Combining multiple information types in Bayesian word segmentation

Gabriel Doyle and **Roger Levy**
Department of Linguistics
University of California, San Diego
La Jolla, CA 92093, USA
{gdoyle,rlevy}@ucsd.edu

Abstract

Humans identify word boundaries in continuous speech by combining multiple cues; existing state-of-the-art models, though, look at a single cue. We extend the generative model of Goldwater et al (2006) to segment using syllable stress as well as phonemic form. Our new model treats identification of word boundaries and prevalent stress patterns in the language as a joint inference task. We show that this model improves segmentation accuracy over purely segmental input representations, and recovers the dominant stress pattern of the data. Additionally, our model retains high performance even without single-word utterances. We also demonstrate a discrepancy in the performance of our model and human infants on an artificial-language task in which stress cues and transition-probability information are pitted against one another. We argue that this discrepancy indicates a bound on rationality in the mechanisms of human segmentation.

1 Introduction

For an adult speaker of a language, word segmentation from fluid speech may seem so easy that it barely needed to be learned. However, pauses in speech and word boundaries are not well correlated (Cole & Jakimik, 1980), word boundaries are marked by a conspiracy of partially-informative cues (Johnson & Jusczyk, 2001), and different languages mark their boundaries differently (Cutler & Carter, 1987). This makes the problem of unsupervised word segmentation acquisition, whether by a computational model or an infant, a daunting task.

Effective segmentation relies on the flexible integration of multiple types of segmentation cues, among them statistical regularities in phonemes and prosody, coarticulation, and allophonic variation. Infants begin using multiple segmentation cues within their first year of life (Johnson & Jusczyk, 2001). Despite this, many state-of-the-art models look at only one type of information: phonemes.

In this study, we expand an existing model to incorporate multiple cues, leading to an improvement in segmentation performance and opening new ways of investigating human segmentation acquisition. On the latter point, we show that rational learners can learn to segment without encountering words in isolation, and that human learners deviate from rationality in certain segmentation tasks.

2 Previous work

The prevailing unsupervised word segmentation systems (e.g., Brent, 1999; Goldwater, Griffiths, & Johnson, 2006; Blanchard & Heinz, 2008) use only phonemic information to segment speech. However, human segmenters use additional information types, notably stress information, in their segmentation. We present an overview of these phonemic models here before discussing the prosodic model expansion. A more complete review is available in Goldwater (2007).

2.1 Goldwater et al (2006)

The Goldwater et al model is related to Brent (1999)'s model, both of which use strictly phonemic information to segment. The model assumes that the corpus is generated by a Dirichlet process over

Proceedings of NAACL-HLT 2013, pages 117–126,
Atlanta, Georgia, 9–14 June 2013.

word bigrams.[1] We present a basic overview here, based on Sect. 5.5 of Goldwater, 2007. To generate the word w_i given the preceding word w_{i-1}:

1. Decide if bigram $b_i = \langle w_{i-1}, w_i \rangle$ is novel
2. If b_i non-novel, draw b_i from bigram lexicon
3. If b_i novel, decide whether w_i is novel
   a. If w_i non-novel, draw w_i from word lexicon
   b. If w_i novel, draw w_i from word-generating distribution P_0.

The Dirichlet process first decides whether to draw a non-novel ("nn") bigram, with probability proportional to the number of times the previous word has appeared in the corpus:

$$p(\langle w_{i-1}, w_i \rangle \text{ nn} | w_{i-1}) = \frac{n_{\langle w_{i-1}, \cdot \rangle}}{n_{\langle w_{i-1}, \cdot \rangle} + \alpha_1}, \qquad (1)$$

where $n_{\langle x,y \rangle}$ is the token count for bigram $\langle x, y \rangle$. If the bigram is non-novel, word w_i is drawn in proportion to the number of times it has appeared after w_{i-1} in the corpus:

$$p(w_i = x | \langle w_{i-1}, w_i \rangle \text{ nn}) = \frac{n_{\langle w_{i-1}, x \rangle}}{n_{\langle w_{i-1}, \cdot \rangle}} \qquad (2)$$

If the bigram is novel, this could either be due to w_i being a novel word or due to w_i being an existing word that had not appeared with w_{i-1} before. The probability of w_i being a non-novel word x is

$$p(w_i = x, w_i \text{ nn} | \begin{array}{c} \langle w_{i-1}, w_i \rangle \\ \text{novel} \end{array}) = \frac{b_{\langle \cdot, w_i \rangle}}{(b_{\langle \cdot, \cdot \rangle} + \alpha_0)}, \qquad (3)$$

where $b_{\langle .,. \rangle}$ is the count of word bigram types. Finally, if w_i is a new word, its phonemic form is generated from a distribution P_0. In the Goldwater et al model, this distribution is simply the product of the unigram probabilities of the phonemes, $P(\sigma_j)$, times the probability of a word boundary, $p_{\#}$, to end the word:

$$p(w_i = \sigma_1 \cdots \sigma_M | \begin{array}{c} w_i \\ \text{novel} \end{array}) = p_{\#}(1 - p_{\#})^{M-1} \prod P(\sigma_j) \qquad (4)$$

To segment an observed corpus, the model Gibbs samples over the possible word boundaries (utterance boundaries are assumed to be word boundaries).[2] The exchangability of draws from a Dirichlet process allows for Gibbs sampling of each possible boundary given all the others.

2.2 A cognitively-plausible variant

Phillips and Pearl (2012) make these Bayesian segmentation models more cognitively plausible in two ways. The first is to move from phonemes to syllables as the base representational unit from which words are constructed, as infants learn to categorize syllables before phonemes (Eimas, 1999). The second is to add memory and processing constraints on the learner. They find that syllable-based segmentation is better than phoneme-based segmentation in the bigram model (though worse in the unigram model), and that, counter-intuitively, the constrained learner outperforms the unconstrained learner. This improvement appears to be driven by better performance in segmenting more common words. In this work, we adopt the syllabified representation but retain the unconstrained rational learner assumption.

2.3 Other multiple-cue models

Some previous models have incorporated multiple cues, specifically the phonemic and stress information that our model will use. Two prominent examples are Christiansen, Allen, and Seidenberg (1998)'s connectionist model and Gambell and Yang (2006)'s algebraic model. The connectionist model places word boundaries where the combination of phonemic and stress information predict likely utterance boundaries, but does not include an explicit sense of "word", and performs only modestly on the segmentation task (boundary F-scores of .40-.45). The algebraic model also underperforms the Bayesian model (Phillips & Pearl, 2012) unless it includes the heuristic that there is a word boundary between any two stressed syllables. Our model presents a more general and completely unsupervised approach to segmentation with multiple cue-types.

[1] We will only discuss the bigram model here because it is more appropriate from both a cognitive perspective (it posits latent hierarchical structure) and engineering perspective (it segments more accurately) than the unigram model.

[2] The model assumes that utterance boundaries are generated just like other words, and includes an adjustable parameter $p_{\$}$ to account for their frequency.

In general, joint inference is becoming more common in language acquisition problems and has been shown to improve performance over single-feature inference. Examples include joint inference of a lexicon and phonetic categories (Feldman, Griffiths, & Morgan, 2009), joint inference of syntactic word order and word reference (Maurits, Perfors, & Navarro, 2009), and joint inference of word meanings and speaker intentions in child-directed speech (Frank, Goodman, & Tenenbaum, 2009).

3 Model design

Our model changes P_0 from a single-cue distribution, generating only phonemes, to a multiple-cue distribution that generates a stress form as well. This can improve segmentation performance and allows the investigation of rational segmentation behavior in a multiple-cue world.

In the original model, $P_0(w_i = \sigma_1 \cdots \sigma_M) \propto \prod_j P(\sigma_j)$, where $P(\sigma_j)$ is the frequency of the phoneme σ_j. In the multiple-cue model, we first generate a phonemic form w_i, then assign a stress pattern s_i to it.

$$\begin{aligned} P_0(w_i, s_i) &= P_W(w_i) P_S(s_i|M) \\ &= p_\#(1-p_\#)^{M-1} \prod_j^M P(\sigma_j) P_S(s_i|M) \quad (5) \end{aligned}$$

The phonemic form w_i has the same product-of-segments probability as the Goldwater et al model, but σ_j are now syllables instead of phonemes. We discuss the rationale behind this change in the next section.

The phonemic form is generated first, and the stress form is then drawn as a multinomial over all possible stress patterns with the same number of syllables as w_i. The stress distribution P_S is a multinomial distribution over word-length stress templates. P_S can be learned by the model based on a Dirichlet prior, but for simplicity in the present implementation, we estimate P_S as the plus-one-smoothed frequency of the stress patterns in the current segmentation. There are two stress levels (stressed or unstressed), and 2^M possible stress templates for a word of length M.[3]

Unlike phonemic forms, stress patterns are drawn as a whole word. This allows the model to capture a wide range of stress biases, although it prevents the model from generalizing biases across different word lengths. A potential future change to P_S that would allow for better generalization is discussed in Section 6.

3.1 On syllabification and stress

We change from segmenting on phonemes to segmenting on syllables in order to more easily implement stress information, which is a supersegmental feature most appropriately located on syllables. Syllabified data has been used in some previous models of segmentation, especially those using stress information or syllable-level transition probabilities (Christiansen et al., 1998; Swingley, 2005; Gambell & Yang, 2006; Phillips & Pearl, 2012).

For studying human word segmentation, Phillips and Pearl argue syllabified speech may be a more cognitively plausible testing ground. 3-month-old infants appear to have categorical representations of syllables (Eimas, 1999), three months before word segmentation appears (Borfeld, Morgan, Golinkoff, & Rathbun, 2005), and seven months before phoneme categorization (Werker & Tees, 1984). In addition, syllabification is assumed in much work on human word segmentation, especially in artificial-language studies (e.g., Thiessen & Saffran, 2003), which calculate statistical cues at the syllable level.

The assumption that syllable boundaries are known affects the baseline performance of the model, as it reduces the number of possible word boundary locations (since a word boundary is necessarily a syllable boundary). As such performance over syllabified data cannot be directly compared to performance on non-syllabified data.

It may seem that syllabification is so closely tied to word segmentation that including the former in a model of the latter leaves little to the model. However, the determinants of syllable boundaries are not the same as those for word boundaries. The prob-

[3] We do not assume that each word has one and only one stressed syllable, which would reduce the number of possible stress templates to M, for two reasons. First, in the current corpus, some words have citation forms with multiple stressed syllables. Second, in actual speech this assumption will not hold (e.g., many function words go unstressed).

lem of assigning syllable boundaries is a question of deciding where a boundary goes between two syllable nuclei, with the assumption that there must be a boundary there. The problem of assigning word boundaries is a question of deciding whether there is a boundary between two syllable nuclei, and if so, where it is. Knowing the syllable boundaries reduces the set of possible word boundaries, but does not directly address the question of how likely a boundary is. The difference in these tasks is supported by the three-month gap between syllable and word identification in infants.

4 Data

We use the Korman (1984) training corpus, as compiled by Christiansen et al. (1998), in this study. This is a 24493-word corpus of English spoken by adults to infants aged 6–16 weeks.[4] Phonemes, stresses, and syllable boundaries are the same as those used by Christiansen et al, which were based on citation forms in the MRC Psycholinguistic Database. All monosyllabic words were coded as stressed. Only utterances for which all words had citation forms were included.

This corpus is largely monosyllabic (87.3% of all word tokens), and heavily biased toward initial stress (89.2% of all multisyllable word tokens). No word is longer than three syllables, and most words have only one stressed syllable. A breakdown of the corpus by stress pattern is given in Table 1. This monosyllabic bias is an inherent property of English, not idiosyncratic to this corpus. The Bernstein-Ratner child-directed corpus is also over 80% monosyllabic. We expect that the results of segmentation on child-directed data will extend to adult speech, as the adult-directed corpus used by Gambell and Yang (2006) has an average word length of 1.17 syllables.

5 Experiments

We test the model on three problems. First, we show that the addition of stress information improves segmentation performance compared to a stress-less model. Next, we apply the model to a question in human segmentation acquisition. Finally, we look at a task where the rational model deviates from human performance.

Types		**Tokens**	
Stress pattern	Count	Stress pattern	Count
S	21402	S	523
SW	2231	SW	208
SS	389	WS	40
WS	284	SWW	24
SWW	182	SS	7
WSW	33	WSW	7
Other	5	Other	2

Table 1: Corpus stress patterns by types and tokens, showing an initial-stress bias in all lengths.

5.1 Parameter setting

The model has four free parameters: α_0 and α_1, which affect the likelihood of new words and bigrams, respectively, and $p_\#$ and $p_\$$, which affect the expected likelihood of word and utterance boundaries. Following Goldwater, Griffiths, and Johnson (2009), we set $\alpha_0 = 20$, $\alpha_1 = 100$, $p_\# = 0.8$ and $p_\$ = 0.5$ in all experiments.[5]

In all cases, the model performed five independent runs of 20000 iterations of Gibbs sampling the boundaries for the full corpus. Simulated annealing was performed during the burn-in period to improve convergence. All performance measures are reported as the mean of these five runs.

Performance is measured as word, boundary, and lexicon precision, recall, and F-scores. A word is matched iff both of its true boundaries are marked as boundaries and no internal boundaries are marked as word boundaries. Boundary counts omit utterance boundaries, which are assumed to be word boundaries. Lexical counts are based on word type counts.

5.2 Stress improves performance

We begin by showing that including a second cue type improves segmentation performance. We compare segmentation on a corpus with the attested stress patterns to that of a corpus without stress. With stress information included in the model, word/boundary/lexicon F-scores are

[4] Approximately 150 word tokens from the original corpus were omitted in our version of the corpus due to a disparity between recorded number of syllables and number of stresses.

[5] Performance was similar for a range of settings between 1 and 100 for α_0 and between 10 and 200 for α_1.

	With stress			Without stress		
	Word	Bnd	Lex	Word	Bnd	Lex
Prec	.76	.99	.75	.76	.99	.72
Rec	.61	.70	.87	.60	.69	.84
F	.68	.82	.80	.67	.82	.77

Table 2: Precision, recall, and F-score over corpora with and without stress information available. Stress information especially improves lexical performance.

.68/.82/.80. Without stress, performance drops to .67/.82/.77.[6] Full results are given in Table 2.

Stress information primarily improves lexicon performance, along with a small improvement in token segmentation. Accounting for stress reduces both false positives and negatives in the lexicon; the fact that the lexical improvement is greater than that for words or boundaries suggests that much of the improvement rests is on rare words.

These effects are small but significant. For word token performance, we performed a paired *t*-test on utterance token F-scores between the with- and without-stress models. This difference was significant ($t = 11.28, \mathit{df} = 8125, p < .001$). We performed a similar utterance-by-utterance test on boundaries; again a small singificant improvement was found ($t = 8.92, \mathit{df} = 6084, p < .001$). To assess lexicon performance, we calculated for each word type in the gold-standard lexicon the proportion of the five trials in which that word appeared in the learned lexicon for the two models. We then examined the words where the proportions differed between the models. 89 true words appeared more often in the with-stress lexicons; 40 appeared more often in the without-stress lexicons. (683 appeared equally often in both.) By a sign test, this is significant at $p < .001$. We also tested lexicon performance with a binomial test on the two models' lexicon accuracy; this result was marginal ($p = .06$).

The explicit tracking of stress information also improves the model's acquisition of the stress bias of the language. Acquisition of the stress bias is potentially useful for generalization; stress patterns can be used for an initial segmentation if few or none of the words are familiar. In practice, we see children use their stress biases to segment new words from English speech (Jusczyk, Houston, & Newsome, 1999) as well as artificial languages (Thiessen & Saffran, 2003).

We assess the learned stress bias by dividing up the corpus as the model has segmented it, and count the number of tokens with SW versus WS stress patterns.[7] With stress representation, the learned stress bias is 6.77:1, and without stress representation, the stress bias is lower, at 6.33:1. Although these are both underestimates of the corpus's true stress bias (7.86:1), the stressed model is stronger and a better estimate of the true value.

The model's performance can be compared to various baselines, but perhaps the strongest is one with every syllable boundary being a word boundary. This baseline represents a shift from boundary *precision* being at ceiling (as in the model) to boundary *recall* being at ceiling. In fact, due to the preponderance of monosyllabic words in English child-directed speech, this baseline outperforms the model on word and boundary F-scores (.68 and .82 in the model, .82 and .91 in the baseline). However, the baseline's lexicon is much worse than the model's (F=.80 in the with-stress model, F=.64 in the baseline), and the baseline fails to learn anything about the language's stress biases. In addition, the baseline oversegments, whereas both the model and infant segmenters undersegment (Peters, 1983). This raises an important question about what the model should seek to optimize: though the baseline is more accurate by token, no structure is learned; type performance is more important if we want to learn the underlying structure.

5.3 Are isolated words necessary?

We next use this model to test the necessity of isolated words in rational word segmentation. It is not immediately obvious how human learners begin to segment words from fluid speech. Stress biases and other phonological cues are dominant in all but the earliest of infant word segmentation (Johnson & Jusczyk, 2001). This raises a chicken-and-egg problem; if the cues infants favor to segment words, such as stress biases, are dependent on the words of the

[6] Recall that due to the syllabified data, these results are not directly comparable to unsyllabified results in previous work.

[7] Note this defines a stress bias for the stressless model as well.

language, how do they learn enough words to determine the cues' biases?

One existing proposal is that human learners develop their stress biases based on words frequently heard in isolation (Jusczyk et al., 1999). In English, these include names and common diminutives (e.g., *mommy*, *kitty*) that generally have initial stress. These single-word utterances could offer the segmenter an initial guess of the stress bias, by supposing that short utterances are single words and recording their stress patterns. The most common stress patterns in short utterances could then be used as an initial guess at the stress bias to bootstrap other words and thereby improve the learned stress bias.

We test the rational learner's need for such explicit bootstrapping by learning to segment a corpus with all single-word utterances removed. The corpus is produced by excising all single-word utterances from the Korman corpus. This results in a 22081-word corpus, 10% fewer tokens than in the original. However, it does not substantially change the lexicon; the number of distinct word types only drops from 811 to 806.

We compare performance only on ambiguous boundaries and lexicon, as these are comparable between the corpora, and find that the model performs almost equally well. Without single-word utterances, boundary and lexical F-scores are .81 and .80, compared to .82 and .80 with single-word utterances. This shows that rational learners are able to segment even without the possibility of bootstrapping stress patterns from single-word utterances.

5.4 Bounded rationality in human segmentation

Lastly, we use this model to examine rational performance in a multiple-cue segmentation task. We show that humans' segmentation does not adhere to these predictions, suggesting a bound on human rationality in word segmentation.

We consider an artificial language study by Thiessen and Saffran (2003). In this study, infants are exposed to an artificial language consisting of four bisyllabic word types uttered repeatedly without pauses. Each syllable appears in only one word type, so within-word transition probabilities are always 1, while across-word transition probabilities are less than 0.5. Segmentation strategies that hypothesize word boundaries at low transition probabilities or that seek to minimize the lexicon size will segment out the four word types as expected.

Against bias, with TP			
AB	CD	CD	AB
WS	WS	WS	WS

With bias, against TP				
A	BC	DC	DA	B
W	SW	SW	SW	S

Table 3: Examples of segmenting an artificial language according to transition probabilities (top) or stress bias (bottom), when the true words have weak-strong stress. Vertical lines represent word boundaries. The top segmentation produces a smaller lexicon, but the bottom segmentation produces primarily words with the preferred stress pattern.

Segmentation in the experiment is complicated by the presence of stress in the artificial language. Depending on the condition, the words are either all strong-weak or all weak-strong. In the first condition, segmenting according to transition probabilities, lexicon size, or English stress bias favors the same segmentation. In the second condition, though, segmenting by the English stress bias to yield a lexicon of strong-weak words requires boundaries in the middle of the words. The segmenter must decide whether transition probabilities or preferred stress patterns are more important in segmentation. This situation is illustrated in Table 3, with a corpus consisting of two word types, AB and CD, each with weak-strong stress.

Thiessen and Saffran found that seven-month-old English-learning infants consistently segmented according to the transition probabilities, regardless of stress. However, nine-month-olds segmented according to the English stress bias, even if this meant going against the transition probabilities.

Intuitively, this could be rational behavior according to our model. A child's increasing age means more exposure to data, potentially leading the child to develop more confidence in the stress bias. As confidence in the stress bias increases, the cost of segmenting against it increases as well. A sufficiently strong stress preference could lead the segmenter to accept a large lexicon, all of whose words have the preferred stress pattern, over a small lexicon, all of whose words have the dispreferred stress pattern.

To judge by the Korman corpus, English has a stress bias of approximately 7:1 in favor of SW bisyllabic stress over WS.[8] If human segmentation behavior follows the rational model, the model should predict segmentation to favor strong-weak words over the transition probabilities when the stress bias is approximately this strong.

We test this rationality hypothesis with a smaller version of the Thiessen and Saffran artificial language, consisting of 48 tokens.[9] In one version, all tokens have the preferred SW pattern, and in the other all tokens have the dispreferred WS pattern. We then adjust the P_S distribution such that $P_S(SW|M = 2) = b * P_S(WS|M = 2)$, where b is the bias ratio. We run the model otherwise the same as in the previous experiments, except with 10 runs instead of 5.

Contrary to this hypothesis, the model's segmentation with $b = 7$ was the same whether the true words were strong-weak or weak-strong. In all ten runs, transition probabilities dictated the segmentation. To switch to stress-based segmentation, the bias must be orders of magnitude greater than the English bias. Figure 1 shows the proportions of runs in the weak-strong condition that show segmentation according to the stress bias, as the bias increases by factors of 10. When $b = 10000$, three of the ten runs segmented according to the stress bias; below that, the stress bias did not affect the rational model's segmentation.

Why is this? In the Bayesian model, the stress bias of a language affects only the $P_S(s_i|M)$ term in the P_0 distribution, so non-novel words are not penalized for their stress pattern. The model pays only once to create a word; once the word is generated, no matter how a priori implausible the word was, it may be cheaply drawn again as a non-novel word. This effect can be illustrated with a brief calculation.

Consider a corpus built from four bisyllabic word types (AB, CD, EF, GH), each appearing N times. If the corpus is segmented against the transition probabilities, the resulting lexicon will have 16 bisyllabic word types (BA, BC, BE, BG, DA, etc.), each occurring approximately $\frac{N}{4}$ times.

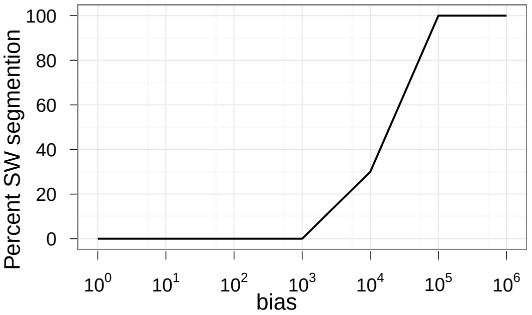

Figure 1: Percentage of runs segmented with the stress bias, against transition probabilities, as bias varies. At English-level biases, the rational model still overrules the stress bias when segmenting.

The probability of the against-bias corpus (C_{WS}) is proportional to the probability of generating the four word types, and then drawing them non-novelly from the lexicon.[10] (To simplify the calculations, we use the unigram version of the Goldwater et al model.)

$$p(C_{WS}) \propto P_W^4 P_S(WS)^4 (N!)^4 \frac{1}{4N!} \qquad (6)$$

The first two terms are the probability of generating the four word types (Eqn. 5);[11] the second two terms are the Dirichlet process draws from the existing lexicon N times each (Eqn. 2). By comparison, the probability of the with-bias corpus C_{SW} depends on generating the 16 word types, and drawing each non-novelly $\frac{N}{4}$ times.

$$p(C_{SW}) \propto P_W^{16} P_S(SW)^{16} \left(\frac{N}{4}!\right)^{16} \frac{1}{4N!} \qquad (7)$$

Given an SW bias b and a uniform distribution over syllables (so $P_W = \frac{1}{64}$), we find:

$$\frac{p(C_{WS})}{p(C_{SW})} = 64^{12}\frac{(b+1)^{12}}{b^{16}}\frac{(N!)^4}{(\frac{N}{4}!)^{16}} \qquad (8)$$

[8] The specific bias varies from corpus to corpus, but this appears to be a representative value.

[9] The 48 tokens come from four word types, with two types appearing 16 times and the other two appearing 8 times, mimicking the relative frequencies of Thiessen and Saffran's languages. Their test language had 270 tokens.

[10] It is also possible to generate this corpus by re-drawing the words novelly, but this is much less likely than non-novel draws.

[11] Because all syllables have equal unigram probabilities, the probability of all words' phonemic forms are equal, and will be written as P_W.

This equation shows that the rational model is heavily biased toward the segmentation that fits the transition probabilities. Increasing the stress bias b or decreasing the number of observed word tokens makes the rational model more likely to segment with the stress bias (against transition probabilities), but as we see in the experimental results, the stress bias must be very strong to overcome the efficient lexicon that the transition probability segmentation provides.

Since humans do not show this same inherent bias (or quickly lose it as they acquire the stress bias), we can ask how humans deviate from rationality. One possibility is that humans simply do not segment in this Bayesian manner. However, previous work (Frank, Goldwater, Griffiths, & Tenenbaum, 2010) has shown that human word segmentation shows similar behavior to a resource-limited Bayesian model. Equation 8 suggests that human segmentation could deviate from rationality by having an effectively stronger bias than English would suggest (reducing the first fraction)[12] or, as with Phillips and Pearl's constrained learners, by having effectively less input than the model assumes (reducing the second fraction).

6 Future work

Introducing stress into the Bayesian segmentation model suggests a few additional expansions. One possibility is to add other cues into the generative model via P_0. Any cue that is based on the word itself can be added in this way, with little change to the general model structure. Phonotactics can be added using an n-gram distribution for P_0 (Blanchard & Heinz, 2008). Coarticulation between adjacent phonemes is also used in human segmentation (Johnson & Jusczyk, 2001), so the P_0 distribution could predict higher within-word coarticulation. Integrating additional cues used by human segmenters extends the investigation of the bounds on rationality in human segmentation and in balancing multiple conflicting cues.

A more complex view of the stress system of a language may also be useful. One possibility is to place a Dirichlet prior over the stress templates and allow P_S to be learned as a latent variable in the model. Another possibility is to treat the stress templates more generally; in the present implementation, knowledge of the preferred stress patterns for word of one length tells the segmenter nothing about preferred stress patterns in another length. Cross-linguistically common stress rules (e.g., those that place stress a certain number of syllables from the left or right edge of a word) can be coded into P_S to improve generalization. Each rule dictates a specific stress pattern for each word length. When a word is generated in the Dirichlet process, the generative model would decide whether to assign stress according to one of these rules or to assign lexical stress from a default multinomial distribution. (This "default" distribution would handle idiosyncratic stress assignments, as one might see with names or morphologically complex words, like Spanish reflexive verbs.) A sparse prior over these rules, asymmetrically weighted against the default category, will encourage the model to explain as much of the observed stress patterns as possible with a few dominant rules, improving the phonological structure that the segmenter learns.

Improving the realism of the data is also important. The corpora used in much of segmentation research are idealized representations of the true data, and the dictionary-based phoneme and stress patterns used in this study are no exception. This ideal setting may paint a skewed picture of the segmentation problem, by providing a more consistent and learnable data source than humans actually receive. Elsner, Goldwater, and Eisenstein (2012)'s model unifying lexical and phonetic acquisition takes a significant step in showing that a rational segmenter can handle noisy input by recognizing phonetic variants of a base form. In terms of stress representations, dictionary-based stress has been standard in previous work (Christiansen et al., 1998; Gambell & Yang, 2006; Rytting, Brew, & Fosler-Lussier, 2010), but it is important to confirm such results against a (currently nonexistent) corpus with stresses based on the actual utterances. Effective use of stress in a less idealized setting may require a more complex representation of stress in the model.

[12] A potential source of an inflated bias is infants' preference for strong-weak patterns. Jusczyk, Cutler, and Redanz (1993) found English-hearing infants listened longer to strong-weak patterns than weak-strong. This could lead to overestimation of the stress bias by making possible strong-weak segmentations more prominent in the segmenter's mind.

7 Conclusion

Effective word segmentation combines multiple factors to make predictions about word boundaries. We extended an existing Bayesian segmentation model to account for two factors, phonemes and stress, when segmenting. This improves segmentation performance and opens up new possibilities for comparing rational segmentation and human segmentation.

Acknowledgments

This research was partially supported by an Alfred P. Sloan Fellowship to RL and by NSF award 0830535. We also appreciate the feedback of the reviewers and the members of the UCSD Computational Psycholinguistics Lab.

References

Blanchard, D., & Heinz, J. (2008). Improving word segmentation by simultaneously learning phonotactics. In *Proceedings of CoNLL* (pp. 65–72).

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

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

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

Cole, R., & Jakimik, J. (1980). A model of speech perception. In *Perception and production of fluent speech* (pp. 136–163). Hillsdale, NJ: Erlbaum.

Cutler, A., & Carter, D. (1987). The predominance of strong initial syllables in the English vocabulary. *Comp. Speech Lang.*, *2*, 133–142.

Eimas, P. (1999). Segmental and syllabic representations in the perception of speech by young infants. *Journal of the Acoustic Society of America*, *105*, 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 ACL.*

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

Frank, M., Goldwater, S., Griffiths, T., & Tenenbaum, J. (2010). Modeling human performance in statistical word segmentation. *Cognition*.

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.* (Unpublished manuscript)

Goldwater, S. (2007). *Nonparametric Bayesian models of lexical acquisition*. Unpublished doctoral dissertation, Brown Univ.

Goldwater, S., Griffiths, T., & Johnson, M. (2006). Contextual dependencies in unsupervised word segmentation. In *Proceedings of Coling/ACL.*

Goldwater, S., Griffiths, T. L., & Johnson, M. (2009). A Bayesian framework for word segmentation: Exploring the effects of context. *Cognition*, *112*, 21–54.

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

Jusczyk, P., Cutler, A., & Redanz, N. (1993). Preference for predominant stress patterns of English words. *Child Development*, *64*, 675–687.

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

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

Maurits, L., Perfors, A., & Navarro, D. (2009). Joint acquisition of word order and word reference. In *Proceedings of 31st annual conference of the Cognitive Science Society.*

Peters, A. (1983). *The units of language acqui-*

sition: Monographs in applied psycholinguistics. Cambridge Univ. Press.

Phillips, L., & Pearl, L. (2012). "less is more" in Bayesian word segmentation: When cognitively plausible learners outperform the ideal. In *Proceedings of the 34th annual conference of the cognitive science society.*

Rytting, C. A., Brew, C., & Fosler-Lussier, E. (2010). Segmenting words from natural speech: subsegmental variation in segmental cues. *Journal of Child Language*, *37*, 513–543.

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

Thiessen, E. D., & Saffran, J. R. (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 and Development*, *7*, 49–63.