Exploring the role of stress in Bayesian word segmentation using Adaptor Grammars

Benjamin Börschinger
Mark Johnson

Macquarie University
Sydney, Australia

June 2014

Talk summary

- High-level goal: use computational models to study human language acquisition
- Most computational models focus on an extremely idealised version of language acquisition problem
  - much previous work treats input as sequences of segments
  - ignores cues that psycholinguists think are important in human language acquisition
- We use Adaptor Grammars to study role of *stress in word learning*, including:
  - the *interaction of stress with phonotactic constraints*
  - how the *contribution of stress varies with size of input*
  - *learning a preference for word-initial stress in English*

MACQUARIE UNIVERSITY

Outline

Stress and word segmentation

Computational models of word segmentation

Experiments

Conclusions and future work

MACQUARIE UNIVERSITY

Word segmentation and language acquisition

- Speech is not cleanly segmented into words
  - children have to learn how to segment utterances into words
- Elman (1996) and Brent (1999) studied a simplified *word segmentation* problem where the data is prepared by:
  - looking up each word in a child-directed speech transcript in a pronouncing dictionary
  - concatenating the most frequent pronunciations to get an utterance pronunciation

j △ u ▲ w △ ɑ △ n △ t ▲ t △ u ▲ s △ i ▲ ð △ ə ▲ b △ ɔ △ k

ju wɑnt tu si ðə bɔk

"you want to see the book"

- Model's goal: determine location of word boundaries

⇒ identifies the pronunciations of words in the transcript

MACQUARIE UNIVERSITY

Stress in English and other languages

- Stress is the "accentuation of syllables within words"
  - phonetic correlates vary within and across languages
- Stress placement in English must be learned:
  - 2-syllable words with initial stress: *GIant, PICture, HEAting*
  - 2-syllable words with final stress: *toDAY, aHEAD, aLLOW*
- In other languages stress depends on syntax (e.g., French)
- English has a *strong preference for initial-syllable stress* (Cutler 1987)
  - roughly 50% of tokens and 85% of types are initial stress
  - but: roughly 50% of tokens and 5% of types are unstressed
- Psycholinguistic work shows English-speaking children use stress in word segmentation

MACQUARIE UNIVERSITY

Adding stress to word-segmentation data

- We *annotate stress on the vowel nucleii of stressed syllables*
  - Johnson and Demuth (2010) annotated tone in Chinese in same way

j △ u ▲ w △ ɑ* △ n △ t ▲ t △ u ▲ s △ i* ▲ ð △ ə ▲ b △ ɔ* △ k

- We marked-up three corpora with dictionary stress
  - we *treat function words as unstressed*
  - results for Alex portion of the Providence corpus
    results on other corpora are very similiar

MACQUARIE UNIVERSITY

Outline

Stress and word segmentation

Computational models of word segmentation

Experiments

Conclusions and future work

MACQUARIE UNIVERSITY

Computational models that exploit stress

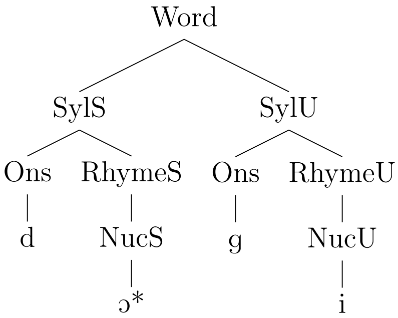
- Yang (2004), Lignos and Yang (2010), Lignos (2011)
  - non-statistical models
  - hard-coded Unique Stress Constraint (at most one stressed syllable per word)
  - pre-syllabified input
  - high segmentation accuracy
- Doyle and Levy (2013)
  - extension of Goldwater's Bigram model
  - pre-syllabified input
  - small but significant improvement by adding stress
- Motivation for this work: how much impact does stress have in Bayesian word segmentation?

Useful cues for word segmentation

- *Vocabulary of the language*
  - no obvious upper bound $\Rightarrow$ *non-parametric* learning
- *Exhaustive parsing* (no unparsed speech)
- *Phonotactics* (e.g., syllable structure constraints)
- *Distributional cues* (e.g., collocations)
- Semantic constraints (e.g., word-topic mappings)
- Social cues (e.g., care-giver's eye-gaze)
- Morpho-syntax, e.g., function words
  (see Johnson et al, this conference)
- *Prosodic cues*, specifically: *stress* (this paper)

MACQUARIE UNIVERSITY

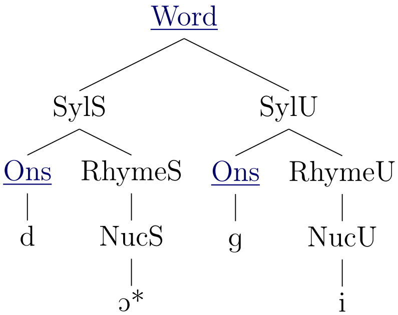
Weaknesses of PCFGs for word segmentation



- PCFG rules can capture stress patterns within words
  - $\mathrm{P}(\mathrm{Word} \rightarrow \mathrm{SylS}\ \mathrm{SylU})$ is probability of 2-syllable words with stressed-unstressed stress pattern
- But this PCFG *can't learn* that /dɔ*gi/ is a word

MACQUARIE UNIVERSITY

Adaptor grammars memoise entire subtrees



- *Adaptor grammars* learn probability of *adapted nonterminals* expanding to *entire subtrees* (as well as rule probabilities)
  - adapted nonterminals depicted as underlined and highlighted
  - e.g. probability of $\underline{\mathrm{Word}} \Rightarrow^{+} dOgi$ and $\underline{\mathrm{Word}} \rightarrow \mathrm{SylS}\ \mathrm{SylU}$
  - each adapted nonterminal is associated with a Pitman-Yor Process (PYP)
    - PCFG rules specify *base distributions*

  $\Rightarrow$ defines a *hierarchy of PYPs*

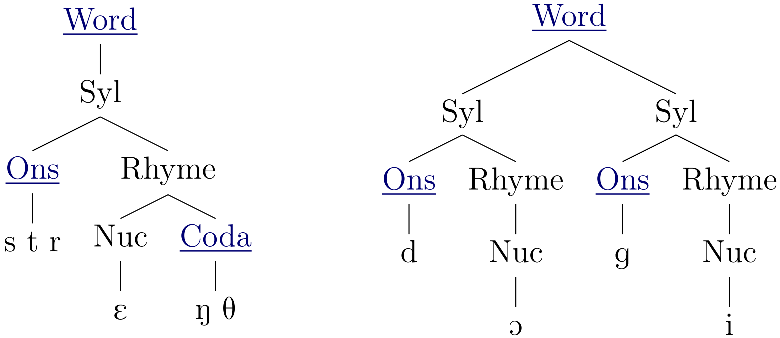
Baseline model 1: no stress or phonotactics

$$
\begin{array}{rcl}
\text{Sentence} & \rightarrow & \underline{\text{Colloc3}}^{+} \\
\underline{\text{Colloc3}} & \rightarrow & \underline{\text{Colloc2}}^{+} \\
\underline{\text{Colloc2}} & \rightarrow & \underline{\text{Colloc}}^{+} \\
\underline{\text{Colloc}} & \rightarrow & \underline{\text{Word}}^{+} \\
\underline{\text{Word}} & \rightarrow & \text{Syll}^{1:4} \\
\text{Syll} & \rightarrow & (\underline{\text{Onset}})\ \text{Rhyme} \\
\underline{\text{Onset}} & \rightarrow & \text{Consonant}^{+} \\
\text{Rhyme} & \rightarrow & \text{Nucleus}\ (\underline{\text{Coda}}) \\
\text{Nucleus} & \rightarrow & \text{Vowel}^{+} \\
\underline{\text{Coda}} & \rightarrow & \text{Consonant}^{+}
\end{array}
$$

- Same as *syllable collocation grammar* of Johnson (2008):

MACQUARIE UNIVERSITY

Sample parses of "no stress or phonotactics" grammar

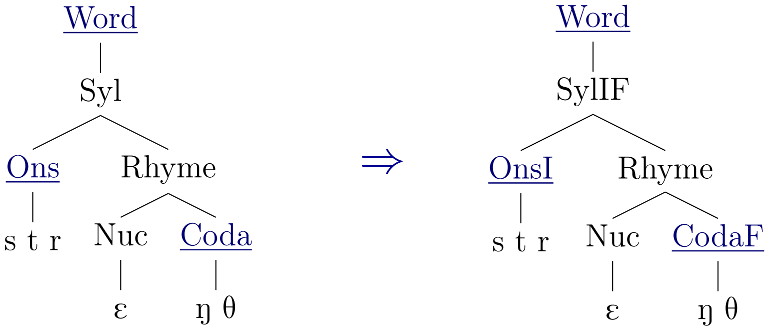


- Model learns a syllabification even though input is not syllabified

MACQUARIE UNIVERSITY

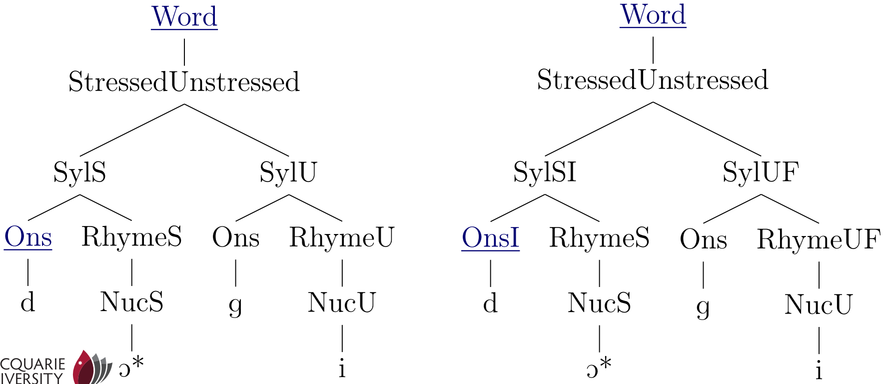
Baseline model 2: phonotactic but no stress generalisations

- Same as above, except that model distinguishes initial onsets OnsI and final codas CodaF
  - ⇒ model learns *word initial and word final clusters*
    - same as Johnson and Goldwater (2009)



Models that learn stress patterns

- Distinguishes stressed from unstressed syllables
  - input distinguishes stressed and unstressed vowels
- Learns all possible stress patterns (up to 4 syllables)
- Stress pattern probabilities are *learned jointly with segmentation*
- Can be combined with models that learn phonotactic generalisations



Outline

Stress and word segmentation

Computational models of word segmentation

Experiments

Conclusions and future work

MACQUARIE UNIVERSITY

Computational set-up

- All models use the same Adaptor Grammar software with the same hyperparameter settings
  - only the adaptor grammars vary

$\Rightarrow$ Any observed differences are due to differences in the models as encoded in the grammars (not implementation differences)

- Computational details (same as in Johnson and Goldwater 2009):
  - AG software uses a MCMC Metropolis-within-Gibbs algorithm
  - slice sampling for all Pitman-Yor hyperparameters with "vague priors"
  - 8 MCMC runs for each setting, each with 2,000 sweeps of training data
  - collect every 10th sweep of last 1,000 sweeps
  - identify most frequent segmentation for each utterance from these 800 samples

Experiment 1: training and testing on entire corpus

- Train and evaluate on entire corpus
- Also evaluate on held-out set of 1000 utterances
- Evaluate segmentation quality with *token f-score*

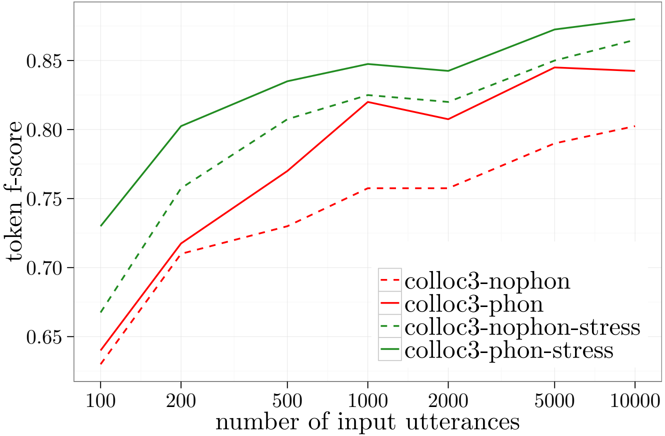
	phon	stress	train	held-out
baselines			.81	.81
	•		.85	.84
stress models		•	.86	.87
	•	•	**.88**	**.88**

⇒ Stress by itself improves segmentation accuracy slightly more than phonotactics (more so on held-out data)

MACQUARIE UNIVERSITY

Experiment 2: varying amount of training data

- Goal: Compare impact of stress on *inputs of different size*
  - perform inference over prefixes of corpus
  - evaluate on held-out data

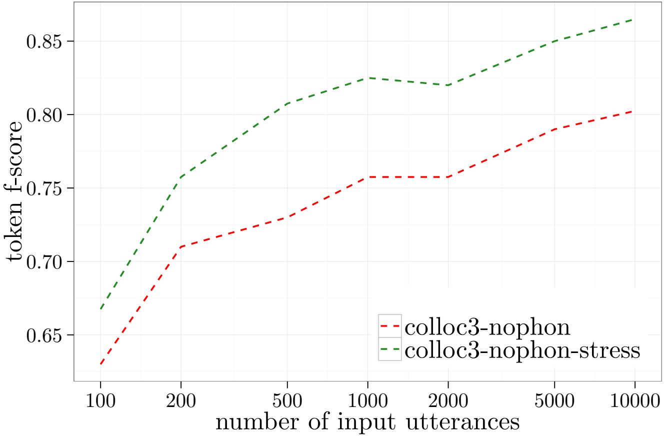


MACQUARIE UNIVERSITY

Stress without phonotactics

- Except on 100 utterances, *consistent improvement of 6-8%*

⇒ Quickly becomes powerful cue that aids segmentation

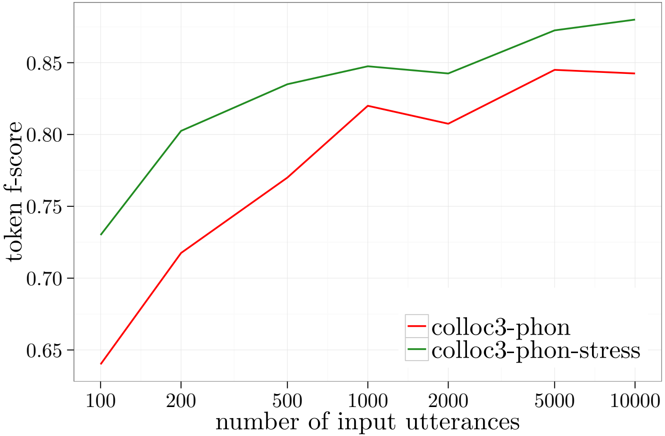


MACQUARIE UNIVERSITY

Interaction of stress and phonotactics

- Stress useful early on, but *relative importance diminishes with more data*
  - On full data, only 4% improvement (c.f., 7% without phonotactics)

⇒ Phonotactics partially redundant with stress with larger data



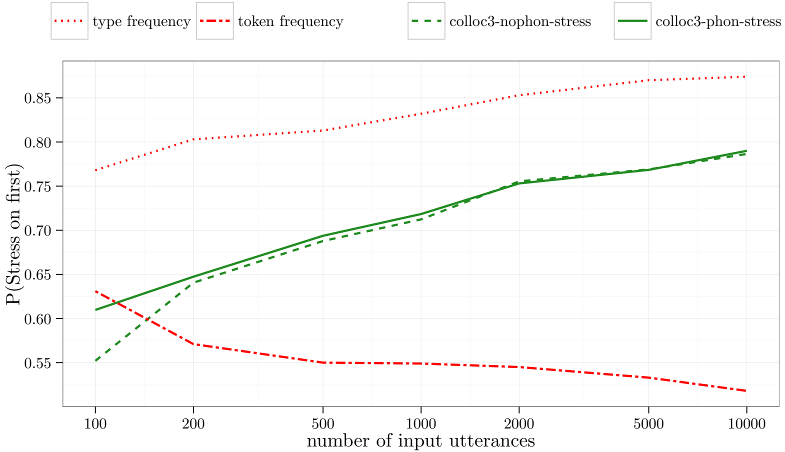
MACQUARIE UNIVERSITY

Identifying the stress patterns of a language

- Goal: identify the stress generalisations of a language
  - extract inferred posterior probabilities of Word expansions
    - e.g., $\mathrm{P}(\underline{\mathrm{Word}} \rightarrow \mathrm{StressedUnstressed})$ is probability of a word consisting of a Stressed followed by an Unstressed syllable
  - compare to empirical token / type fraction of each pattern
- This is a very simplified model of English stress
  - ignores interactions of stress with syllable weight, syntax, etc.

Induced stress patterns reflect type frequency

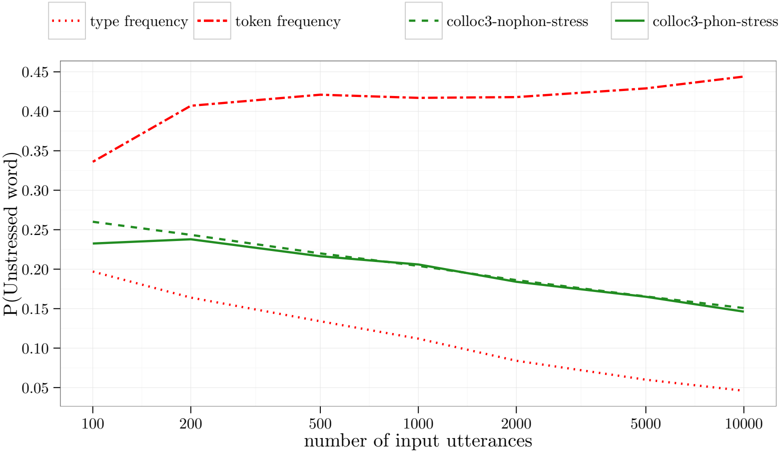
- Model's probability of initial stress reflects type rather than token frequency
  - these PCFG rules define the *base distribution* of the Word PYP



MACQUARIE UNIVERSITY

Unstressed words

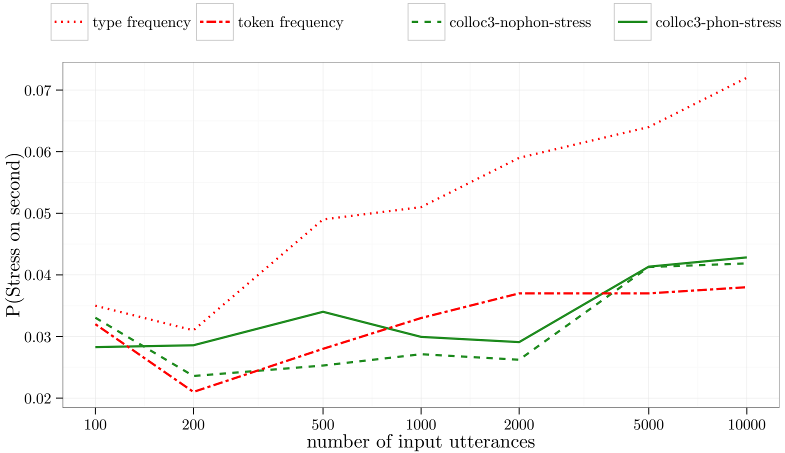
- Typically high token frequency function words
- True token / type fraction of pattern in red



MACQUARIE UNIVERSITY

Stress on second syllable

- Model does not identify low frequency stress-second pattern
- Consistent with observation that infants' struggle with this pattern



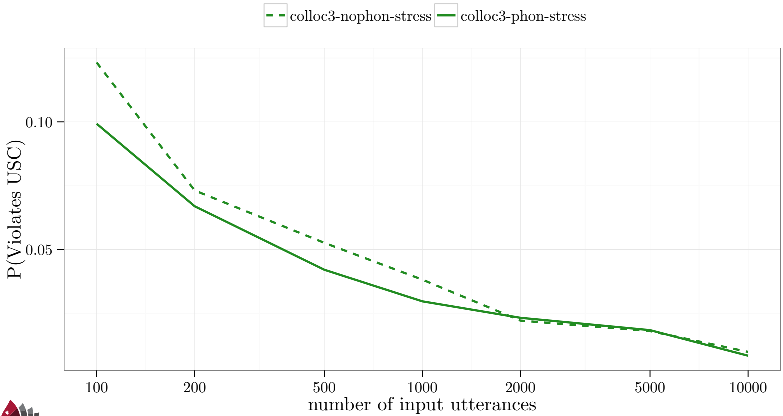
MACQUARIE UNIVERSITY

Unique stress constraint

- Probability of words with multiple stressed syllables approaches 0

⇒ Model learns that there is at most one stressed syllable per word

⇒ The Unique Stress Constraint (Yang 2004) can be acquired and does not need to be built in (?)



MACQUARIE UNIVERSITY

Outline

Stress and word segmentation

Computational models of word segmentation

Experiments

Conclusions and future work

MACQUARIE UNIVERSITY

Conclusions

- Adaptor Grammar models can exploit stress cues
  - consistent benefit by using stress (c.f. Yang / Lignos models)
  - acquires something like the Unique Stress Constraint
- Studied the interaction of stress and phonotactic cues
  - relative contribution of stress varies over time
- Bayesian learners can jointly infer the stress pattern of the language and use it to improve segmentation

MACQUARIE UNIVERSITY

Future work

- Cross-linguistic exploration of stress and other cues in languages besides English
- Use more realistic information rather than dictionary stress
- Providence corpus provides audio and video to derive ‘less idealized’ corpora
    - acoustic correlates of stress differ cross-linguistically
    - can we learn what (if anything?) corresponds to stress?