Bayesian inference as a viable cross-linguistic word segmentation strategy: It's all about what's useful

Lawrence Phillips and Lisa Pearl
Department of Cognitive Sciences
University of California, Irvine
{lawphill, lpearl}@uci.edu

Abstract

Identifying useful items from fluent speech is one of the first tasks children must accomplish during language acquisition. Typically, this task is described as word segmentation, with the idea that words are the basic useful unit that scaffolds future acquisition processes. However, it may be that other useful items are available and easy to segment from fluent speech, such as sub-word morphology and meaningful word combinations. A successful early learning strategy for identifying words in English is statistical learning, implemented via Bayesian inference (Goldwater, Griffiths, & Johnson, 2009; Pearl, Goldwater, & Steyvers, 2011; Phillips & Pearl, 2012). Here, we test this learning strategy on child-directed speech from seven languages, and discover it is effective cross-linguistically, especially when the segmentation goal is expanded to include these other kinds of useful units. We also discuss which useful units are easy to segment from the different languages using this learning strategy, as the useful unit varies across languages.

Keywords: language acquisition; Bayesian learning; word segmentation; cross-linguistic; segmentation metrics

Introduction

Segmenting useful items, typically words, from fluent speech is one of the first tasks children face in learning their native language. The earliest evidence of infant word segmentation comes at six months (Bortfeld, Morgan, Golinkoff, & Rathbun, 2005) when familiar names are used to segment adjacent words. By seven and a half months, infants are beginning to segment words using the most common stress pattern in their language (Jusczyk, Cutler, & Redanz, 1993; Jusczyk & Aslin, 1995; Echols, Crowhurst, & Childers, 1997), and by nine months infants also utilize phonotactics (Mattys, Jusczyk, & Luce, 1999), metrical stress patterns (Morgan & Saffran, 1995), and coarticulation effects (Johnson & Jusczyk, 2001) to identify words. Importantly, these later segmentation strategies use cues that vary cross-linguistically (e.g., metrical stress: English words tend to have word-initial stress while French words tend to have word-final stress). In order to identify the relevant cues for these strategies, infants need a pool of words from which to learn the language-specific cue.

While knowing some words is necessary to infer useful language-specific cues to word segmentation, there are other units that may be useful to segment. For example, sub-word morphology can be useful for grammatical categorization (e.g., *-ing* for identifying the word as a verb). Similarly, meaningful word combinations could be useful for early structure learning (e.g., *could+I* functioning as a kind of yes/no question marker). So, while it is important to know how children could segment words, it is likely that segmenting other units is helpful for acquisition.

Proposals for early segmentation strategies have centered on language-independent cues that do not need to be derived from knowledge of existing words, such as transitional probability between syllables (Saffran, Aslin, & Newport, 1996). Experimental evidence also suggests that statistical cues like transitional probability are used earlier than language-specific cues like metrical stress (Thiessan & Saffran, 2003).

Bayesian inference for early statistical word segmentation has been shown to be successful for identifying words in English, whether the salient perceptual units are phonemes (Goldwater et al., 2009; Pearl et al., 2011) or syllables (Phillips & Pearl, 2012), and whether the inference process is optimal (Goldwater et al., 2009; Pearl et al., 2011) or constrained by cognitive limitations that children may share (Pearl et al., 2011; Phillips & Pearl, 2012). Notably, however, there is little evidence that current Bayesian word segmentation approaches succeed cross-linguistically (though see Johnson, 2008 and Fourtassi et al., 2013 for some examples).

If Bayesian segmentation is meant to be a universal early strategy, cross-linguistic success is crucial. Interestingly, there is some evidence that English may be inherently easier to segment into words than other languages (Fourtassi, Börschinger, Johnson, & Dupoux, 2013). We therefore evaluate the Bayesian learners of Phillips & Pearl (2012) on seven languages with different linguistic profiles: English, German, Spanish, Italian, Farsi, Hungarian, and Japanese. Because experimental evidence suggests that infants younger than seven and a half months categorically represent syllable-like units, but not phonemes (Jusczyk & Derrah, 1987; Eimas, 1999) and that phonological representations are still developing at this age (Werker & Tees, 1984), we follow previous modeling studies (Swingley, 2005; Gambell & Yang, 2006; Lignos & Yang, 2010; Phillips & Pearl, 2012) and assume that the relevant perceptual units for word segmentation are syllables.

We show that Bayesian segmentation is indeed a successful cross-linguistic learning strategy, especially if we define success in a more practical way than previous segmentation studies have done. In particular, we consider a segmentation strategy successful if it identifies units useful for subsequent language acquisition processes (e.g., grammatical categorization, structure learning). Thus, not only are the orthographic words traditionally used as the "gold standard" in word segmentation tasks acceptable (e.g., in "I am eating in the kitchen": the orthographic words are *I*, *am*, *eating*, *in*, *the*, and *kitchen*), but also productive morphology (e.g., *-ing*) and coherent chunks made up of multiple function words (e.g.,

Iam, *inthe*) similar to some of the errors attested by Brown (1973; e.g., *that'sa, what'sthat*).

The Bayesian learning strategy

Bayesian models are well suited to questions of language acquisition because they naturally distinguish between the learner's pre-existing beliefs (prior) and how the learner evaluates incoming data (likelihood), using Bayes' theorem:

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

The Bayesian learners we evaluate are the optimal learners of Goldwater et al. (2009) and the constrained learners of Pearl et al. (2011). All learners are based on the same underlying generative models developed by Goldwater et al. (2009). The first of these models assumes words do not depend on previous words (a unigram assumption) while the second assumes that a word depends only on the word before it (a bigram assumption). While both are clearly overly-simplistic ideas about how language is generated, they may serve as a reasonable approximation of an infant's first guesses about language structure. To encode these assumptions into the model, Goldwater et al. (2009) use a Dirichlet process (Ferguson, 1973), assuming that the observed sequence of words $w_1 \ldots w_n$ is generated sequentially using a probabilistic generative process. In the unigram case, the identity of the i^{th} word is chosen according to:

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

where n_{i-1} is the number of times w appears in the previous i–1 words, α is a free parameter of the model, and P_0 is a base distribution specifying the probability that a novel word will consist of the perceptual units $x_1 \ldots x_m$:

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

In the bigram case, a hierarchical Dirichlet Process (Teh, Jordan, Beal, & Blei, 2006) is used. This model also tracks the frequencies of two-word sequences and is defined as:

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

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

where $n_{i-1}(w', w)$ is the number of times the bigram (w', w) has occurred in the first i–1 words, $b_{i-1}(w)$ is the number of times w has occurred as the second word of a bigram, b is the total number of bigrams, and β and γ are free model parameters.[1]

[1]Parameters for the models utilized by all learners were chosen to maximize the word token F-score of the unigram and bigram BatchOpt learner. English: $\alpha = 1, \beta = 1, \gamma = 90$; German: $\alpha = 1, \beta = 1, \gamma = 100$; Spanish: $\alpha = 1, \beta = 200, \gamma = 50$; Italian: $\alpha = 1, \beta = 20, \gamma = 200$; Farsi: $\alpha = 1, \beta = 200, \gamma = 500$; Hungarian: $\alpha = 1, \beta = 300, \gamma = 500$; Japanese: $\alpha = 1, \beta = 300, \gamma = 100$.

In both the unigram and bigram case, this generative model implicitly incorporates preferences for smaller lexicons by preferring words that appear frequently (due to (2), (4), and (5)) and preferring shorter words in the lexicon (due to (3)), both of which may be thought of as domain-general parsimony biases.

Learners: Implementing Bayesian inference

The **BatchOpt** learner for this model is taken from Goldwater et al. (2009) and utilizes Gibbs sampling (Geman & Geman, 1984) to run over the entire input in a single batch, sampling every potential word boundary 20,000 times to decide if a word boundary is present. This represents the most idealized learner, since Gibbs sampling is guaranteed to converge on the segmentation which best fits the underlying model. Notably, this learner incorporates no cognitive processing or memory constraints. Because of this, we also evaluate the most successful constrained learner developed by Pearl et al. (2011) that incorporates processing and memory constraints, providing a test of the utility of the model's learning assumptions when inference is not guaranteed to be optimal.

The **Online-Mem** learner is taken from Pearl et al. (2011), and is similar to the BatchOpt learner in that it samples boundaries during learning. However, the Online-Mem learner operates utterance by utterance and does not sample all potential boundaries equally. Instead, it implements a Decayed Markov Chain Monte Carlo algorithm (Marthi, Pasula, Russell, & Peres, 2002), sampling s previous boundaries using the decay function b^{-d} to select the boundary to sample; b is the number of potential boundary locations between the boundary under consideration b_c and the end of the current utterance while d is the decay rate. So, the further b_c 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. All results presented here use a set, non-optimized value for d of 1.5, which was chosen to implement a heavy memory constraint (e.g., 90% of samples come from the current utterance, while 96% are in the current or previous utterance). Having sampled a set of boundaries[2], the learner can then update its beliefs about those boundaries and subsequently update its lexicon before moving on to the next utterance.

Perceptual units

While the original model by Goldwater et al. (2009) used phonemes as the basic perceptual unit for word segmentation, the learning model can operate on any unit. Based on experimental evidence, we chose syllables as a more realistic unit of representation for six- and seven-month-old infants just beginning segmentation. By three months, infants seem to possess categorical perception of syllable-like units, but not of phones (Jusczyk & Derrah, 1987; Eimas, 1999). Moreover, infants continue to distinguish non-native consonant contrasts

[2]All Online-Mem learners sample $s = 20{,}000$ boundaries per utterance. For a syllable-based learner, this works out to approximately 74% less processing than the BatchOpt learner (Phillips & Pearl, 2012).

until ten to twelve months (Werker & Tees, 1984), which is much later than when early segmentation begins (although vowels do begin this process around six months: Polka & Werker 1994).[3]

This means that syllables are viewed as atomic units, and the learner loses access to all phonotactic information within a syllable. This assumption is supported by experimental evidence showing that three-month-olds do not recognize sub-syllabic similarities between syllables (Jusczyk & Derrah, 1987). For instance, while three-month-olds distinguish /ba/ from /bu/ and /ba/ from /du/, they do not regard /ba/ as more similar to /bu/ than /du/, though /ba/ and /bu/ share an initial phoneme while /ba/ and /du/ do not. This may suggest that infants disregard sub-syllabic information at this early age.

From a learning perspective, using syllables as the basic perceptual input has both potential benefits and drawbacks. The learning problem is somewhat easier because boundaries cannot occur within syllables, which limits the number of possible boundary locations per utterance. However, the model loses access to all phonotactic information in the language, which can provide useful statistical cues to boundary locations (Blanchard, Heinz, & Golinkoff, 2010).

Cross-linguistic input

We evaluate the Bayesian learner on corpora of child-directed speech in seven languages: English, German, Spanish, Italian, Farsi, Hungarian and Japanese. All corpora were taken from the CHILDES database (MacWhinney, 2000) and are briefly summarized in Table 1. When corpora were available only in orthographic form, they were converted into an appropriate phonemic form by native speakers. Afterwards, unsyllabified corpora were syllabified. Where possible, we utilized adult syllabification judgments (Baayen, Piepenbrock, & Gulikers, 1996). All other words were syllabified using the Maximum-Onset principle, which states that the beginning of a syllable should be as large as possible, without violating a language's phonotactic constraints. We note that this serves only as an approximation of the infant representation, given the lack of clear data on infant syllabification at this age.

Our corpora vary in a number of important ways. While most of our corpora are Indo-European languages (English, German, Spanish, Italian, Farsi), we also use data from two non-Indo-European languages (Hungarian, Japanese). Languages were chosen such that available native speakers could give guidance regarding the phonemic encoding and segmentation results for each language. Though the learning task we model is one which occurs in the first seven months, child-directed speech corpora are not always easily available in this age range. So, while many of our corpora do consist entirely of early child-directed speech (e.g., English, Japanese), some corpora contain speech directed to older children as well (e.g., German, Farsi). Likewise, the same amount of data is not easily available for each language. Our shortest corpus (German) consists of 9,378 utterances, while the longest (Farsi) consists of 31,657. Notably, corpus size does not seem to affect segmentation performance noticeably (see Table 3) which may indicate that performance for this type of Bayesian segmentation strategy plateaus relatively quickly.

[3]We note that utilizing syllables does not address one potential concern: if at six to seven months, infants are still distinguishing non-native contrasts, this may indicate that all representational units at that age are phonetically narrower than adult representations. For example, infants may categorically represent both /ta/ and /t^ha/.

	Corpora	**(age range)**	**# Utt**	**# Syl**
English	Brent	(0;6-0;9)	28391	2330
German	Caroline	(0;10-4;3)	9378	1683
Spanish	JacksonThal	(0;10-1;8)	16924	524
Italian	Gervain	(1;0-3;4)	10473	1158
Farsi	Family, Samadi	(1;8-5;2)	31657	2008
Hungarian	Gervain	(1;11-2;11)	15208	3029
Japanese	Noji, Miyata, Ishii	(0;2-1;8)	12246	526

Table 1: Summary of cross-linguistic corpora from CHILDES, including age range of children the speech was directed at, the number of child-directed speech utterances, and the number of unique syllables.

The languages themselves also contain many differences that potentially affect syllable-based word segmentation. While our English and Hungarian corpora contain 2,330 and 3,029 unique syllables, respectively, Japanese and Spanish contain only 526 and 524. Because the generative model prefers units to appear frequently, languages with fewer syllables will tend to have those syllables appear often, potentially causing the learner to identify individual syllables as words. This can lead to oversegmentation errors, such as *kissing* segmented as *kiss* and *-ing*. In addition, these languages also differ in their syntax and morphology. For example, Hungarian and Japanese are both agglutinative languages that have more regular morphological systems, while English, German, Spanish, Italian and Farsi are all fusional languages to varying degrees. If a language has regular morphology, the learner might reasonably segment morphemes rather than words, and later language learning will depend on successful segmentation of morphemes. This highlights the need for a more flexible metric of segmentation performance: A segmentation strategy that identifies useful morphology in agglutinative languages would be at a disadvantage if the "gold standard" of orthographic words is used to evaluate it, even though useful units have been identified.

Learning results & discussion

We first analyze our results in terms of word token F-score, the harmonic mean of token precision (P) and recall (R): $F = 2 * \frac{P*R}{P+R}$. Precision measures the probability that a word segmented is a true word (# identified true / # identified) and recall measures the probability that any true word was correctly identified (# identified true / total # true). F-scores range from 0 to 100, with higher values indicating better performance. Performance on all languages is presented in

Table 3. The non-bolded F-scores represent performance against the "gold standard" of orthographic words typically used in word segmentation modeling studies (Goldwater et al., 2009; Pearl et al., 2011; Blanchard et al., 2010; Lignos & Yang, 2010). This provides a simple, easy-to-implement metric for comparison to previous segmentation models, though it has its conceptual shortcomings as the target state for early segmentation, as discussed previously.

While English and German both perform very well against the gold standard (Bigram BatchOpt: 77.06 and 73.05), all other languages have somewhat lower performance (e.g., Spanish: 64.75 and Japanese: 66.53). Still, our results far outperform a random-guess baseline segmenter (21.37–38.15). We also compare our results to the subtractive segmenter with beam search from Lignos (2011), which provides a good baseline since it is also syllable-based and performs extremely well on English. This learner goes through the input segmenting any word which it has previously recognized. When there are multiple possible words, the word previously encountered most often is chosen. While this method works well for English (87.77) and German (82.37), it fairs much more poorly with the remaining languages (30.09 – 58.25).

One important factor noticed by Fourtassi et al. (2013) is that English is less ambiguous with respect to segmentation than other languages. Fourtassi et al. (2013) compare phonemically-encoded corpora of English and Japanese, demonstrating with an Adaptor Grammar (Johnson et al., 2007) that performance is much higher for English than Japanese. They explain their results in terms of Normalized Segmentation Entropy (NSE), defined for any utterance as:

$$NSE = -\sum_i P_i \log_2(P_i)/(N-1) \quad (6)$$

where P_i represents the probability of a particular segmentation i and N represents the length of the utterance in terms of perceptual units (e.g., phonemes in their analysis). In essence, given knowledge of all the true words and their frequencies, NSE quantifies how ambiguous a particular utterance remains with respect to segmentation. For example, Fourtassi et al. (2013) note that the phrase /ajskɹim/ has two possible segmentations that produce only English words: "I scream" (/aj skɹim/) and "ice cream" (/ajs kɹim/).

Using our own unigram and bigram models to stand in for the probability of any given segmentation, we replicate Fourtassi et al.'s (2013) findings that English segmentation is less ambiguous than Japanese (see Table 2). Notably however, we find that ambiguity does not correlate with our unigram results (r = -.0510) and correlates only moderately with our bigram results (r = -.3871).

Given that the differences in segmentation performance could not be attributed solely to varying segmentation ambiguity, we investigated the types of errors made across languages. It turned out that many errors fell into one of three categories of "useful errors", described below. Table 3 shows token F-scores when compared against an adjusted gold standard that does not penalize certain "useful errors". Table 4 presents common examples of each type of useful error.

	Uni NSE	Uni F	Bi NSE	Bi F
Ger	0.000257	60.33	0.000502	73.05
Ita	0.000348	61.85	0.000604	71.25
Hun	0.000424	59.90	0.000694	66.20
Eng	0.000424	53.12	0.000907	77.06
Far	0.000602	66.63	0.00111	69.63
Spa	0.00128	55.03	0.00103	66.53
Jpn	0.00126	66.63	0.00239	69.63

Table 2: NSE scores compared against the BatchOpt token F-score for a language, when compared against the gold standard word segmentation. 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 word token identification.

First, we adjusted for mis-segmentations resulting in **real words**. For example, /alɹajt/ (*alright*) might be oversegmented as /al/ /ɹajt/ (*all right*), resulting in two actual English words. All languages show errors of this type, often occurring for the bigram model, with the fewest in English (BatchOpt: 4.52% of all errors) and the most in Spanish (BatchOpt: 23.97%). These errors are likely due to the model's preference to segment frequently-occuring words it has already seen.

Another reasonable error is productive **morphology**. Because the perceptual unit is the syllable, only syllabic morphology can be identified in this manner. This likely explains why languages like English, Spanish, and Italian have relatively few errors that produce morphemes (e.g., BatchOpt: 0.13%, 0.05%, and 1.13% of all errors respectively), while Japanese, with more syllabic morphology, has more such errors (e.g., BatchOpt: 4.69%). Prefixes and suffixes were only identified as useful morphological errors when they appeared at the beginning or end of a segmented word, respectively. For instance, the prefix /ɹi/ as in *redo*, would not be counted as a useful error if *very* were to be segmented as /vɛ ɹi/.

A third reasonable error type was common sequences of **function words**. For example, a learner might identify *is that a* as a single word *isthata*, similar to the errors reported by Brown (1973). These errors tend to be more common for unigram than bigram learners. This is intuitive from a statistical standpoint because the unigram model is unable to account for commonly occurring sequences of words (since it assumes all words are independent) and so accounts for these frequently occurring sequences by combining them into a single word. Still, function word sequence errors are relatively uncommon in every language except German (e.g., BatchOpt: 21.73%; vs. English: 4.30%, Farsi: 2.12%).

For all useful errors, F-scores were adjusted so that the "correct" portions of the error were not penalized. For instance, if a learner mis-segmented *oopsie* as *oop* and *see*, *see* would be counted as correct because it is a real English word

		Eng	Ger	Spa	Ita	Far	Hun	Jpn
Unigram	Batch-Opt	53.12	60.33	55.03	61.85	66.63	59.90	63.19
		55.70	**73.43**	**64.28**	**70.48**	**72.48**	**64.01**	**69.11**
	Online-Mem	55.12	60.27	56.12	58.58	59.57	54.54	63.73
		58.68	**73.85**	**67.78**	**66.77**	**67.31**	**60.07**	**70.49**
Bigram	Batch-Opt	77.06	73.05	64.75	71.25	69.63	66.20	66.53
		80.19	**84.15**	**80.34**	**79.36**	**76.01**	**70.87**	**73.11**
	Online-Mem	86.26	82.56	60.22	60.87	62.46	59.51	63.32
		89.58	**88.83**	**83.27**	**74.08**	**73.98**	**69.48**	**73.24**
Baselines	Subtractive Seg.	87.77	82.37	58.25	39.95	35.14	49.83	30.09
	Random	38.15	34.23	28.92	22.88	21.37	25.68	23.84

Table 3: Word token F-scores for each Bayesian learner across English, German, Spanish, Italian, Farsi, Hungarian, and Japanese. Results are given both for Unigram and Bigram learners, and include both the BatchOpt and Online-Mem learners. The F-score when compared against orthographic words is shown, with the adjusted F-score that includes "useful errors" in **bold**. A random-guess baseline is given along with model results for the subtractive segmenter with beam search from Lignos (2011). Higher token F-scores indicate better performance.

while *oop* would still be counted as incorrect since it is not.

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

Table 4: Examples of useful errors (with English glosses) made by learners in different languages. **True** words refer to the segmentation in the original corpus, while **Model** output represents the segmentation leading to a useful error.

Languages that fared more poorly when compared against the original "gold standard" benefit the most from the useful error analysis, underscoring the utility of this more nuanced metric. Focusing on the useful error results, we find as previous studies did that the bigram learners outperform the unigram learners. This suggests that the knowledge that words depend on previous words continues to be a useful one (as Goldwater et al. 2009, Pearl et al 2011, and Phillips & Pearl 2012 found for English), though this difference may be smaller for some languages (e.g., Farsi, Japanese). As with the unadjusted results, performance for English and German is very high (best score: 89.58), while for other languages the learners tend to fare less well (best score: 70.87–83.27), though still quite good if the goal is to generate a set of useful units from which to bootstrap further language acquisition.

Incorporating cognitive constraints into Bayesian learning with the Online-Mem learner touches somewhat on the "Less is More" (LiM) hypothesis (Newport, 1990), which supposes that cognitive limitations help – rather than hinder – language acquisition. Pearl et al. (2011) and Phillips & Pearl (2012) found that the constrained Online-Mem learner outperformed its ideal BatchOpt equivalent. Cross-linguistically, this pattern is less robust (unigram learners: only in English, Spanish, and Japanese (e.g. Spanish BatchOpt: 64.28 vs. Online-Mem 67.78); bigram learners: only in English, German, and Spanish (e.g. German BatchOpt: 84.15 vs. Online-Mem 88.83)). Thus, while there is some support for the idea that incorporating cognitive considerations into Bayesian learners might improve word segmentation results, it is not true for every language. Still, for all languages it does appear that adding psychological constraints does not significantly harm performance, especially once "useful errors" are taken into account. This suggests that Bayesian inference is a viable strategy for word segmentation cross-linguistically, even for learners who cannot perform optimal inference.

Importantly, the goal of early segmentation is not for the infant to segment perfectly as an adult would, but to provide a way to get the word segmentation process started. Given this goal, Bayesian segmentation seems effective for all these languages. Moreover, because our learners are looking for useful units, which can be realized in different ways across languages, they can identify foundational aspects of a language that are both smaller and larger than orthographic words.

Conclusion

We have demonstrated that Bayesian segmentation performs quite well as an initial learning strategy for many different languages, especially if the learner is measured by whether it identifies useful units. This not only supports Bayesian segmentation as a viable cross-linguistic strategy, but also suggests that a useful methodological norm for word segmentation research should be how well a learning strategy identifies units that can scaffold future language acquisition. By taking into account reasonable errors that identify such units, we bring our model evaluation into alignment with the actual goal of early word segmentation.

Acknowledgments

We 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, Jon Sprouse, Barbara Sarnecka, Michael Lee, Alex Ihler, and UCLA's phonology seminar for their helpful comments.

References

Baayen, R., Piepenbrock, R., & Gulikers, L. (1996). Celex2. *Linguistic Data Consortium*.

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.

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

Echols, C., Crowhurst, M., & Childers, J. (1997). The perception of rhythmic units in speech by infants and adults. *Journal of Memory and Language*, *36*, 202–225.

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

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

Fourtassi, A., Börschinger, B., Johnson, M., & Dupoux, E. (2013). Whyisenglishsoeasytosegment. In *Cognitive modeling and computational linguistics 2013* (pp. 1–10).

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.

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

Johnson, M. (2008). Unsupervised word segmentation for sesotho using adaptor grammars. In *Proceedings of the tenth meeting of the acl special interest group on computational morphology and phonology* (pp. 20–27).

Johnson, M., 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., & Aslin, R. (1995). Infants' detection of the sound patterns of words in fluent speech. *Cognitive Psychology*, *29*, 1–23.

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

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

Lignos, C. (2011). Modeling infant word segmentation. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning* (pp. 29–38).

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

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

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

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

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

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

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

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).

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.

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

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

Thiessan, 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.

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.