	0.712	0.751	0.714	0.751
	Subtractive Seg	**0.916**	0.872	0.692	0.595	0.632	0.663	0.645
Baseline	Random Guess	0.646	0.594	0.424	0.310	0.317	0.331	0.418

Table 13: Adjusted word token F-scores for a selection of learners across all languages when all three reasonable error types are counted as correct segmentation. Higher token F-scores indicate better performance. The best score for each language is in bold.

Notably, the Bayesian learners remain consistently very good across all languages, particu-

larly the bigram learners (minimum F-score: 0.712 (Italian), maximum: 0.901 (English)). The main difference is simply that the average token F-score is higher than before (e.g. unadjusted bigram: 0.666; adjusted bigram: 0.760), which is also considerably higher than the standard performance reported in previous segmentation studies (e.g., Batchelder, 2002; Fleck, 2008; Blanchard et al., 2010; Boruta et al., 2011). Also, we note that the OnlineMem bigram learner performs as well as or better than the BatchOpt bigram learner on most languages. In cases where it performs below the BatchOpt (e.g., Spanish, Italian), it does not perform so poorly that we would consider it unsuccessful. This indicates that constrained inference does not 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.

Another important finding of the reasonable error analysis is that, as we take into account reasonable errors, differences in performance across languages tend to diminish (as represented by standard deviations across all languages, shown in Table 14). This is especially true for the bigram Bayesian and subtractive segmentation learners: performance variability across languages decreases (Bayesian Bigram: 0.09 to 0.06, Subtractive Seg: 0.21 to 0.12). This demonstrates how variation in model performance across many languages can be partially accounted for by examining model output in a more nuanced way.

	Gold Standard (std. dev.)	Reasonable Errors (std. dev.)
Bayesian Unigram	0.05	0.06
Bayesian Bigram	0.09	0.06
Subtractive Seg	0.21	0.12
Random Oracle	0.13	0.13

Table 14: Standard deviation of word token F-score results across all languages for different classes of learners, comparing evaluation against the adult orthographic segmentation (Gold Standard) and against the segmentation where reasonable errors are viewed as correct (Reasonable Errors). Lower standard deviations indicate less variability in performance across languages.

In addition, the Bayesian strategy is more consistent in its cross-linguistic performance than

the subtractive segmentation strategy, regardless of how the Bayesian strategy is implemented and the manner in which it is evaluated (Bayesian variability: 0.05–0.09; Subtractive variability: 0.12-0.21). This is because the subtractive strategy does exceptionally well on English and German (outperforming most of the Bayesian learners, in fact) while doing less well – though still very good with a nuanced evaluation – on the remaining languages. This again highlights the utility of using a more nuanced evaluation of segmentation success, as the subtractive learner suffered considerably on all languages but English and German when the adult orthographic segmentation was the sole segmentation target. The random guesser also has this bias for good performance on English and German, though its performance on the remaining languages is not nearly as high.

7.3. Summary of results

Our initial findings comparing against the adult orthographic segmentation typically used for segmentation evaluation demonstrated that the Bayesian segmentation strategy yielded consistent and reasonably good segmentation performance across the seven languages investigated. In contrast, the subtractive segmentation strategy suffered on languages other than English or German. This underscores the Bayesian strategy's cross-linguistic versatility. Nonetheless, once we allow a more nuanced evaluation metric that considered useful errors to be reasonable segmentation output, the differences between the languages significantly lessened. Both the Bayesian strategy and the subtractive segmentation strategy fare well, though the subtractive segmentation strategy retains a performance bias favoring English and German.

More generally, the segmented units identified by both the Bayesian strategy and the subtractive segmentation strategy tend to be reasonable, even if they do not match the orthographic words of the language. So, these units may be helpful for six- or seven-month-old infants who are just learning to segment their language's input, even if these units are not the speaker's intended words. Interestingly, the two successful strategies make different predictions about the

precise nature of these reasonable errors. The Bayesian strategy suggests these errors will be distributed across real words, productive morphology, and function word sequences, with the exact distribution depending on the language. In contrast, the subtractive segmentation strategy suggests that most errors will be real words followed by productive morphology, with almost no function word sequences, irrespective of language. This is in part due to the different segmentation biases the two strategies have, with the subtractive segmentation strategy having a far stronger oversegmentation bias for all languages than any of the Bayesian learners. This suggests that infants using the subtractive segmentation strategy should exhibit a strong oversegmentation bias, no matter the language. These are predictions that can be tested experimentally in young infants, and can thus differentiate empirically between the two strategies.

Still, as with any model, there are simplifying assumptions incorporated into the model implementations that yielded these results. First, experimental evidence has yet to converge on the exact representation infants have for fluent speech. Given this, it is possible that using syllables to represent speech does not in fact match how infants represent speech. Second, even if infants do represent fluent speech using syllables, they may not represent the syllables as we have done here, i.e., using both adult syllabic representations and syllables created using the maximum-onset principle. Future experimental work is needed to address whether these modeling assumptions match infant representations, and update the model implementations as necessary.

8. Conclusion

We have proposed that universal early language learning strategies should satisfy four criteria: (i) operate over realistic input, (ii) be usable by infants, (iii) yield useful output, and (iv) succeed across all languages children might have to learn. To investigate two strategies intended to be universal, we have evaluated modeled learners across seven languages, and found

that these strategies do indeed satisfy these criteria. In particular, our modeled Bayesian and subtractive segmentation learners (i) perceive the input in a cognitively plausible way and learn from a realistic input distribution, (ii) have a cognitively constrained inference process, (iii) generate useful units that can be smaller or larger than orthographic words, and (iv) do so on a variety of languages with different linguistic properties. One difference between the strategies is that subtractive segmentation achieves cross-linguistic success only when useful units smaller than orthographic words are considered acceptable segmentation output, while the Bayesian strategy succeeds regardless. Additionally, each strategy makes different empirically-testable predications about the kinds of errors infants should make during early segmentation. Infants using a Bayesian strategy should produce errors that reflect the structure of the language, while infants using the subtractive segmentation strategy should produce errors that are roughly similar across languages.

To our knowledge, these are the first segmentation strategies to be validated using all four proposed criteria for a universal strategy, though we expect that the segmentation corpora we have assembled across seven languages will aid researchers who are interested in evaluating cross-linguistic success more thoroughly. In addition, our work contributes to how language acquisition models might be evaluated. While adult knowledge as implemented in gold standards can provide a simple comparison metric across modeling studies, achieving adult knowledge may not be a plausible goal for infants during the early stages of acquisition. Instead, researchers can add to their evaluation toolkit by looking at more nuanced definitions of acquisition success. By doing so, we may find that strategies that looked insufficient when evaluated against the gold standard merit further consideration.

Acknowledgments

The authors would like to thank Caroline Wagenaar, James White, Galia Barsever, Tiffany Ng, Alicia Yu, Nazanin Sheikhan, and Sebastian Reyes for their help with corpus preparation. In addition, we are very grateful to Robert Daland, Constantine Lignos, Amy Perfors, Naomi Feldman, Abdellah Fourtassi, Jon Sprouse, Barbara Sarnecka, Michael Lee, Alex Ihler, UCLA's phonology seminar, and audiences at the CogACLL 2014 workshop and CogSci 2014 for their feedback.

Aslin, R., Saffran, J., Newport, E., 1998. Computation of conditional probability statistics by 8-month-old infants. Psychological Science 9, 321–324.

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

Bernstein-Ratner, N., 1984. Patterns of vowel modification in motherese. Journal of Child Language 11, 557–578.

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.

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

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

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

Boruta, L., Peperkamp, S., Crabbé, B., Dupoux, E., 2011. Testing the robustness of online word segmentation: Effects of linguistic diversity and phonetic variation. In: Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics at the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. pp. 1–9.

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

Broen, P., 1972. The verbal environment of the language learning child. ASHA monographs 17.

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

Cole, R., Jakimik, J., 1980. Perception and production of fluent speech. Erlbaum, Hillsdale, NJ, pp. 133–163.

Comrie, B., 1989. Language Universals and Linguistic Typology: Syntax and Morphology. The University of Chicago Press, Chicago, IL.

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

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

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

Elsner, M., Goldwater, S., Eisenstein, J., 2012. Bootstrapping a unified model of lexical and phonetic acquisition. In: Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1. pp. 184–193.

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

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

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

Fleck, M., 2008. Lexicalized phonotactic word segmentation. In: Proceedings of ACL-08: HLT. pp. 130–138.

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.

Gambell, T., Yang, C., 2006. Word segmentation: Quick but not dirty.

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

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

Hirsh-Pasek, K., Nelson, D. K., Jusczyk, P., Cassidy, K., Druss, B., Kennedy, L., 1987. Clauses are perceptual units for young infants. Cognition 26, 269–286.

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., 2008a. 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., 2008b. Using adaptor grammars to identify synergies in the unsupervised acquisition of linguistic structure. In: Proceedings of ACL-08: HLT. pp. 398–406.

Johnson, M., Demuth, K., 2010. Unsupervised phonemic chinese word segmentation using adaptor grammars. In: Proceedings of the 23rd International Conference on Computational Linguistics. Association for Computational Linguistics, pp. 528–536.

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

Jusczyk, P., 1997. The Discovery of Spoken Language. MIT Press, Cambridge, MA.

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

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

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

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

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

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

Lignos, C., 2011. Modeling infant word segmentation. In: Proceedings of the Fifteenth Conference on Computational Natural Language Learning. Association for Computational Linguistics, pp. 29–38.

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 Edition. Lawrence Erlbaum Associates, Mahwah, NJ.

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.

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

Morgan, J., Saffran, J., 1995. Emerging integration of sequential and suprasegmental information in preverbal speech segmentation. Child Development 66 (4), 911–936.

Pearl, L., 2011. When unbiased probabilistic learning is not enough: Acquiring a parametric system of metrical phonology. Language Acquisition 18 (2), 87–120.

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.

Pearl, L., Sprouse, J., 2013. Syntactic islands and learning biases: Combining experimental syntax and computational modeling to investigate the language acquisition problem. Language Acquisition 20, 23–68.

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

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

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

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

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

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., 2014. 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., 2015. 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.

Phillips, L., Pearl, L., in press. The utility of cognitive plausibility in language acquisition modeling: Evidence from word segmentation. Cognitive Science.

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

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

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

Selkirk, E., 1981. English compounding and the theory of word structure. In: Moortgat, M., van der Hulst, H., Hoestra, T. (Eds.), The Scope of Lexical Rules. Foris, Dordrecht, pp. 229–277.

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

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

Venkataraman, A., 2001. A statistical model for word discovery in transcribed speech. Computational Linguistics 27 (3), 351–372.

Wang, H., Mintz, T., 2008. A dynamic learning model for categorizing words using frames. In: Proceedings of BUCLD. Vol. 32. pp. 525–536.

Werker, J., Lalonde, C., 1988. Cross-language speech perception: Initial capabilities and developmental change. Developmental Psychology 24 (5), 672–683.

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

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