Syllable weight encodes mostly the same information for English word segmentation as dictionary stress

John K Pate **Mark Johnson**
Centre for Language Technology
Macquarie University
Sydney, NSW, Australia
{john.pate,mark.johnson}@mq.edu.au

Abstract

Stress is a useful cue for English word segmentation. A wide range of computational models have found that stress cues enable a 2-10% improvement in segmentation accuracy, depending on the kind of model, by using input that has been annotated with stress using a pronouncing dictionary. However, stress is neither invariably produced nor unambiguously identifiable in real speech. Heavy syllables, i.e. those with long vowels or syllable codas, attract stress in English. We devise Adaptor Grammar word segmentation models that exploit either stress, or syllable weight, or both, and evaluate the utility of syllable weight as a cue to word boundaries. Our results suggest that syllable weight encodes largely the same information for word segmentation in English that annotated dictionary stress does.

1 Introduction

One of the first skills a child must develop in the course of language acquisition is the ability to segment speech into words. Stress has long been recognized as a useful cue for English word segmentation, following the observation that words in English are predominantly stress-initial (Cutler and Carter, 1987), together with the result that 9-month-old English-learning infants prefer stress-initial stimuli (Jusczyk et al., 1993). A range of statistical (Doyle and Levy, 2013; Christiansen et al., 1998; Börschinger and Johnson, 2014) and rule-based (Yang, 2004; Lignos and Yang, 2010) models have used stress information to improve word segmentation. However, that work uses stress-marked input prepared by marking vowels that are listed as stressed in a pronouncing dictionary. This pre-processing step glosses over the fact that stress identification itself involves a non-trival learning problem, since stress has many possible phonetic reflexes and no known invariants (Campbell and Beckman, 1997; Fry, 1955; Fry, 1958). One known strong correlate of stress in English is syllable weight: heavy syllables, which end in a consonant or have a long vowel, attract stress in English. We present experiments with Bayesian Adaptor Grammars (Johnson et al., 2007) that suggest syllable weight encodes largely the same information for word segmentation that dictionary stress information does.

Specifically, we modify the Adaptor Grammar word segmentation model of Börschinger and Johnson (2014) to compare the utility of syllable weight and stress cues for finding word boundaries, both individually and in combination. We describe how a shortcoming of Adaptor Grammars prevents us from comparing stress and weight cues in combination with the full range of phonotactic cues for word segmentation, and design two experiments to work around this limitation. The first experiment uses grammars that provide parallel analyses for syllable weight and stress, and learns initial/non-initial phonotactic distinctions. In this first experiment, syllable weight cues are actually more useful than stress cues at larger input sizes. The second experiment focuses on incorporating phonotactic cues for typical word-final consonant clusters (such as inflectional morphemes), at the expense of parallel structures. In this second experiment, weight cues merely match stress cues at larger input sizes, and the learning curve for the combined weight-and-stress grammar follows almost perfectly with the stress-only grammar. This second experiment suggests that the advantage of weight over stress in the first experiment was purely due to poor modeling of word-final consonant clusters by the stress-only grammar, not weight *per se*. All together, these results indicate that syllable weight

Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 844–853,
October 25-29, 2014, Doha, Qatar.

is highly redundant with dictionary-based stress for the purposes of English word segmentation; in fact, in our experiments, there is no detectable difference between relying on syllable weight and relying on dictionary stress.

2 Background

Stress is the perception that some syllables are more prominent than others, and reflects a complex, language-specific interaction between acoustic cues (such as loudness and duration), and phonological patterns (such as syllable shapes). The details on how stress is assigned, produced, and perceived vary greatly across languages. Three aspects of the English stress system are relevant for this paper. First, although English stress can shift in different contexts (Liberman and Prince, 1977), such as from the first syllable of 'fourteen' in isolation to the second syllable when followed by a stressed syllable, it is largely stable across different tokens of a given word. Second, most words in English end up being stress-initial on a type and token basis. Third, heavy syllables (those with a long vowel or a consonant coda) attract stress in English.

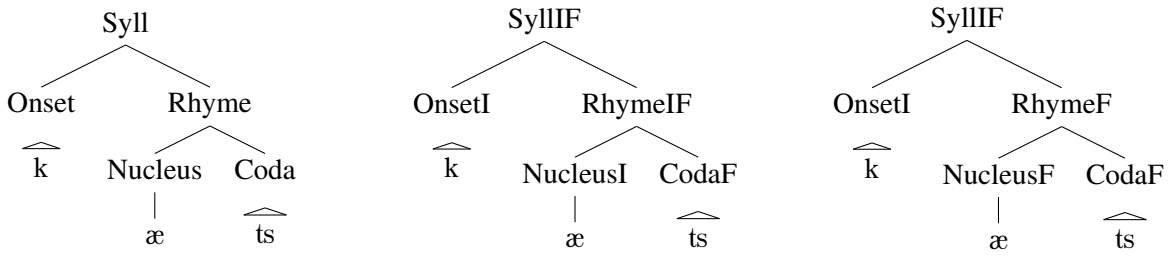
There is experimental evidence that English-learning infants prefer stress-initial words from around the age of seven months (Jusczyk et al., 1993; Juszcyk et al., 1999; Jusczyk et al., 1993; Thiessen and Saffran, 2003). A variety of computational models have subsequently been developed that take stress-annotated input and use this regularity to improve segmentation accuracy. The earliest Simple Recurrent Network (SRN) modeling experiments of Christiansen et al. (1998) and Christiansen and Curtin (1999) found that stress improved word segmentation from about 39% to 43% token f-score (see Evaluation). Rytting et al. (2010) applied the SRN model to probability distributions over phones obtained from a speech recognition system, and found that the entropy of the probability distribution over phones, as a proxy to local hyperarticulation and hence a stress cue, improved token f-score from about 16% to 23%. In a deterministic approach using pre-syllabified input, Yang (2004), with follow-ups in Lignos and Yang (2010) and Lignos (2011; 2012), showed that a 'Unique Stress Constraint' (USC), or assuming each word has at most one stressed syllable, leads to an improvement of about 2.5% boundary f-score.

Among explicitly probabilistic models, Doyle and Levy (2013) incorporated stress into Goldwater et al.'s (2009) Bigram model. They did this by modifying the base distribution over lexical forms to generate not simply phone strings but a sequence of syllables that may or may not be stressed. The resulting model can learn that some sequences of syllables (in particular, sequences that start with a stressed syllable) are more likely than others. However, observed stress improved token f-score by only 1%. Börschinger and Johnson (2014) used Adaptor Grammars (Johnson et al., 2007), a generalization of Goldwater et al.'s (2009) Bigram model that will be described shortly, and found a clearer 4-10% advantage in token f-score, depending on the amount of training data.

Together, the experimental and computational results suggest that infants in fact pay attention to stress, and that stress carries useful information for segmenting words in running speech. However, stress identification is itself a non-trivial task, as stress has many highly variable, context-sensitive, and optional phonetic reflexes. However, one strong phonological cue in English is syllable weight: heavy syllables attract stress. Heavy syllables, in turn, are syllables with a coda and/or a long vowel, which, in English, are tense vowels. Turk et al. (1995) replicated the Jusczyk et al. (1993) finding that English-learning infants prefer stress-initial stimuli (using non-words), and then examined how stress interacted with syllable weight. They found that syllable weight was not a necessary condition to trigger the preference: infants preferred stress-initial stimuli even if the initial syllable was light. However, they also found that infants most strongly preferred stimuli whose first syllable was both stressed and heavy: infants preferred stress-initial and heavy-initial stimuli to stress-initial and light-initial stimuli. This result suggests that infants are sensitive to syllable weight in determining typical stress and rythmic patterns in their language.

2.1 Models

We will adopt the Adaptor Grammar framework used by Börschinger and Johnson (2014) to explore the utility of syllable weight as a cue to word segmentation by way of its covariance with stress. Adaptor Grammars are Probabilistic Context Free Grammars (PCFGs) with a spe-



(a) Basic syllable. (b) Mono-syllable with initial Rhyme. (c) Mono-syllable with final Rhyme.

Figure 1: Different ways to incorporate phonotactics. It is not possible to capture word-final codas and word initial rhymes in monosyllabic words with factors the size of a PCFG rule.

cial set of *adapted* non-terminal nodes. We underline adapted non-terminals (X) to distinguish them from non-adapted non-terminals (Y). While a vanilla PCFG can only directly model regularities that are expressed by a single re-write rule, an Adaptor Grammar model caches entire subtrees that are rooted at adapted non-terminals. Adaptor Grammars can thus learn the internal structure of words, such as syllables, syllable onsets, and syllable rhymes, while still learning entire words as well.

In Adaptor Grammars, parameters are associated with PCFG rules. While this has been a useful factorization in previous work, it makes it difficult to integrate syllable weight and syllable stress in a linguistically natural way. A syllable is typically analyzed as having an optional onset followed by a rhyme, with the rhyme rewriting to a nucleus (the vowel) followed by an optional coda, as in Figure 1a. We expect stress and syllable weight to be useful primarily because initial syllables tend to be different from non-initial syllables. However, distinguishing final from non-final codas should be useful as well, due to the frequency of suffixes in English, and the importance of edge phenomena in phonology more generally (Brent and Cartwright, 1996). These principles come into conflict when modeling monosyllabic words. If we say that a monosyllable is an Initial and Final SyllIF, and has an initial Onset and an initial Rhyme, as in Figure 1b, then we can learn the initial/non-initial generalization about stressed or heavy rhymes at the expense of the generalization about final and non-final codas. If we say that a monosyllable is an initial onset with a final rhyme, the reverse occurs: we can learn the final/non-final coda generalization at the expense of the initial/non-initial regularities. If we split the symbols further, we'd generalize even less: we'd essentially have to learn the initial/non-initial patterns separately for monosyllables and polysyllables.

The most direct solution would introduce factors that are 'smaller' than a single PCFG rule. Essentially, we would compute the score of a PCFG rule in terms of multiple features of its right-hand side, rather than a single 'one-hot' feature identifying the expansion. We left this direction for future work and instead carried out two experiments using Adaptor Grammars that were designed to work around this limitation.

Our first experiment focuses on modeling the initial/non-initial distinction, leaving the final/non-final coda distinction unmodeled. The models in this experiment assume parallel structures for syllable weight and stress, and focus on providing the *most direct* comparison between syllable weight and stress with a strictly initial/non-initial distinction. This first experiment shows that observing dictionary stress is better early in learning, but that modeling syllable weight is better later in learning. However, it is possible that syllable weight was more useful because modeling syllable weight involves modeling the characteristics of codas; the advantage may not have been due to weight *per se* but due to having learned something about the effects of suffixes on final codas.

Our second experiment focuses on modeling some aspects of final codas at the expense of maintaining a rigid parallelism in the structures for syllable weight and stress. The models in this experiment split only those symbols that are necessary to bring stress or weight patterns into the expressive power of the model, and focus on comparing *richer* models of syllable weight and stress that account for inital/internal/final distinctions. This second experiment shows that observing dictionary stress is better early in learning, and that modeling syllable weight merely catches up to

$$\text{Sentence} \rightarrow \underline{\text{Collocations3}}^{+} \quad (1)$$
$$\underline{\text{Collocations3}} \rightarrow \underline{\text{Collocations2}}^{+} \quad (2)$$
$$\underline{\text{Collocations2}} \rightarrow \underline{\text{Collocation}}^{+} \quad (3)$$
$$\underline{\text{Collocation}} \rightarrow \underline{\text{Word}}^{+} \quad (4)$$

Figure 2: Three levels of collocation; symbols followed by $^{+}$ may occur one or more times.

stress without surpassing it. Moreover, a combined stress-and-weight model does no better than a stress model, suggesting that the weight grammar's contribution is fully redundant, for the purposes of word segmentation, with the stress observations.

Together, these experiments suggest that syllable weight eventually encodes everything about word segmentation that dictionary stress does, and that any advantage that syllable weight has over observing dictionary stress is entirely redundant with knowledge of word-final codas.

3 Experiments

3.1 Adaptor Grammars

We follow Börschinger and Johnson (2014) in using a 3-level collocation Adaptor Grammar, as introduced by Johnson and Goldwater (2009) and presented in Figure 2, as the backbone for all models, including the baseline. A 3-level collocation grammar assumes that words are grouped into collocations of words that tend to appear with each other, and that the collocations themselves are grouped into larger collocations, up to three levels of collocations. This collocational structure allows the model to capture strong word-to-word dependencies without having to group frequently-occuring word sequences into a single, incorrect, undersegmented 'word' as the unigram model tends to do (Johnson and Goldwater, 2009)

Word rewrites in different ways in Experiment I and Experiment II, which will be explained in the relevant experiment section.

3.2 Experimental Set-up

We applied the same experimental set-up used by Börschinger and Johnson (2014), to their dataset, as described below. To understand how different modeling assumptions interact with corpus size, we train on prefixes of each corpus with increasing input size: 100, 200, 500, 1,000, 2,000, 5,000, and 10,000 utterances. Inference closely followed Börschinger and Johnson (2014) and Johnson and Goldwater (2009). We set our hyperparameters to encourage onset maximization. The hyperparameter for syllable nodes to rewrite to an onset followed by a rhyme was 10, and the hyperparameter for syllable nodes to rewrite to a rhyme only was 1. Similarly, the hyperparameter for rhyme nodes to include a coda was 1, and the hyperparameter for rhyme nodes to exclude the coda was 10. All other hyperparameters specified vague priors. We ran eight chains of each model for 1,000 iterations, collecting 20 samples with a lag of 10 iterations between samples and a burn-in of 800 iterations. We used the same batch-initialization and table-label resampling to encourage the model to mix.

After gathering the samples, we used them to perform a single minimum Bayes risk decoding of a separate, held-out test set. This test set was constructed by taking the last 1,000 utterances of each corpus. We use a common test-set instead of just evaluating on the training data to ensure that performance figures are comparable across input sizes; when we see learning curves slope upward, we can be confident that the increase is due to learning rather than easier evaluation sets.

We measured our models' performance with the usual token f-score metric (Brent, 1999), the harmonic mean of how many proposed word tokens are correct (token precision) and how many of the actual word tokens are recovered (token recall). For example, a model may propose "the in side" when the true segmentation is "the inside." This segmentation would have a token precision of $\frac{1}{3}$, since one of three predicted words matches the true word token (even though the other predicted words are valid word types), and a token recall of $\frac{1}{2}$, since it correctly recovered one of two words, yield a token f-score of 0.4.

3.3 Dataset

We evaluated on a dataset drawn from the Alex portion of the Providence corpus (Demuth et al., 2006). This dataset contains $17,948$ utterances with $72,859$ word tokens directed to one child from the age of 16 months to 41 months. We used a version of this dataset that contained annotations of primary stress that Börschinger and Johnson (2014) added to this input using an extended

RhymeI → HeavyRhyme
RhymeI → LightRhyme
Rhyme → HeavyRhyme
Rhyme → LightRhyme
HeavyRhyme → LongVowel
HeavyRhyme → Vowel Coda
LightRhyme → ShortVowel

(a) Weight-sensitive grammar

RhymeI → RhymeS
RhymeI → RhymeU
Rhyme → RhymeS
Rhyme → RhymeU
RhymeS → Vowel Stress (Coda)
RhymeU → Vowel (Coda)

(b) Stress-sensitive grammar

RhymeI → Vowel (Coda)
Rhyme → Vowel (Coda)

(c) Baseline grammar

RhymeI → HeavyRhymeS
RhymeI → HeavyRhymeU
RhymeI → LightRhymeS
RhymeI → LightRhymeU
Rhyme → HeavyRhymeS
Rhyme → HeavyRhymeU
Rhyme → LightRhymeS
Rhyme → LightRhymeU
HeavyRhymeS → LongVowel Stress
HeavyRhymeS → LongVowel Stress Coda
HeavyRhymeU → LongVowel
HeavyRhymeU → LongVowel Coda
LightRhymeS → ShortVowel Stress
LightRhymeU → ShortVowel

(d) Combined grammar

Figure 3: Experiment I Grammars

version of CMUDict (cmu, 2008).[1] The mean number of syllables per word token was 1.2, and only three word tokens had more than five syllables. Of the $40,323$ word tokens with a stressed syllable, $27,258$ were monosyllabic. Of the $13,065$ polysyllabic word tokens with a stressed syllable, $9,931$ were stress-initial. Turning to the $32,536$ word tokens with no stress (i.e., the function words), all but 23 were monosyllabic (the 23 were primarily contractions, such as "couldn't").

3.4 Experiment I: Parallel Structures

The goal of this first experiment is to provide the most direct comparison possible between grammars that attend to stress cues and grammars that attend to syllable weight cues. As these are both hypothesized to be useful by way of an initial/non-initial distinction, we defined a word to be an initial syllable SyllI followed by zero to three syllables, and syllables to consist of an optional onset and a rhyme:

$$\underline{\text{Word}} \rightarrow \text{SyllI} \quad (\text{Syll})^{\{0,3\}} \tag{5}$$

$$\text{SyllI} \rightarrow (\underline{\text{OnsetI}}) \quad \text{RhymeI} \tag{6}$$

$$\text{Syll} \rightarrow (\underline{\text{Onset}}) \quad \text{Rhyme} \tag{7}$$

In the baseline grammar, presented in Figure 3c, rhymes rewrite to a vowel followed by an optional consonant coda. Rhymes then rewrite to be heavy or light in the weight grammar, as in Figure 3a, to be stressed or unstressed in the stress grammar, as in Figure 3b. In the combination grammar, rhymes rewrite to be heavy or light and stressed or unstressed, as in Figure 3d. LongVowel and ShortVowel both re-write to all vowels. An additional grammar that restricted them to rewrite to long and short vowels, respectively, led to virtually identical performance, suggesting that vowel quantity can be learned for the purposes of word segmentation from distributional cues. We will also present evidence that the model did manage to learn most of the contrast.

Figure 4 presents learning curves for the grammars in this parallel structured comparison. We see that observing stress without modeling weight

[1] This dataset and these Adaptor Grammar models are available at: `http://web.science.mq.edu.au/~jpate/stress/`

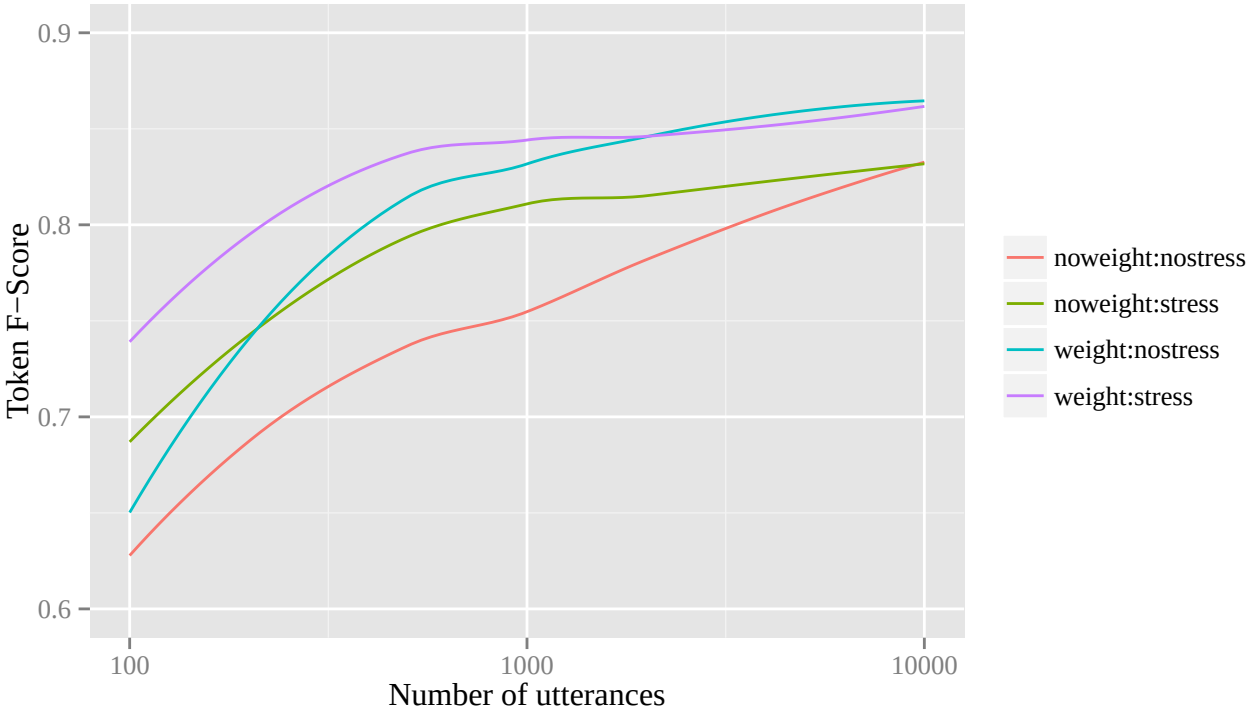


Figure 4: Learning curves on the Alex corpus for Experiment I grammars with parallel distinctions between Stressed/Unstressed and Heavy/Light syllable rhymes.

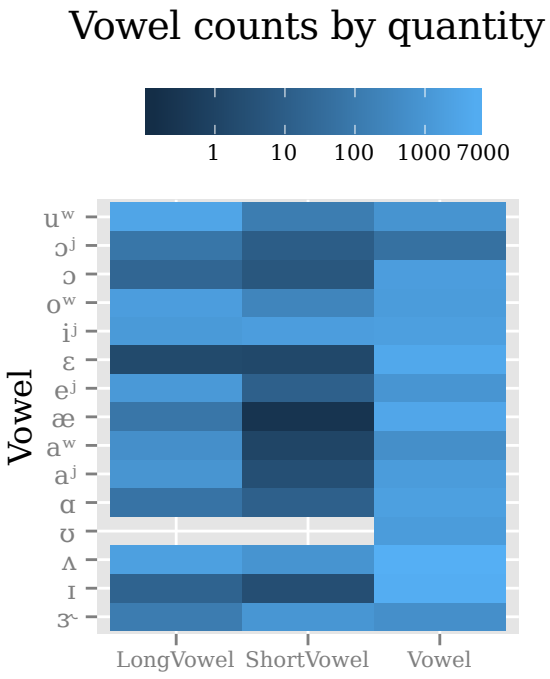


Figure 5: Heatmap of learned vowels in the Experiment I weight-only grammar. Each cell corresponds to the count of a particular vowel being analyzed as one of the three vowel types. Diphthongs are rarely ShortVowel.

outperforms both the baseline and the weight-only grammar early in learning. The weight-only grammar rapidly improves in performance at larger training data sizes, increasing its advantage over the baseline, while the advantage of the stress-only grammar slows and appears to disappear at the largest training data size. At 10,000 utterances, the improvement of the weight-only grammar over the stress-only grammar is significant according to an independent samples t-test ($t = 7.2, p < 0.001$, 14 degrees of freedom). This pattern suggests that annotated dictionary stress is easy to take advantage of at low data sizes, but that, with sufficient data, syllable weight can provide even more information about word boundaries. The best overall performance early in learning is obtained by the combined grammar, suggesting that syllable weight and dictionary stress provide information about word segmentation that is not redundant.

An examination of the final segmentation suggests that the weight grammar has learned that initial syllables tend to be heavy. Specifically, across eight runs, 98.1% of RhymeI symbols rewrote to HeavyRhyme, whereas only 54.5% of Rhyme symbols (i.e. non-initial rhymes) rewrote to HeavyRhyme.

Model	Mean TF	Std. Dev.
noweight:nostress	0.830	0.005
noweight:stress	0.831	0.008
weight:nostress	0.861	0.008
weight:stress	0.861	0.008

Table 1: Segmentation Token F-score for Experiment I at 10,000 utterances across eight runs.

We also examined the final segmentation to see well the model learned the distinction between long vowels and short vowels. Figure 5 presents a heatmap, with colors on a log-scale, showing how many times each vowel label rewrote to each possible vowel in the (translated to IPA). Although the quantity generalisations are not perfect, we do see a general trend where ShortVowel rarely rewrites to diphthongs.

3.5 Experiment II: Word-final Codas

Experiment I suggested that, under a basic initial/non-initial distinction, syllable weight eventually encodes more information about word boundaries than does dictionary stress. This is a surprising result, since we initially investigated syllable weight as a noisy proxy for dictionary stress. One possible source of the 'extra' advantage that the syllable weight grammar exhibited has to do with the importance of word-final codas, which can encode word-final morphemes in English (Brent and Cartwright, 1996). Even though the grammars did not explicitly model them, the weight grammar could implicitly capture a bias for or against having a coda in non-initial position, while the stress grammar could not. This is because most word tokens are one or two syllables, and only one of the two rhyme types of the weight grammar included a coda. Thus, the HeavyRhyme symbol could simultaneously capture the most important aspects of both stress and coda constraints.

To see if the extra advantage of the syllable weight grammar can be attributed to the influence of word-final codas, we formulated a set of grammars that model word-final codas and also can learn stress and/or syllable weight patterns. These grammars are more similar in structure to the ones that Börschinger and Johnson (2014) used. For the baseline and weight grammar, we again defined words to consist of up to four syllables with an initial SyllI syllable, but this time distinguished final syllables SyllF in polysyllabic words. The non-stress grammars use the following rules for producing syllables:

$$\underline{\text{Word}} \rightarrow \text{SyllIF} \tag{8}$$

$$\underline{\text{Word}} \rightarrow \text{SyllI} \quad (\text{Syll})^{\{0,2\}} \quad \text{SyllF} \tag{9}$$

$$\text{SyllIF} \rightarrow (\text{OnsetI}) \quad \text{RhymeI} \tag{10}$$

$$\text{SyllI} \rightarrow (\text{OnsetI}) \quad \text{RhymeI} \tag{11}$$

$$\text{Syll} \rightarrow (\text{Onset}) \quad \text{Rhyme} \tag{12}$$

$$\text{SyllF} \rightarrow (\text{Onset}) \quad \text{RhymeF} \tag{13}$$

For the stress grammar, we followed Börschinger and Johnson (2014) in distinguishing stressed and unstressed syllables, rather than simply stressed rhymes as in Experiment I, to allow the model to learn likely stress patterns at the word level. A word can consist of up to four syllables, and any syllable and any number of syllables may be stressed, as in Figure 6a.

The baseline grammar is similar to the previous one, except it distinguishes word-final codas, as in Figure 6b. The weight grammar, presented in Figure 6c, rewrites rhymes to a nucleus followed by an optional coda and distinguishes nuclei in open syllables according to their position in the word. The stress grammar, presented in Figure 6d, is the all-stress-patterns model (without the unique stress constraint) Börschinger and Johnson (2014). This grammar introduces additional distinctions at the syllable level to learn likely stress patterns, and distinguishes final from non-final codas. The combined model is identical to the stress model, except Vowel non-terminals in closed and word-internal syllables are replaced with Nucleus non-terminals, and Vowel non-terminals in word-inital (-final) open syllables are replaced with NucleusI (NucleusF) non-terminals.

To summarize, the stress models distinguish stressed and unstressed syllables in initial, final, and internal position. The weight models distinguish the vowels of initial open syllables, the vowels of final open syllables, and other vowels, allowing them to take advantage of an important cue from syllable weight for word segmentation: if an initial vowel is open, it should usually be long.

Figure 7 shows segmentation performance on the Alex corpus with these more complete models. While the performance of the weight grammars is virtually unchanged compared to Figure 4, the two grammars that do not model syllable weight improve dramatically. This result supports our proposal that much of the advantage of the weight

$$\underline{\text{Word}} \rightarrow \{\text{SyllUIF}|\text{SyllSIF}\}$$
$$\underline{\text{Word}} \rightarrow \{\text{SyllUI}|\text{SyllSI}\} \quad \{\text{SyllU}|\text{SyllS}\}^{\{0,2\}} \quad \{\text{SyllUF}|\text{SyllSF}\}$$

(a) The all-patterns stress model

$$\text{Rhyme} \rightarrow \text{Vowel} \quad (\underline{\text{Coda}})$$
$$\text{RhymeF} \rightarrow \text{Vowel} \quad (\underline{\text{CodaF}})$$

(b) Baseline grammar

$$\text{RhymeI} \rightarrow \underline{\text{NucleusI}}$$
$$\text{RhymeI} \rightarrow \underline{\text{Nucleus}} \quad \underline{\text{Coda}}$$
$$\text{Rhyme} \rightarrow \underline{\text{Nucleus}} \quad (\underline{\text{Coda}})$$
$$\text{RhymeF} \rightarrow \underline{\text{NucleusF}}$$
$$\text{RhymeF} \rightarrow \underline{\text{Nucleus}} \quad \underline{\text{CodaF}}$$

(c) Weight-sensitive grammar

$$\text{SyllSIF} \rightarrow \underline{\text{OnsetI}} \quad \text{RhymeSF}$$
$$\text{SyllUIF} \rightarrow \underline{\text{OnsetI}} \quad \text{RhymeUF}$$
$$\text{SyllSI} \rightarrow \underline{\text{Onset}} \quad \text{RhymeS}$$
$$\text{SyllUI} \rightarrow \underline{\text{Onset}} \quad \text{RhymeU}$$
$$\text{SyllSF} \rightarrow \underline{\text{Onset}} \quad \text{RhymeSF}$$
$$\text{SyllUF} \rightarrow \underline{\text{Onset}} \quad \text{RhymeUF}$$
$$\text{RhymeSI} \rightarrow \text{Vowel} \quad \text{Stress} \quad (\underline{\text{Coda}})$$
$$\text{RhymeUI} \rightarrow \text{Vowel} \quad (\underline{\text{Coda}})$$
$$\text{RhymeS} \rightarrow \text{Vowel} \quad \text{Stress} \quad (\underline{\text{Coda}})$$
$$\text{RhymeU} \rightarrow \text{Vowel} \quad (\underline{\text{Coda}})$$
$$\text{RhymeSF} \rightarrow \text{Vowel} \quad \text{Stress} \quad (\underline{\text{CodaF}})$$
$$\text{RhymeUF} \rightarrow \text{Vowel} \quad (\underline{\text{CodaF}})$$

(d) Stress-sensitive grammar

Figure 6: Experiment II Grammars.

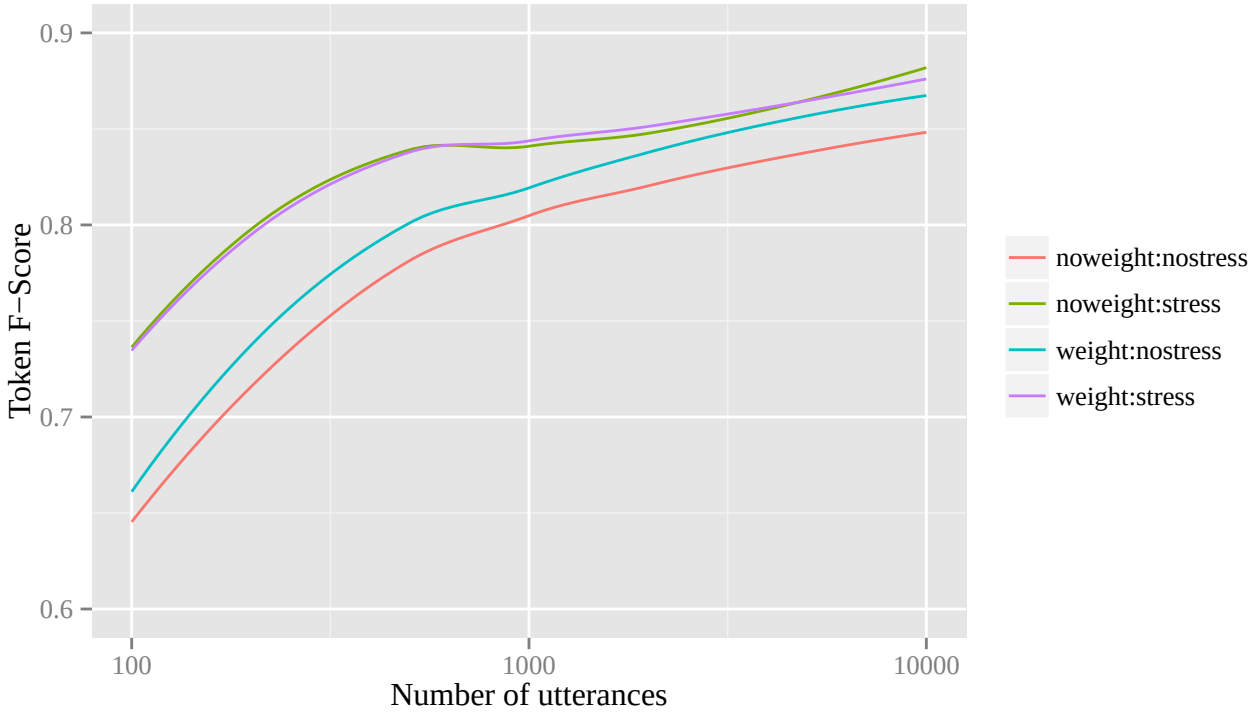


Figure 7: Learning curves on the Alex corpus for Experiment II grammars with word-final phonotactics that exploit Stress and Weight.

Model	Mean TF	Std. Dev.
noweight:nostress	0.846	0.007
noweight:stress	0.880	0.005
weight:nostress	0.865	0.011
weight:stress	0.875	0.005

Table 2: Segmentation Token F-score for Experiment II at 10,000 utterances across eight runs.

grammars over stress in Experiment I was due to modeling of word-final coda phonotactics.

Table 2 presents token f-score at 10,000 training utterances averaged across eight runs, along with the standard deviation in f-score. We see that the noweight:nostress grammar is several standard deviations than the grammars that model syllable weight and/or stress, while the syllable weight and/or stress grammars exhibit a high degree of overlap.

4 Conclusion

We have presented computational modeling experiments that suggest that syllable weight (eventually) encodes nearly everything about word segmentation that dictionary stress does. Indeed, our experiments did not find a persistent advantage to observing stress over modeling syllable weight. While it is possible that a different modeling approach might find such a persistent advantage, this advantage could not provide more than 13% absolute F-score. This result suggests that children may be able to learn and exploit important rhythm cues to word boundaries purely on the basis of segmental input. However, this result also suggests that annotating input with dictionary stress has missed important aspects of the role of stress in word segmentation. As mentioned, Turk et al. (1995) found that infants preferred initial light syllables to be stressed. Such a preference obviously cannot be learned by attending to syllable weight alone, so infants who have learned weight distinctions must also be sensitive to non-segmental acoustic correlates to stress. There was no long-term advantage to observing stress in addition to attending to syllable weight in our models, however, suggesting that annotated dictionary stress does not capture the relevant non-segmental phonetic detail. More modeling is necessary to assess the non-segmental phonetic features that distinguish stressed light syllables from unstressed light syllables.

This investigation also highlighted a weakness of current Adaptor Grammar models: the 'smallest' factors are the size of one PCFG rule. Allowing further factorizations, perhaps using feature functions of a rule's right-hand side, would allow models to capture finer-grained distinctions without fully splitting the symbols that are involved.

References

Benjamin Börschinger and Mark Johnson. 2014. Exploring the role of stress in Bayesian word segmentation using Adaptor Grammars. *Transactions of the ACL*, 2:93–104.

Michael R Brent and Timothy A Cartwright. 1996. Distributional regularity and phonotactic constraints are useful for segmentation. *Cognition*, 61:93–125.

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

Nick Campbell and Mary Beckman. 1997. Stress, prominence, and spectral tilt. In *Proceedings of an ESCA workshop*, pages 67–70, Athens, Greece.

Morten H. Christiansen and Suzanne L 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*, 13:221—268.

2008. The CMU pronouncing dictionary. `http://www.speech.cs.cmu.edu/cgi-bin/cmudict`.

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

Katherine Demuth, Jennifer Culbertson, and Jennifer 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 NAACL 2013*, pages 117–126. Association for Computational Linguistics.

D B Fry. 1955. Duration and intensity as physical correlates of linguistic stress. *J. Acoust. Soc. of Am.*, 27:765–768.

D B Fry. 1958. Experiments in the perception of stress. *Language and Speech*, 1:126–152.

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.

Mark Johnson and Sharon Goldwater. 2009. Improving nonparametric 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 Schoelkopf, J Platt, and T Hoffmann, editors, *Advances in Neural Information Processing Systems*, volume 19. The MIT Press.

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

Mark Liberman and Alan Prince. 1977. On stress and linguistic rhythm. *Linguistic Inquiry*, 8(2):249–336, Spring.

Constantine Lignos and Charles Yang. 2010. Recession segmentation: simpler online word segmentation using limited resources. In *Proceedings of ACL 2010*, 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*.

C Anton Rytting, Chris Brew, and Eric Fosler-Lussier. 2010. Segmenting words from natural speech: subsegmental variation in segmental cues. *Journal of Child Language*, 37(3):513–543.

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):706–716.

Alice Turk, Peter W Jusczyk, and Louann Gerken. 1995. Do English-learning infants use syllable weight to determine stress? *Language and Speech*, 38(2):143–158.

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

EMNLP 2014

The 2014 Conference on Empirical Methods In Natural Language Processing

Proceedings of the Conference

October 25-29, 2014
Doha, Qatar

Sponsors

Diamond

Platinum

Gold

Silver

Bronze

Supporter

IBM **Research**

Sponsor of Student Volunteers

Order copies of this and other ACL proceedings from:

Association for Computational Linguistics (ACL)
209 N. Eighth Street
Stroudsburg, PA 18360
USA
Tel: +1-570-476-8006
Fax: +1-570-476-0860
acl@aclweb.org

ISBN 978-1-937284-96-1

Preface by the General Chair

Welcome to the 2014 Conference on Empirical Methods in Natural Language Processing.

The EMNLP conference series is annually organized by SIGDAT, the Association for Computational Linguistics' special interest group on linguistic data and corpus-based approaches to NLP. This year the conference is being held from October 25, 2014 (Sat.) to October 29, 2014 (Wed.) in Doha, Qatar.

In the past five years, the EMNLP conference attendance has been continuously growing, reaching just over 500 paying attendees in 2013, and it is nowadays considered as one of the leading conferences in Computational Linguistics and Natural Language Processing.

Given the growing trend, we believed it was the right time to lead EMNLP into an organization structure typical of large and important conferences. Therefore, we proposed several novelties: first of all, a large organization committee consisting of twenty (plus twenty-six area chairs) well-known members of the ACL community, who carried out several tasks required by the new achieved scale.

Secondly, as this is the first conference edition spanning five days, in addition to six workshops, we also selected and included for the first time an excellent selection of eight tutorials. We defined a registration policy that allows the participants to attend any of the tutorials and workshops (held on October 25th and 29th) by just paying a low flat rate on top of the registration fee for the main conference. We believe this can greatly increase the spread of advanced technology and promote a unified view of the techniques and foundations of our research field.

Thirdly, as a standalone conference, EMNLP required the definition of new administrative procedures and policies, regarding sponsorship booklets, double submission, scholarship assignment, and the joint EACL-ACL-EMNLP call for workshop proposals.

Next, EMNLP is finding new ways to foster the dissemination of research work by facing the increasing number of papers to be presented at the conference. Our new approach consisted in presenting posters in nine sessions each proposing a small numbers of papers: this way poster presentations can receive the space and consideration that they deserve. Then, we are adding a surprise in terms of paper presentation and dissemination, which will be unveiled only few days before the start of the conference.

Finally, this is the first time that an ACL conference is largely supported by a government research foundation. The Qatar National Research Foundation (QNRF) has included EMNLP 2014 as one of its local funding events. This enabled EMNLP and SIGDAT to perform unprecedented student scholarship support: more than 30 students were sponsored (partially or entirely) for participating in the conference. The obtained funds also allowed for offering a social dinner free of charge to all the attendees and still closing the conference budget in active, thus creating additional resources that SIGDAT can use to support the upcoming conferences.

The novelties above as well as the traditional activities that the EMNLP conference series proposes to its members could not have been organized without the work of our large committee. In this respect, I would like to thank our PC co-chairs Walter Daelemans and Bo Pang, who greatly used their large experience with program committees of our community for selecting an excellent program.

Special thanks go to our publication chair Yuval Marton, who did a terrific job in organizing and preparing the proceedings. As a side effect of his proactive action, workshop organizers and future publication chairs using the SoftConf START/ACLPUB systems can now streamline the inclusion of workshops and conference schedules in the proceedings, without heavy manual customization.

We are very grateful to Enrique Alfonseca and Eric Gaussier for selecting interesting and successful

workshops and to Lucia Specia and Xavier Carreras, who, for the first time, carried out the new task of selecting tutorials for an EMNLP conference. The workshops and tutorials nicely filled the additional two days of EMNLP, making our conference even more valuable.

Many thanks are due to Katrin Erk and Sebastian Padó, who were challenged by the new activity (for EMNLP) of defining policy for the selection and assignment of participation scholarships to the most deserving students. The uncertainty over the final amount of funds and their diverse nature made this task particularly difficult. Nevertheless, they were able to find appropriate and successful solutions.

As any large conference, we could count on the help of publicity co-chairs to advertise the old and new EMNLP features. We give our gratitude to Mona Diab and Irina Matveeva for their professional work.

Fund hunting is a very important activity for conferences, in this respect, I would like to thank our sponsorship co-chairs, Jochen Leidner, Veselin Stoyanov and Min Zhang, for helping us to look for sponsors in three different continents.

Regarding the SIGDAT side, a special thank is devoted to Noah Smith, who promptly answered any question I came out with. I am also grateful to the other SIGDAT officers (past and new): Eugene Charniak, Mark Johnson, Philipp Koehn, Mark Steedman, who were always there to give suggestions and solutions to critical issues that inevitably arise in any large event.

Many thanks also to Tim Baldwin, Anna Korhonen, Graeme Hirst and David Yarowsky who provided much useful information from past conferences. Last but not least, I would like to thank Priscilla Rasmussen for her help and advice, and her undoubtful qualities of soothsayer regarding the estimation of conference numbers.

Coming back to the sponsor topic, we are enormously thankful to QNRF, for accepting our proposal to fund EMNLP: this has made it possible to sponsor an unprecedented number of students and offer a banquet free of charge to all participants (we needed to create a new level of sponsorship for them, namely, Diamond). We are very grateful to The Qatar Computing Research Institute, which in addition to providing the very valuable Platinum sponsorship, also provided the required man power for organizing the event.

In particular, EMNLP could not be organized in Qatar without the work of Kareem Darwish, the local organization chair. We are also very grateful to Kemal Oflazer, local co-chair and Francisco Guzman Herrera, local sponsorship chair, whose work was determinant to obtain the QNRF sponsorship. We are deeply in debt with the other local organizers, Lluís Màrquez, who also edited the conference booklet, Preslav Nakov, Fabrizio Sebastiani and Stephan Vogel for their help with the daily big and little issues.

Special thanks go to The Carnegie Mellon University in Qatar for helping us with the proposal preparation and management of the QNRF funds and also for supporting us with a Gold sponsorship. Additionally, many thanks go to our silver sponsors, Facebook and Yandex and our bronze sponsor iHorizons, who show the increasing interest of industry in the technology of our community for the design of real-world and high-societal impact applications. In this respect, we sincerely thank Google Inc. and IBM Watson, New York, for supporting the student participation with their scholarships.

Finally, and foremost, thanks to all the authors and conference attendees who are the main actors of this event, bringing the real value to it and determining its success. My personal thanks also go to the entire SIGDAT committee, for choosing me as the chair of this fantastic conference, held in a fascinating venue.

Alessandro Moschitti

General Chair of EMNLP 2014

Preface by the Program Committee Co-Chairs

We welcome you to the 2014 Conference on Empirical Methods in Natural Language Processing.

As in the previous EMNLP, we invited both long and short papers with a single submission deadline. Short papers encourage the submission of smaller and more preliminary contributions.

We received 790 submissions (after initial withdrawals of unfinished submissions and removal of duplicates), of which 28 were rejected before review for not adhering to the instructions in the call for papers regarding paper length or anonymity. The remaining 510 long and 252 short papers were allocated to one of the fourteen areas. The most popular areas this year were Machine Translation, Semantics, and Syntax (Tagging, Chunking, and Parsing).

Reviewing for a conference this size involves an army of dedicated professionals volunteering to donate their valuable and scarce time to make sure that the highest possible reviewing standards are reached. We are very grateful to our 26 area chairs and a programme committee of more than 500 for their efforts. We accepted 155 long and 70 short papers, representing a global acceptance rate of just under 30%. Nine papers accepted by the ACL journal TACL were added to the program.

Based on the reviews and on nominations by the area chairs, 5 long papers were shortlisted for the best paper award. The best paper will be presented in a plenary best paper award ceremony. We would like to thank Mark Johnson and Claire Cardie for their willingness to serve in the best paper award committee that was set up and for providing excellent advice and motivation for their choice.

We are grateful to the authors for selecting EMNLP as the venue for their work. Congratulations to the authors of accepted submissions. To the authors of rejected submissions we would like to offer as consolation the fact that because of the competitive nature of the conference and the inevitable time and space limitations, many worthwhile papers could not be included in the program. We hope the feedback of the reviewers will be considered worthwhile by them and lead to successful future submissions.

We are very grateful to our invited speakers Thorsten Joachims and Salim Roukos. Thorsten Joachims is professor at the Computer Science and Information Science departments at Cornell University and shows how integrating microeconomic models of human behavior into the learning process leads to new interaction models and learning algorithms, in turn leading to better performing systems. Salim Roukos is senior manager of multilingual NLP and CTO of Translation Technologies at IBM T.J. Watson research Center and addresses IBM's approach to cognitive computing for building systems and solutions that enable and support richer human-machine interactions, and remaining opportunities in this area for novel statistical models for natural language processing. We thank them for their inspiring talks and presence at the conference.

We would also like to thank our general chair Alessandro Moschitti for his leadership, advice, encouragement, and support, Kareem Darwish and his colleagues for impeccable cooperation from local organization, and Yuval Marton for doing an excellent job assembling these proceedings.

It was an honour to serve as Programme Chairs of EMNLP 2014, and we hope that you will enjoy the conference and be able to think back later and remember a scientifically stimulating conference and a pleasant time in Doha, Qatar.

Bo Pang and Walter Daelemans

EMNLP 2014 Program Chairs