# Exploring the Role of Stress in Bayesian Word Segmentation using Adaptor Grammars

**Benjamin Börschinger**[1,2] **Mark Johnson**[1,3]

[1]Department of Computing, Macquarie University, Sydney, Australia
[2]Department of Computational Linguistics, Heidelberg University, Heidelberg, Germany
[3]Santa Fe Institute, Santa Fe, USA
{benjamin.borschinger|mark.johnson}@mq.edu.au

## Abstract

Stress has long been established as a major cue in word segmentation for English infants. We show that enabling a current state-of-the-art Bayesian word segmentation model to take advantage of stress cues noticeably improves its performance. We find that the improvements range from 10 to 4%, depending on both the use of phonotactic cues and, to a lesser extent, the amount of evidence available to the learner. We also find that in particular early on, stress cues are much more useful for our model than phonotactic cues by themselves, consistent with the finding that children do seem to use stress cues before they use phonotactic cues. Finally, we study how the model's knowledge about stress patterns evolves over time. We not only find that our model correctly acquires the most frequent patterns relatively quickly but also that the Unique Stress Constraint that is at the heart of a previously proposed model does not need to be built in but can be acquired jointly with word segmentation.

## 1 Introduction

Among the first tasks a child language learner has to solve is picking out words from the fluent speech that constitutes its linguistic input.[1] For English, stress has long been claimed to be a useful cue in infant word segmentation (Jusczyk et al., 1993; Jusczyk et al., 1999b), following the demonstration of its effectiveness in adult speech processing (Cutler et al., 1986). Several studies have investigated the role of stress in word segmentation using computational models, using both neural network and "algebraic" (as opposed to "statistical") approaches (Christiansen et al., 1998; Yang, 2004; Lignos and Yang, 2010; Lignos, 2011; Lignos, 2012). Bayesian models of word segmentation (Brent, 1999; Goldwater, 2007), however, have until recently completely ignored stress. The sole exception in this respect is Doyle and Levy (2013) who added stress cues to the Bigram model (Goldwater et al., 2009), demonstrating that this leads to an improvement in segmentation performance. In this paper, we extend their work and show how to integrate stress cues into the flexible Adaptor Grammar framework (Johnson et al., 2007). This allows us to both start from a stronger baseline model and to investigate how the role of stress cues interacts with other aspects of the model. In particular, we find that phonotactic cues to word-boundaries interact with stress cues, indicating synergistic effects for small inputs and partial redundancy for larger inputs. Overall, we find that stress cues add roughly 6% token f-score to a model that does not account for phonotactics and 4% to a model that already incorporates phonotactics. Relatedly and in line with the finding that stress cues are used by infants before phonotactic cues (Jusczyk et al., 1999a), we observe that phonotactic cues require more input than stress cues to be used efficiently. A closer look at the knowledge acquired by our models shows that the Unique Stress Constraint of Yang (2004) can be acquired jointly with segmenting the input instead

[1] The datasets and software to replicate our experiments are available from http://web.science.mq.edu.au/~bborschi/

Transactions of the Association for Computational Linguistics, 2 (2014) 93–104. Action Editor: Stefan Riezler.
Submitted 12/2013; Published 2/2014.

of having to be pre-specified; and that our models correctly identify the predominant stress pattern of the input but underestimate the frequency of iambic words, which have been found to be missegmented by infant learners.

The outline of the paper is as follows. In Section 2 we review prior work. In Section 3 we introduce our own models. In Section 4 we explain our experimental evaluation and its results. Section 5 discusses our findings, and Section 6 concludes and provides some suggestions for future research.

## 2 Background and related work

Lexical stress is the "accentuation of syllables within words" (Cutler, 2005) and has long been argued to play an important role in adult word recognition. Following Cutler and Carter (1987)'s observation that stressed syllables tend to occur at the beginnings of words in English, Jusczyk et al. (1993) investigated whether infants acquiring English take advantage of this fact. Their study demonstrated that this is indeed the case for 9 month olds, although they found no indication of using stressed syllables as cues for word boundaries in 6 month olds. Their findings have been replicated and extended in subsequent work (Jusczyk et al., 1999b; Thiessen and Saffran, 2003; Curtin et al., 2005; Thiessen and Saffran, 2007). A brief summary of the key findings is as follows: English infants treat stressed syllables as cues for the beginnings of words from roughly 7 months of age, suggesting that the role played by stress needs to be acquired, and that this requires antecedent segmentation by non-stress-based means (Thiessen and Saffran, 2007). They also exhibit a preference for low-pass filtered stress-initial words from this age, suggesting that it is indeed stress and not other phonetic or phonotactic properties that are treated as a cue for word-beginnings (Jusczyk et al., 1993). In fact, phontactic cues seem to be used later than stress cues (Jusczyk et al., 1999a) and seem to be outweighed by stress cues (Mattys and Jusczyk, 2000).

The earliest computational model for word segmentation incorporating stress cues we are aware of is the recurrent network model of Christiansen et al. (1998) and Christiansen and Curtin (1999). They only reported a word-token f-score of 44% (roughly, segmentation accuracy: see Section 4), which is considerably below the performance of subsequent models, making a direct comparison complicated. Yang (2004) introduced a simple incremental algorithm that relies on stress by embodying a Unique Stress Constraint (USC) that allows at most a single stressed syllable per word. On pre-syllabified child directed speech, he reported a word token f-score of 85.6% for a non-statistical algorithm that exploits the USC. While the USC has been argued to be near-to-universal and follows from the "culminative function of stress" (Fromkin, 2001; Cutler, 2005), the high score Yang reported crucially depends on every word token carrying stress, including function words. More recently, Lignos (2010, 2011, 2012) further explored Yang's original algorithm, taking into account that function words should not be assumed to possess lexical stress cues. While his scores are in line with those reported by Yang, the importance of stress for this learner were more modest, providing a gain of around 2.5% (Lignos, 2011). Also, the Yang/Lignos learner is unable to acquire knowledge about the role stress plays in the language, e.g. that stress tends to fall on particular positions within words.

Doyle and Levy (2013) extend the Bigram model of Goldwater et al. (2009) by adding stress-templates to the lexical generator. A stress-template indicates how many syllables the word has, and which of these syllables (if any) are stressed. This allows the model to acquire knowledge about the stress patterns of its input by assigning different probabilities to the different stress-templates. However, Doyle and Levy (2013) do not directly examine the probabilities assigned to the stress-templates; they only report that their model does slightly prefer stress-initial words over the baseline model by calculating the fraction of stress-initial word types in the output segmentations of their models. They also demonstrate that stress cues do indeed aid segmentation, although their reported gain of 1% in token f-score is even smaller than that reported by Lignos (2011). Our own approach differs from theirs in assuming phonemic rather than pre-syllabified input (although our model could, trivially, be run on syllabified input as well) and makes use of Adaptor Grammars instead of the Goldwater et al. (2009) Bigram model, providing us with a flexible framework for exploring the usefulness of stress in different models.

Adaptor Grammar (Johnson et al., 2007) is a grammar-based formalism for specifying non-parametric hierarchical models. Previous work explored the usefulness of, for example, syllable-structure (Johnson, 2008b; Johnson and Goldwater, 2009) or morphology (Johnson, 2008b; Johnson, 2008a) in word segmentation. The closest work to our own is Johnson and Demuth (2010) who investigate the usefulness of tones for Mandarin phonemic segmentation. Their way of adding tones to a model of word segmentation is very similar to our way of incorporating stress.

## 3 Models

We give an intuitive description of the mathematical background of Adaptor Grammars in 3.1, referring the reader to Johnson et al. (2007) for technical details. The models we examine are derived from the collocational model of Johnson and Goldwater (2009) by varying three parameters, resulting in 6 models: two baselines that do not take advantage of stress cues and either do or do not use phonotactics, as described in Section 3.2; and four stress models that differ with respect to the use of phonotactics, and as to whether they embody the Unique Stress Constraint introduced by Yang (2004). We describe these models in section 3.3.

### 3.1 Adaptor Grammars

Briefly, an Adaptor Grammar (AG) can be seen as a probabilistic context-free grammar (PCFG) with a special set of *adapted* non-terminals. We use underlining to distinguish adapted non-terminals ($\underline{\mathrm{X}}$) from non-adapted non-terminals ($\mathrm{Y}$). The distribution for each adapted non-terminal $\underline{\mathrm{X}}$ is drawn from a Pitman-Yor Process which takes as its base-distribution the tree-distribution over trees rooted in $\underline{\mathrm{X}}$ as defined by the PCFG. As an effect, each adapted non-terminal can be seen as having associated with it a cache of previously-generated subtrees that can be reused without having to be regenerated using the individual PCFG rules. This allows AGs to learn reusable sub-trees such as words, sequences of words, or smaller units such as Onsets and Codas. Thus, while ordinary PCFGs have a finite number of parameters (one probability for each rule), Adaptor Grammars in addition have a parameter for every possible complete tree rooted in any of its adapted non-terminals, leading to a potentially infinite number of such parameters. The Pitman-Yor Process induces a rich-get-richer dynamics, biasing the model towards identifying a small set of units that can be reused as often as possible. In the case of word segmentation, the model will try to identify as compact a lexicon as possible to segment the unsegmented input.

### 3.2 Baseline models

Our starting point is the state-of-the-art AG model for word segmentation, Johnson and Goldwater (2009)'s colloc3-syll model, reproduced in Figure 1.[2] The model assumes that words are grouped into larger collocational units that themselves can be grouped into even larger collocational units. This accounts for the fact that in natural language, there are strong word-to-word dependencies that need to be accounted for if severe undersegmentations of the form "is in the" are to be avoided (Goldwater, 2007; Johnson and Goldwater, 2009; Börschinger et al., 2012). It also uses a language-independent form of syllable structure to constrain the space of possible words. Finally, this model can learn word-initial onsets and word-final codas. In a language like English, this ability provides additional cues to word-boundaries as certain onsets are much more likely to occur word-initially than medially (e.g. "bl" in "black"), and analogously for certain codas (e.g. "dth" in "width" or "ngth" in "strength").

We define an additional baseline model by replacing rules (5) and (6) by (17), and deleting rules (7) to (12). This removes the model's ability to use phonotactic cues to word-boundaries.

$$\underline{\mathrm{Word}} \rightarrow \mathrm{Syll}\,(\,\mathrm{Syll}\,)\,(\,\mathrm{Syll}\,)\,(\,\mathrm{Syll}\,) \qquad (17)$$

We refer to the model in Figure 1 as the colloc3-phon model, and the model that results from substituting and removing rules as described as the colloc3-nophon model. Alternatively, one could limit the models ability to capture word-to-word dependencies by removing rules (1) to (3). This results

[2] We follow Johnson and Goldwater (2009) in limiting the length of possible words to four syllables to speed up runtime. In pilot experiments, this choice did not have a noticeable effect on segmentation performance.

$$
\begin{aligned}
\text{Collocations3} &\to \underline{\text{Collocation3}}^{+} && (1)\\
\underline{\text{Collocation3}} &\to \underline{\text{Collocation2}}^{+} && (2)\\
\underline{\text{Collocation2}} &\to \underline{\text{Collocation}}^{+} && (3)\\
\underline{\text{Collocation}} &\to \underline{\text{Word}}^{+} && (4)\\
\underline{\text{Word}} &\to \text{SyllIF} && (5)\\
\underline{\text{Word}} &\to \text{SyllI}\,(\,\text{Syll}\,)\,(\,\text{Syll}\,)\,\text{SyllF} && (6)\\
\text{SyllIF} &\to (\,\underline{\text{OnsetI}}\,)\,\text{RhymeF} && (7)\\
\text{SyllI} &\to (\,\underline{\text{OnsetI}}\,)\,\text{Rhyme} && (8)\\
\text{SyllF} &\to (\,\underline{\text{Onset}}\,)\,\text{RhymeF} && (9)\\
\underline{\text{CodaF}} &\to \text{Consonant}^{+} && (10)\\
\text{RhymeF} &\to \text{Vowel}\,(\,\underline{\text{CodaF}}\,) && (11)\\
\underline{\text{OnsetI}} &\to \text{Consonant}^{+} && (12)\\
\text{Syll} &\to (\,\underline{\text{Onset}}\,)\,\text{Rhyme} && (13)\\
\text{Rhyme} &\to \text{Vowel}\,(\,\underline{\text{Coda}}\,) && (14)\\
\underline{\text{Onset}} &\to \text{Consonant}^{+} && (15)\\
\underline{\text{Coda}} &\to \text{Consonant}^{+} && (16)
\end{aligned}
$$

Figure 1: The baseline model. We use regular-expression notation to abbreviate multiple rules. $\text{X}^{\{n\}}$ stands for up to $n$ repetitions of $\text{X}$, brackets indicate optionality, and $\text{X}^{+}$ stands for one or more repetitions of $\text{X}$. $\underline{\text{X}}$ indicates an adapted non-terminal. Rules that introduce terminals for the pre-terminals $\text{Vowel}$, $\text{Consonant}$ are omitted. Refer to the main text for an explanation of the grammar.

in the colloc-model (Johnson, 2008b) that has previously been found to behave similarly to the Bigram model used in Doyle and Levy (2013) (Johnson, 2008b; Börschinger et al., 2012). We performed experiments with the colloc-model as well and found similar results to Doyle and Levy (2013) which are, while overall worse, similar in trend to the results obtained for the colloc3-models. For the rest of the paper, therefore, we will focus on variants of the colloc3-model.

### 3.3 Stress-based models

In order for stress cues to be helpful, the model must have some way of associating the position of stress with word-boundaries. Intuitively, the reason stress helps infants in segmenting English is that a stressed syllable is a reliable indicator of the beginning of a word (Jusczyk et al., 1993). More generally, if there is a (reasonably) reliable relationship between the position of stressed syllables and beginnings (or endings) of words, a learner might exploit this relationship. In a Bayesian model, this intuition can be captured by modifying the lexical generator, that is, the distribution that generates $\underline{\text{Word}}$s.

$$
\begin{aligned}
\underline{\text{Word}} &\to \{\,\text{SSyll} \mid \text{USyll}\,\}^{\{1,4\}} && (18)\\
\text{SSyll} &\to (\,\underline{\text{Onset}}\,)\,\text{RhymeS} && (19)\\
\text{USyll} &\to (\,\underline{\text{Onset}}\,)\,\text{RhymeU} && (20)\\
\text{RhymeS} &\to \text{Vowel} * (\,\underline{\text{Coda}}\,) && (21)\\
\text{RhymeU} &\to \text{Vowel}\,(\,\underline{\text{Coda}}\,) && (22)\\
\underline{\text{Onset}} &\to \text{Consonant}^{+} && (23)\\
\underline{\text{Coda}} &\to \text{Consonant}^{+} && (24)
\end{aligned}
$$

Figure 2: Description of the all-stress-patterns model. We use $\text{X}^{\{m,n\}}$ for "at least $m$ and at most $n$ repetitions of $\text{X}$" and $\{\,\text{X} \mid \text{Y}\,\}$ for "either $\text{X}$ or $\text{Y}$". Stress is associated with a vowel by suffixing it with the special terminal symbol $*$, leading to a distinction between stressed ($\text{SSyll}$) and unstressed ($\text{USyll}$) syllables. A word can consist of any possible sequence of up to four syllables, as indicated by the regular-expression notation. By additionally adding initial and final variants of $\text{SSyll}$ and $\text{USyll}$ as in Figure 1, phonotactics can be combined with stress cues.

Here, changing the lexical generator corresponds to modifying the rules expanding $\underline{\text{Word}}$. A straightforward way to modify it accordingly is to enumerate all possible sequences of stressed and unstressed syllables.[3] A lexical generator like this is given in Figure 2. In the data, stress cues are represented using a special terminal "$*$" that follows a stressed vowel, as illustrated in Figure 3. In the grammar, "$*$" is constrained to only surface following a $\text{Vowel}$, rendering a syllable in which it occurs stressed ($\text{SSyll}$). Syllables that do not contain a "$*$" are considered unstressed ($\text{USyll}$). By performing inference for the probabilities assigned to the different expansions of rule (18), our models can, for example, learn that a bi-syllabic word that is stress-initial (a trochee) is more probable than one that puts stress on the second syllable (an iamb). This (partly) captures the tendency of English for stress-initial words and thus provide an additional cue for identifying words; and it is exactly the kind of preference infant learners of English seem to acquire (Jusczyk

[3] This is, in essence, also the strategy chosen by Doyle and Levy (2013).

| grammar | phon | stress | USC |
|---|---|---|---|
| colloc3-nophon | | | |
| colloc3-phon | • | | |
| colloc3-nophon-stress | | • | |
| colloc3-phon-stress | • | • | |
| colloc3-nophon-stress-usc | | • | • |
| colloc3-phon-stress-usc | • | • | • |

Table 1: The different models used in our experiments. "phon" indicates whether phonotactics are used, "stress" whether stress cues are used and "usc" whether the Unique Stress Constraint is assumed.

| | |
|---|---|
| orthographic | the **do**-gie |
| no-stress | dh ah d ao g iy |
| stress | dh ah d ao * g iy |

Figure 3: Illustration of the input-representation we choose. We indicate primary stress according to the dictionary with bold-face in the orthography. The phonemic transcription uses ARPABET and is produced using an extended version of CMUDict. Primary stress is indicated by inserting the special symbol "*" after the vowel of a stressed syllable.

et al., 1993).

We can combine this lexical generator with the colloc3-nophon baseline, resulting in the colloc3-nophon-stress model. We can also add phonotactics to the lexical generator in Figure 2 by adding initial and final variants of SSyll and USyll, analogous to rules (5) to (12) in Figure 1. This yields the colloc3-phon-stress model. We can also add the Unique Stress Constraint (USC) (Yang, 2004) by excluding all variants of rule (18) that generate two or more stressed syllables. For example, while the lexical generator for the colloc3-nophon-stress model will include the rule Word $\rightarrow$ SSyll SSyll, the lexical generator embodying the USC lacks this rule. We refer to the models that include the USC as colloc3-nophon-stress-usc and colloc3-phon-stress-usc models. A compact overview of the six different models is given in Table 1.

## 4 Experiments

We evaluate our models on several corpora of child directed speech. We first describe the corpora we used, then the experimental methodology employed and finally the experimental results. As the trend is comparable across all corpora, we only discuss in detail results obtained on the Alex corpus. For completeness, however, Table 3 reports the "standard" evaluation of performing inference over all of the three corpora.

### 4.1 Corpora and corpus creation

Following Christiansen et al. (1998) and Doyle and Levy (2013), we use the Korman corpus (Korman, 1984) as one of our corpora. It comprises child-directed speech for very young infants, aged between 6 and 16 weeks and, like all other corpora used in this paper, is available through the CHILDES database (MacWhinney, 2000). We derive a phonemicized version of the corpus using an extended version of CMUDict (Carnegie Mellon University, 2008)[4], as we were unable to obtain the stress-annotated version of this corpus used in previous experiments. The phonemicized version is produced by replacing each orthographic word in the transcript with the first pronunciation given by the dictionary. CMUDict also annotates lexical stress, and we use this information to add stress cues to the corpus. We only code primary lexical stresses in the input, ignoring secondary stresses in line with experimental work that indicates that human listeners are capable of reliably distinguishing primary and secondary stress (Mattys, 2000). Due to the very low frequency of words with 3 or more syllables in these corpora, this choice has very little effect on the number of stress cues available in the input. Our version of the Korman corpus contains, in total, 11413 utterances. Unlike Christiansen et al. (1998), Yang (2004), and Doyle and Levy (2013), we follow Lignos and Yang (2010) in making the more realistic assumption that the 94 mono-syllabic function words listed by Selkirk (1984) never surface with lexical stress. As function words account for roughly 50% of the tokens but only roughly 5% of the types in our corpora, this means that the type and token distribution of stress patterns differs dramatically in all our corpora, as can be seen from Table 2.

We also added stress information to the Brent-Bernstein-Ratner corpus (Bernstein-Ratner, 1987; Brent, 1999), following the procedure just outlined. This corpus is a de-facto standard for evaluat-

[4] `http://svn.code.sf.net/p/cmusphinx/code/trunk/cmudict/cmudict.0.7a`

| Pattern | brent | | korman | | alex | |
|---|---|---|---|---|---|---|
| | Tok | Typ | Tok | Typ | Tok | Typ |
| $W^+$ | .48 | .07 | .47 | .08 | .44 | .05 |
| $SW^*$ | .49 | .86 | .49 | .86 | .52 | .87 |
| $WSW^*$ | .03 | .07 | .03 | .06 | .04 | .07 |
| Other | .00 | .00 | .00 | .00 | .00 | .00 |

Table 2: Relative frequencies for stress patterns for the corpora used in our study. $X^*$ stands for 0 or more, $X^+$ for one or more repetitions of $X$, and S for a stressed and W for an unstressed syllable. Note the stark asymmetry between type and token frequencies for unstressed words. Up to two-decimal places, patterns other than the ones given have relative frequency 0.00 (frequencies might not sum to 1 as an artefact of rounding to 2 decimal places).

ing models of Bayesian word segmentation (Brent, 1999; Goldwater, 2007; Goldwater et al., 2009; Johnson and Goldwater, 2009), comprising in total 9790 utterances.

As our third corpus, we use the Alex portion of the Providence corpus (Demuth et al., 2006; Börschinger et al., 2012). A major benefit of the Providence corpus is that the video-recordings from which the transcripts were produced are available through CHILDES alongside the transcripts. This will allow future work to rely on even more realistic stress cues that can be derived directly from the acoustic signal. While beyond the scope of this paper, we believe choosing a corpus that makes richer information available will be important for future work on stress (and other acoustic) cues. Another major benefit of the Alex corpus is that it provides longitudinal data for a single infant, rather than being a concatenation of transcripts collected from multiple children, such as the Korman and the Brent-Bernstein-Ratner corpus. In total, the Alex corpus comprises 17948 utterances.

Note that despite the differences in age of the infants and overall make-up of the corpora, the distribution of stress patterns across the corpora is roughly the same, as shown by Table 2 for the first 10,000 utterances of each of the corpora. This suggests that the distribution of stress patterns both at a token and type level is a robust property of English child-directed speech.

### 4.2 Evaluation procedure

The aim of our experiments is to understand the contribution of stress cues to the Bayesian word segmentation models described in Section 3. To get an idea of how input size interacts with this, we look at prefixes of the corpora with increasing sizes (100, 200, 500, 1000, 2000, 5000, and 10,000 utterances). In addition, we are interested in understanding what kind of stress pattern preferences our models acquire. For this, we also collect samples of the probabilities assigned to the different expansions of rule (18), allowing us to examine this directly. The standard evaluation of segmentation models involves having them segment their input in an unsupervised manner and evaluating performance on how well they segmented that input. We additionally evaluate the models on a test set for each corpus. Use of a separate test set has previously been suggested as a means of testing how well the knowledge a learner acquired generalizes to novel utterances (Pearl et al., 2011), and is required for the kind of comparison across different sizes of input we are interested in to determine whether there the role of stress cues interacts with the input size.

We create the test-sets by taking the final 1000 utterances for each corpus. These 1000 utterances will be segmented by the model after it has performed inference on its input, without making any further changes to the lexicon that the model has induced. In other words, the model will have to segment each of the test utterances using only the lexicon (and any additional knowledge about co-occurrences, phonotactics, and stress) it has acquired from the training portion of the corpus during inference.

We measure segmentation performance using the standard metric of token f-score (Brent, 1999) which is the harmonic mean of token precision and recall. Token f-score provides an overall impression of how accurate individual word tokens were identified. To illustrate, if the gold segmentation is “the dog”, the segmentation “th e dog” has a token precision of $\frac{1}{3}$ (one out of three predicted words is correct); a token recall of $\frac{1}{2}$ (one of the two gold words was correctly identified); and a token f-score of 0.4.

### 4.3 Inference

For inference, we closely follow Johnson and Goldwater (2009): we put vague priors on all the hyper-

| p | s | usc | alex |  | korman |  | brent |  |
|---|---|---|---|---|---|---|---|---|
|  |  |  | train | test | train | test | train | test |
|  |  |  | .81 | .81 | .85 | .83 | .82 | .82 |
| • |  |  | .85 | .84 | .86 | .84 | .86 | .86 |
|  | • |  | .86 | .87 | .87 | .86 | .86 | .87 |
| • | • |  | **.88** | **.88** | **.88** | .87 | **.87** | .87 |
|  | • | • | .87 | **.88** | .87 | **.88** | .86 | .87 |
| • | • | • | **.88** | **.88** | **.88** | .87 | **.87** | **.88** |

Table 3: Token f-scores on both train and test portions for all three corpora when inference is performed over the full corpus. Note that the benefit of stress is clearer when evaluating on the test set, and that overall, performance of the different models is comparable across all three corpora. Models are coded according to the key in Table 1.

parameters of our models and run 4 chains for 1000 iterations, collecting 20 samples from each chain with a lag of 10 iterations between each sample after a burn-in of 800 iterations, using both batch-initialization and table-label resampling to ensure good convergence of the sampler. We construct a single segmentation from the posterior samples using their minimum Bayes risk decoding, providing a single score for each condition.

### 4.4 Experimental conditions

Each of our six models is evaluated on inputs of increasing size, starting at 100 and ending at 10,000 utterances, allowing us to investigate both how performance and "knowledge" of the learner varies as a function of input size. For completeness, we also report the "standard" evaluation, i.e. performance of our models on all corpora when trained on the entire input in Table 3. We will focus our discussion on the results obtained on the Alex corpus, which are depicted in Figure 4, where the input size is depicted on the x-axis, and the segmentation f-score for the test-set on the y-axis.

## 5 Discussion

We find a clear improvement for the stress-models over both the colloc3-nophon and the colloc3-phon models. As can be seen in Table 3, the overall trend is the same for all three corpora, both when evaluating on the input and the separate test-set.[5] Note how the relative gain for stress is roughly 1% higher when evaluating on the test-set; this might have to do with Jusczyk (1997)'s observation that the advantage of stress "might be more evident for relatively unexpected or unfamiliarized strings" (Jusczyk, 1997). A closer look at Figure 4 indicates further interesting differences between the colloc3-nophon and the colloc3-phon models that only become evident when considering different input sizes.

### 5.1 Stress cues without phonotactics

For the colloc3-nophon models, we observe a relatively stable improvement by adding stress cues of 6-7%, irrespective of input size and whether or not the Unique Stress Constraint (USC) is assumed. The sole exception to this occurs when the learner only gets to see 100 utterances: in this case, the colloc-nophon-stress model only shows a 3% improvement, whereas the colloc3-nophon-stress-usc model obtains a boost of roughly 8%. Noticeable consistent differences between the colloc3-nophon-stress and colloc3-nophon-stress-usc model, however, all but disappear starting from around 500 utterances. This is somewhat surprising, considering that it is the USC that was argued by Yang (2004) to be key for taking advantage of stress.[6]

We take this behaviour to indicate that even with as little evidence as 200 to 500 utterances, a Bayesian ideal learner can effectively infer that something like the USC is true of English. This also becomes clear when examining how the learners' preferences for different stress patterns evolve over time, as we do in Section 5.3 below.

### 5.2 Stress cues and phonotactics

Overall, the models including phonotactic cues perform better than those that do not rely on phonotactics. However, the overall gain contributed by stress to the colloc3-phon baseline is smaller, al-

[5] We performed Wilcox rank sum tests on the individual scores of the 4 independent chains for each model on the full training data sets and found that the stress-models were always significantly more accurate ($p < 0.05$) than the baseline models except when evaluating on the training data for the Korman and Brent corpora.

[6] On data in which function words are marked for stress (as in Yang (2004) and Doyle and Levy (2013)), the USC yields extremely high scores across all models, simply because roughly every second word is a function word. Given that this assumption is extremely unnatural, we do not take this as an argument for the USC.

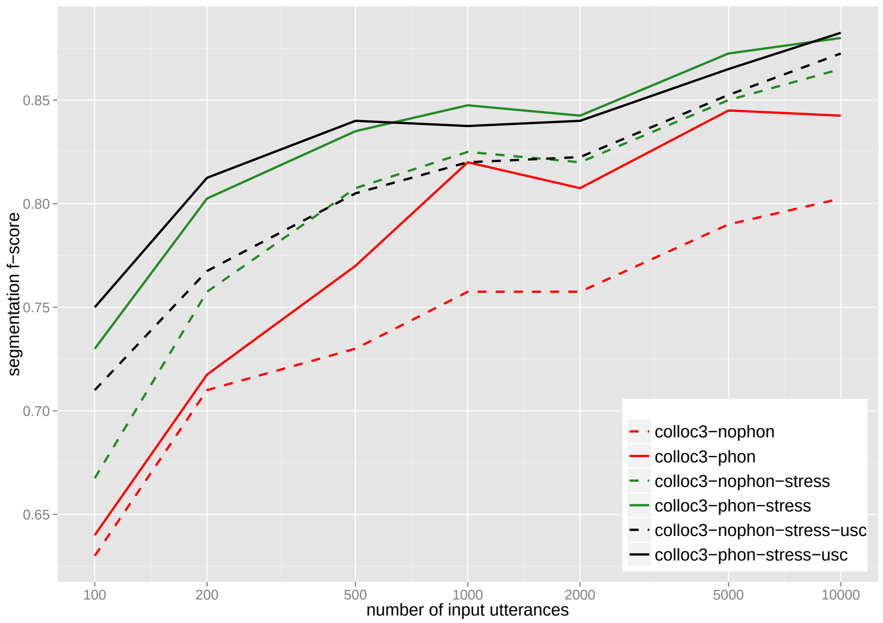

Figure 4: Segmentation performance of the different models, across different input sizes and as evaluated on the test-set for the Alex corpus. The no-stress baselines are given in red, the stress-models without the Unique Stress Constraint (USC) in green and the ones including the USC in black. Solid lines indicate models that use, dashed lines models that do not use phonotactics. Refer to the text for discussion.

though this seems to depend on the size of the input. While phonotactics by itself appears to be a powerful cue, yielding a noticeable 4-5% improvement over the colloc3-nophon baseline, the learner seems to require at least around 500 utterances before the colloc3-phon model becomes clearly more accurate than the colloc3-nophon model. In contrast, even for only 100 utterances stress cues by themselves provide a 3% improvement to the colloc3-nophon model, indicating that they can be taken advantage of earlier. While the number of utterances processed by a Bayesian ideal learner is not directly related to developmental stages, this observation is consistent with the psycholinguists' claim that phonotactics are used by infants for word segmentation after they have begun to use stress for segmentation (Jusczyk et al., 1999a).

Turning to the interaction between stress and phonotactics, we see that there is no consistent advantage of including the USC in the model. This is, in fact, even clearer than for the colloc3-nophon model where at least for small inputs of size 100, the USC added almost 5% in performance. For the colloc3-phon models, we only observe a 1-2% improvement by adding the USC up until 500 utterances. This further strengthens the point that even in the absence of such an innate constraint, a statistical learner can take advantage of stress cues and, as we show below, actually acquire something like the USC from the input.

The 4% difference between the colloc3-phon-stress / colloc3-phon-stress-usc models to the colloc3-phon baseline is smaller than the 7% difference between the colloc3-nophon and colloc3-nophon-stress models. This shows that there is a redundancy between phonotactic and stress cues in large amounts of data, as their joint contribution to the colloc3-nophon baseline is less than the sum of their individual contributions at 10,000 utterances, of 4% (for phonotactics) and 7% (for stress).

Unlike for the colloc3-nophon models, we also see a clear impact of input size. In particular, at 100 utterances the addition of stress cues leads to an 8 – 10% improvement, depending on whether or not the USC is assumed, whereas for the colloc3-nophon model we only observed a 3 – 8% improvement. This is particularly striking when we consider that by themselves, the phonotactic cues only contribute a 1% improvement to the colloc3-nophon baseline when trained on the 100 utterance corpus, indicating a synergistic interaction (rather than redundancy) between phonotactics and stress for small inputs. This effect disappears starting from around 1000 utterances; for inputs of size 1000 and larger, the net-gain of stress drops from roughly 10% to a 3–4% improvement. That is, while we did not notice any relationship between input size and impact of stress cues for the colloc3-nophon model, we do see such an interaction for the combination of phonotactics and stress cues which, taken together, lead to a larger relative gain in performance on smaller inputs than on large ones.

### 5.3 Acquisition of stress patterns

In addition to acquiring a lexicon, the Bayesian learner acquires knowledge about the possible stress patterns of English words. The fact that this knowledge is explicitly represented through the PCFG rules and their probabilities that define the lexical generator allows us to study the generalisations about stress the model actually acquires. While Doyle and Levy (2013) suggest carrying out such an analysis, they restrict themselves to estimating the fraction of stress patterns in the segmented output. As shown in Table 2, however, the type and token distributions of stress patterns can differ substantially. We therefore investigate the stress preferences acquired by our learner by examining the probabilities assigned to the different expansions of rule (18), aggregating the probabilities of the individual rules into patterns. For example, the rules $\underline{\text{Word}} \rightarrow \text{SSyll}\,(\,\text{USyll}\,)^{\{0,3\}}$ correspond to the pattern “Stress on the first syllable”, whereas the rules $\underline{\text{Word}} \rightarrow \text{USyll}^{\{1,4\}}$ correspond to the pattern “Unstressed word”. By computing the respective probabilities, we get the overall probability assigned by a learner to the pattern.

Figure 5 provides this information for several different rule patterns. Additionally, these plots include the empirical type (red dotted) and token proportions (red double-dashed) for the input corpus. Note how for the two major patterns, all models successfully track the type, rather than the token frequency, correctly developing a preference for stress-initial over unstressed words, despite the comparable token frequency of these two patterns. This is compatible with a recent proposal by Thiessen and Saffran (2007), who argue that infants infer the stress pattern over their lexicon. For a Bayesian model such as ours or Goldwater et al. (2009)’s, there is no need to pre-specify that the distribution ought to be learned over types rather than tokens, as the models automatically interpolate between type and token statistics according to the properties of their input (Goldwater et al., 2006). In addition, a Bayesian framework provides a simple answer to the question of how a learner might identify the role of stress in its language without already having acquired at least some words. By combining different kinds of cues, e.g. distributional, phonotactic and prosodic, in a principled manner a Bayesian learner can jointly segment its input and learn the appropriate role of each cue, without having to pre-specify specific preferences that might differ across languages.

The iambic rule pattern that puts stress on the second syllable is much more infrequent on a token level. All models track this low token frequency, underestimating the type frequency of this pattern by a fair amount. This suggests that learning this pattern correctly requires considerably more input than for the other patterns. Indeed, the iambic pattern is known to pose problems for infants when they start using stress as an effective cue. It is only from roughly 10 months of age that infants successfully segment iambic words (Jusczyk et al., 1999b). Not surprisingly, the USC doesn’t aid in learning about this pattern because it is completely silent on where stress might fall (and does not noticeably improve segmentation performance to begin with).

Finally, we can also investigate whether the models that lack the USC nevertheless learn that words contain at most one lexically stressed syllable. The bottom-right graph in Figure 5 plots the probability assigned by the models to patterns that violate the USC. This includes, for example, the rules $\underline{\text{Word}} \rightarrow \text{SyllS SyllS}$ and $\underline{\text{Word}} \rightarrow \text{SyllS SyllU SyllS}$. Note how the probabilities assigned to these rules approaches zero, indicating that the learner becomes more certain that there are no words that contain more than one syllable with lexical stress. As we argued above, this suggests that a Bayesian learner can acquire the USC from a modest amount of data — it will properly infer that the unnatural patterns are simply not supported by the input. To summarize, by examining the internal

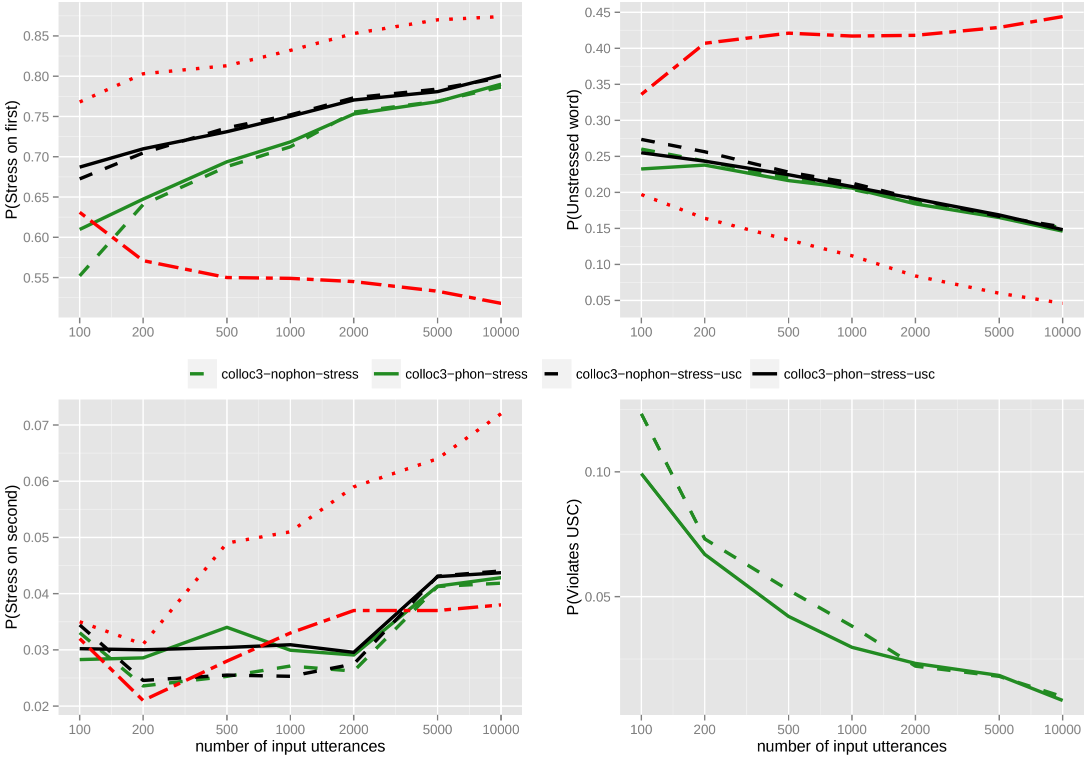

Figure 5: Evolution of the knowledge the learner acquires on the Alex corpus. The red dotted line indicates the empirical type distribution of a specific pattern, and the double-dashed line the empirical token distribution. Top-Left: Stress-initial pattern, Top-Right: Unstressed Words, Bottom-Left: Stress-second pattern, Bottom-Right: Patterns that violate the USC.

state of the Bayesian learners we can characterise how their knowledge about the stress preferences of their languages develops, rather than merely measuring how well they perform word segmentation. We find that the iambic pattern that has been observed to pose problems for infant learners also is harder for the Bayesian learner to acquire, arguably due to its extremely low token-frequency.

## 6 Conclusion and Future Work

We have presented Adaptor Grammar models of word segmentation that are able to take advantage of stress cues and are able to learn from phonemic input. We find that phonotactics and stress interact in interesting ways, and that stress cues makes a stable contribution to existing word segmentation models, improving their performance by 4-6% token f-score. We also find that the USC introduced by Yang (2004) need not be prebuilt into a model but can be acquired by a Bayesian learner from the data. Similarly, we directly investigate the stress preferences acquired by our models and find that for stress-initial and unstressed words, they track type rather than token frequencies. The rare stress-second pattern seems to require more input to be properly acquired, which is compatible with infant development data.

An important goal for future research is to evaluate segmentation models on typologically different languages and to study the relative usefulness of different cues cross-lingually. For example, languages such as French lack lexical stress; it would be interesting to know whether in such a case, phonotactic (or other) cues are more important. Relatedly, recent work such as Börschinger et al. (2013) has found that artificially created data often masks the complexity exhibited by real speech. This suggests that future work should use data directly derived from the acoustic signal to account for contextual effects, rather than using dictionary look-up or other heuristics. In using the Alex corpus, for which good quality audio is available, we have taken a first step in this direction.

## Acknowledgements

This research was supported by the Australian Research Council's Discovery Projects funding scheme (project numbers DP110102506 and DP110102593). We'd like to thank Professor Dupoux and our other colleagues at the Laboratoire de Sciences Cognitives et Psycholinguistique in Paris for hosting us while this research was performed, as well as the Mairie de Paris, the fondation Pierre Gilles de Gennes, the Ecole des Hautes Etudes en Sciences Sociales, the Ecole Normale Supérieure, The Region Ile de France, the European Research Council (ERC-2011-AdG-295810 BOOTPHON), the Agence Nationale pour la Recherche (ANR-2010-BLAN-1901-1 BOOTLANG, ANR-10-IDEX-0001-02 and ANR-10-LABX-0087) and the Fondation de France. We'd also like to thank three anonymous reviewers for helpful comments and suggestions.

## References

N. Bernstein-Ratner. 1987. The phonology of parent-child speech. In K. Nelson and A. van Kleeck, editors, *Children's Language*, volume 6. Erlbaum, Hillsdale, NJ.

Benjamin Börschinger, Katherine Demuth, and Mark Johnson. 2012. Studying the effect of input size for Bayesian word segmentation on the Providence corpus. In *Proceedings of the 24th International Conference on Computational Linguistics (Coling 2012)*, pages 325–340. Coling 2012 Organizing Committee.

Benjamin Börschinger, Mark Johnson, and Katherine Demuth. 2013. A joint model of word segmentation and phonological variation for English word-final /t/-deletion. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1508–1516. Association for Computational Linguistics.

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

M. Christiansen and S. Curtin. 1999. The power of statistical learning: No need for algebraic rules. In *Proceedings of the 21st Annual Conference of the Cognitive Science Society.*

Morten H Christiansen, Joseph Allen, and Mark S Seidenberg. 1998. Learning to segment speech using multiple cues: A connectionist model. *Language and Cognitive Processes*, 13(2-3):221–268.

Suzanne Curtin, Toben H Mintz, and Morten H Christiansen. 2005. Stress changes the representational landscape: Evidence from word segmentation. *Cognition*, 96(3):233–262.

Anne Cutler and David M Carter. 1987. The predominance of strong initial syllables in the English vocabulary. *Computer Speech and Language*, 2(3):133–142.

Anne Cutler, Jacques Mehler, Dennis Norris, and Juan Segui. 1986. The syllable's differing role in the segmentation of French and English. *Journal of Memory and Language*, 25(4):385 – 400.

Anne Cutler. 2005. Lexical stress. In David B. Pisoni and Robert E. Remez, editors, *The Handbook of Speech Perception*, pages 264–289. Blackwell Publishing.

K. Demuth, J. Culbertson, and J. Alter. 2006. Word-minimality, epenthesis, and coda licensing in the acquisition of English. *Language and Speech*, 49:137–174.

Gabriel Doyle and Roger Levy. 2013. Combining multiple information types in Bayesian word segmentation. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 117–126. Association for Computational Linguistics.

Victoria Fromkin, editor. 2001. *Linguistics: An Introduction to Linguistic Theory*. Blackwell, Oxford, UK.

Sharon Goldwater, Tom Griffiths, and Mark Johnson. 2006. Interpolating between types and tokens by estimating power-law generators. In Y. Weiss, B. Schölkopf, and J. Platt, editors, *Advances in Neural Information Processing Systems 18*, pages 459–466. MIT Press.

Sharon Goldwater, Thomas L. Griffiths, and Mark Johnson. 2009. A Bayesian framework for word segmentation: Exploring the effects of context. *Cognition*, 112(1):21–54.

Sharon Goldwater. 2007. *Nonparametric Bayesian Models of Lexical Acquisition*. Ph.D. thesis, Brown University.

Mark Johnson and Katherine Demuth. 2010. Unsupervised phonemic Chinese word segmentation using Adaptor Grammars. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 528–536. Coling 2010 Organizing Committee.

Mark Johnson and Sharon Goldwater. 2009. Improving nonparameteric Bayesian inference: experiments on unsupervised word segmentation with adaptor grammars. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational*

*Linguistics*, pages 317–325. Association for Computational Linguistics.

Mark Johnson, Thomas L. Griffiths, and Sharon Goldwater. 2007. Adaptor Grammars: A framework for specifying compositional nonparametric Bayesian models. In B. Schölkopf, J. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems 19*, pages 641–648. MIT Press, Cambridge, MA.

Mark Johnson. 2008a. Unsupervised word segmentation for Sesotho using Adaptor Grammars. In *Proceedings of the Tenth Meeting of ACL Special Interest Group on Computational Morphology and Phonology*, pages 20–27. Association for Computational Linguistics.

Mark Johnson. 2008b. Using Adaptor Grammars to identify synergies in the unsupervised acquisition of linguistic structure. In *Proceedings of the 46th Annual Meeting of the Association of Computational Linguistics*, pages 398–406. Association for Computational Linguistics.

Peter W Jusczyk, Anne Cutler, and Nancy J Redanz. 1993. Infants' preference for the predominant stress patterns of English words. *Child Development*, 64(3):675–687.

Peter W. Jusczyk, E. A. Hohne, and A. Bauman. 1999a. Infants' sensitivity to allophonic cues for word segmentation. *Perception and Psychophysics*, 61:1465–1476.

Peter W. Jusczyk, Derek M. Houston, and Mary Newsome. 1999b. The beginnings of word segmentation in English-learning infants. *Cognitive Psychology*, 39(3-4):159–207.

Peter Jusczyk. 1997. *The discovery of spoken language*. MIT Press, Cambridge, MA.

Myron Korman. 1984. Adaptive aspects of maternal vocalizations in differing contexts at ten weeks. *First Language*, 5:44–45.

Constantine Lignos and Charles Yang. 2010. Recession segmentation: simpler online word segmentation using limited resources. In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning*, pages 88–97. Association for Computational Linguistics.

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

Constantine Lignos. 2012. Infant word segmentation: An incremental, integrated model. In *Proceedings of the West Coast Conference on Formal Linguistics 30*.

Brian MacWhinney. 2000. The CHILDES project: Tools for analyzing talk: Volume I: Transcription format and programs, volume II: The database. *Computational Linguistics*, 26(4):657–657.

Sven L Mattys and Peter W Jusczyk. 2000. Phonotactic cues for segmentation of fluent speech by infants. *Cognition*, 78(2):91–121.

Sven L Mattys. 2000. The perception of primary and secondary stress in English. *Perception and Psychophysics*, 62(2):253–265.

Lisa Pearl, Sharon Goldwater, and Mark Steyvers. 2011. Online learning mechanisms for Bayesian models of word segmentation. *Research on Language and Computation*, 8(2):107–132.

Elisabeth O. Selkirk. 1984. *Phonology and Syntax: The Relation Between Sound and Structure*. MIT Press.

Erik D Thiessen and Jenny R Saffran. 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.

Erik D Thiessen and Jenny R Saffran. 2007. Learning to learn: Infants acquisition of stress-based strategies for word segmentation. *Language Learning and Development*, 3(1):73–100.

Carnegie Mellon University. 2008. The CMU pronouncing dictionary, v.0.7a.

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