Exploring the role of stress in word segmentation using Adaptor Grammars

Mark Johnson
joint work with Benjamin Börschinger and Katherine Demuth

Macquarie University
Sydney, Australia

December 2014

MACQUARIE UNIVERSITY

Computational questions about language acquisition

- What *sources of information* are available to a language-learning child?
- *How do these information sources interact?*
  - *are any essential?* (e.g., convey information not available from other sources)
  - *are there redundancies?* (i.e., the same information can be obtained from two or more sources)
  - *are there dependencies?* (e.g., you have to learn the stress patterns of the language before they can be used to learn words)
  - *are there synergies?* (e.g., learning the stress patterns of the language helps identify word boundaries, and accurately learning word boundaries helps learn the stress patterns)
- These are *computational questions* that can be addressed with computational models

Talk overview

- This talk compares *four computational models of word segmentation* that are identical except that:
  - two can learn *phonotactic constraints* on possible words, and
  - two can learn *stress patterns* of possible words
- Questions this work addresses:
  - *does stress information help word segmentation?* (yes)
  - *is stress useful for word segmentation even if function words are unstressed?* (yes)
  - *are there synergies jointly learning stress and phonotactics* (yes, but decreases with more data)
  - *is stress redundant with phonotactics?* (not initially, but eventually both supply similiar information)
  - *can the English preference for word-initial stress be learnt from data?* (yes, it can be learnt)
  - *can the "one primary stress per word" constraint be learnt from data, or does it have to be innate?* (yes, it can be learnt)

Outline

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 ▵ ʊ ▵ k

ju wɑnt tu si ðə bʊk

"you want to see the book"

- Model's goal: determine location of word boundaries

  $\Rightarrow$ identifies the pronunciations of words in the transcript (a first step in learning the lexical entries)

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: *Glant*, *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

Adding stress to word-segmentation data

- We *annotate stress on the vowel nucleii of stressed syllables*

  j ▵ u ▲ w ▵ ɑ* ▵ n ▵ t ▲ t ▵ u ▲ s ▵ i* ▲ ð ▵ ə ▲ b ▵ ʊ* ▵ k

  - Johnson and Demuth (2010) annotated tone in Chinese in same way

- 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

Outline

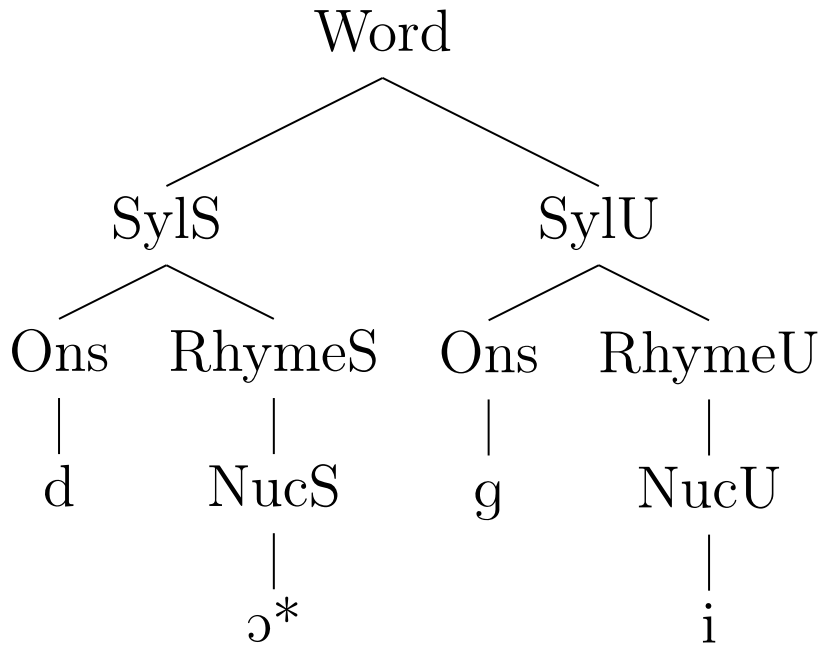
Computational models that exploit stress

- 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 (stress has a much larger effect in our model)

Useful cues for word segmentation

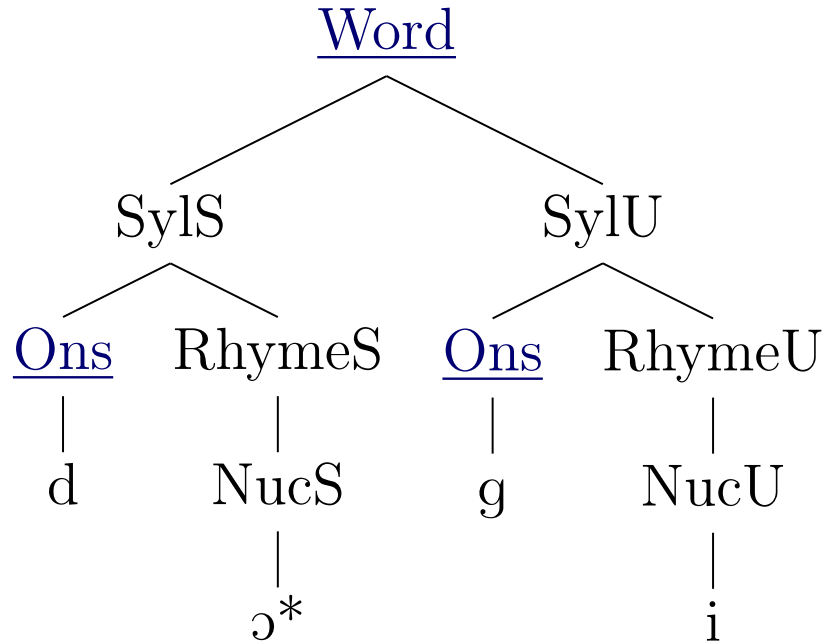
- *Vocabulary of the language*
  - no obvious upper bound $\Rightarrow$ *non-parametric* learning
- *Exhaustive parsing* (no unparsed speech)
- *Phonotactics* (e.g., syllable structure constraints)
- *Distributional cues* (e.g., collocations)
- Semantic constraints (e.g., word-topic mappings)
- Social cues (e.g., care-giver's eye-gaze)
- Morpho-syntax, e.g., function words
  (see Johnson et al 2014)
- *Prosodic cues*, specifically: *stress* (this paper)

Weaknesses of PCFGs for word segmentation



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

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 $\underline{\text{Word}} \Rightarrow^{+} \text{dɔ*gi}$ and $\underline{\text{Word}} \rightarrow \text{SylS SylU}$
  - each adapted nonterminal is associated with a Pitman-Yor Process (PYP)
    - PCFG rules specify *base distributions*

    $\Rightarrow$ defines a *hierarchy of PYPs*

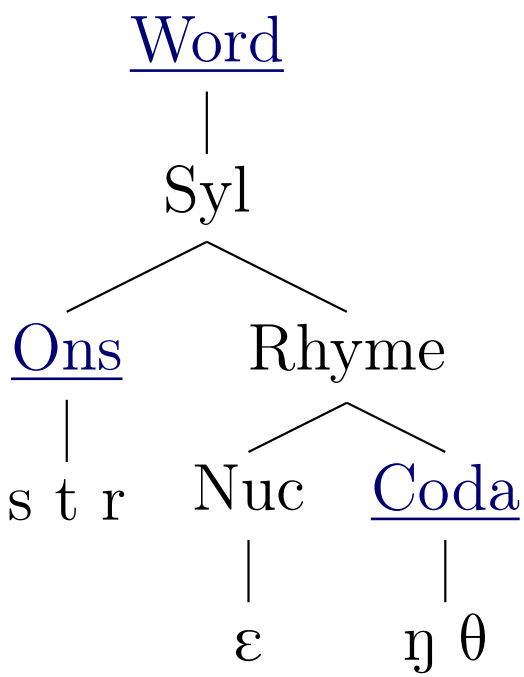
MACQUARIE UNIVERSITY

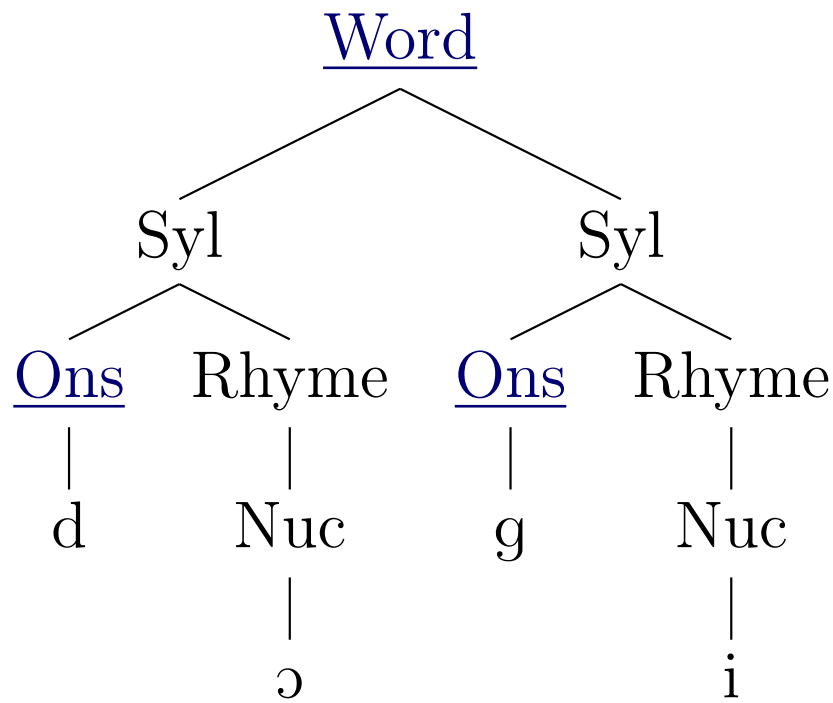
Baseline model 1: no stress or phonotactics

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

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

Sample parses of "no stress or phonotactics" grammar

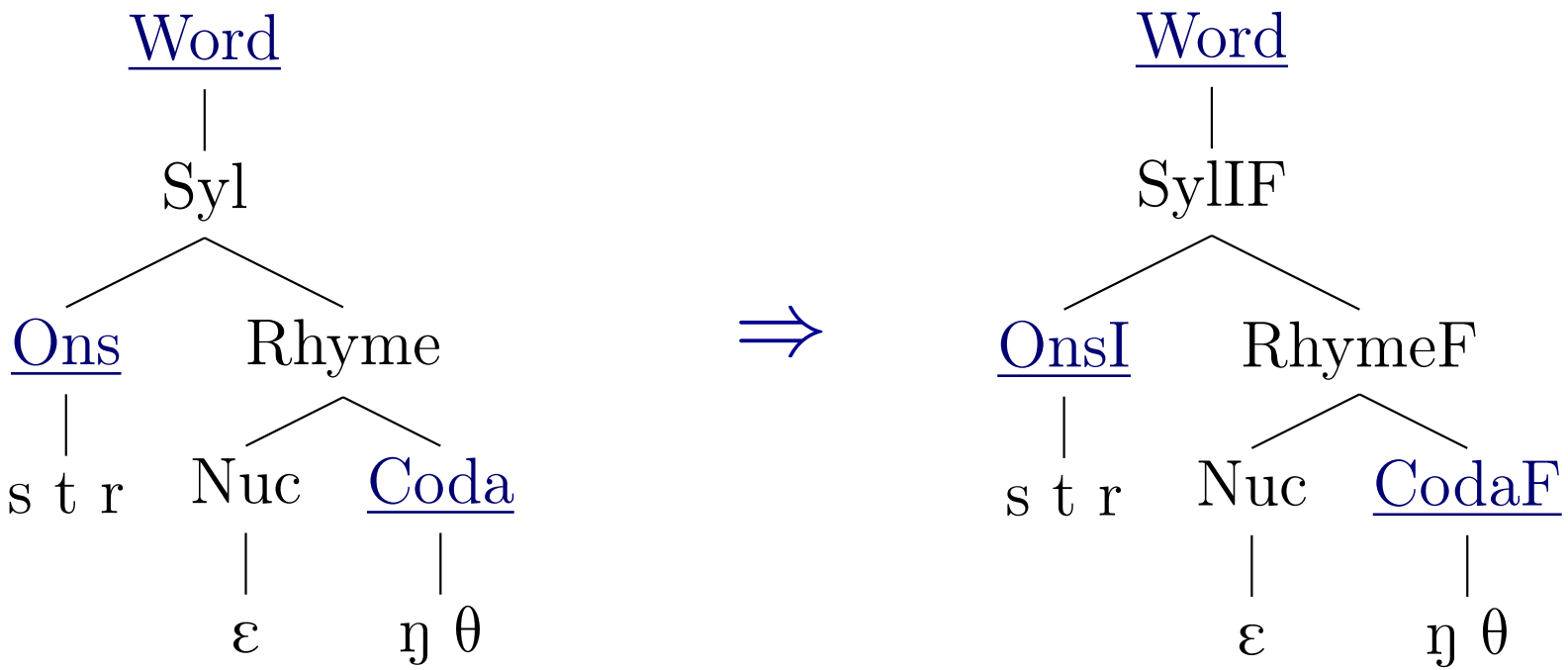




- Model learns a syllabification even though input is not syllabified

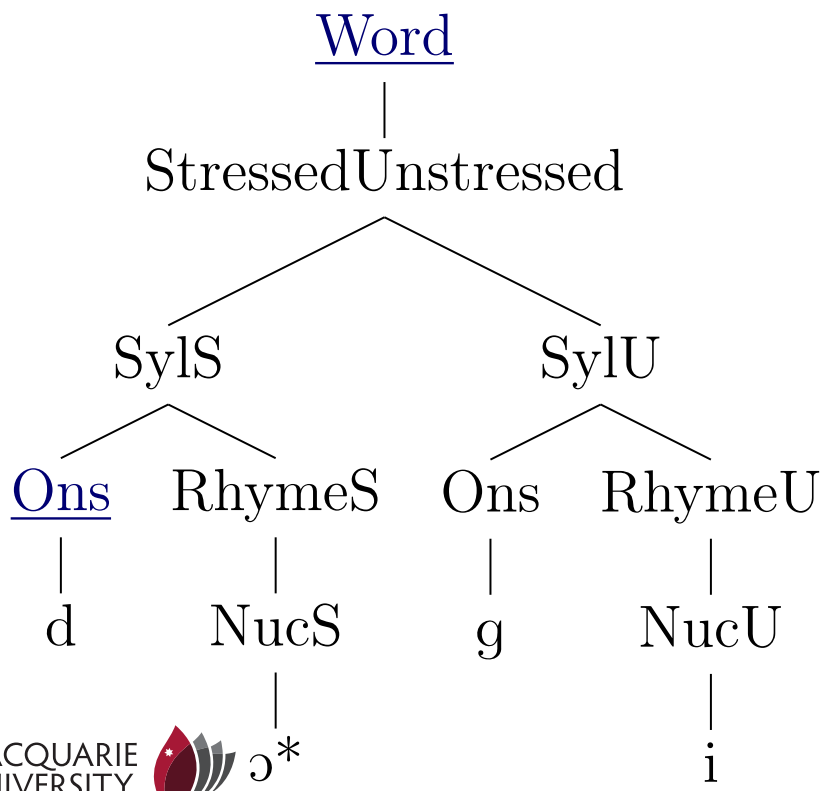
Baseline model 2: phonotactic but no stress generalisations

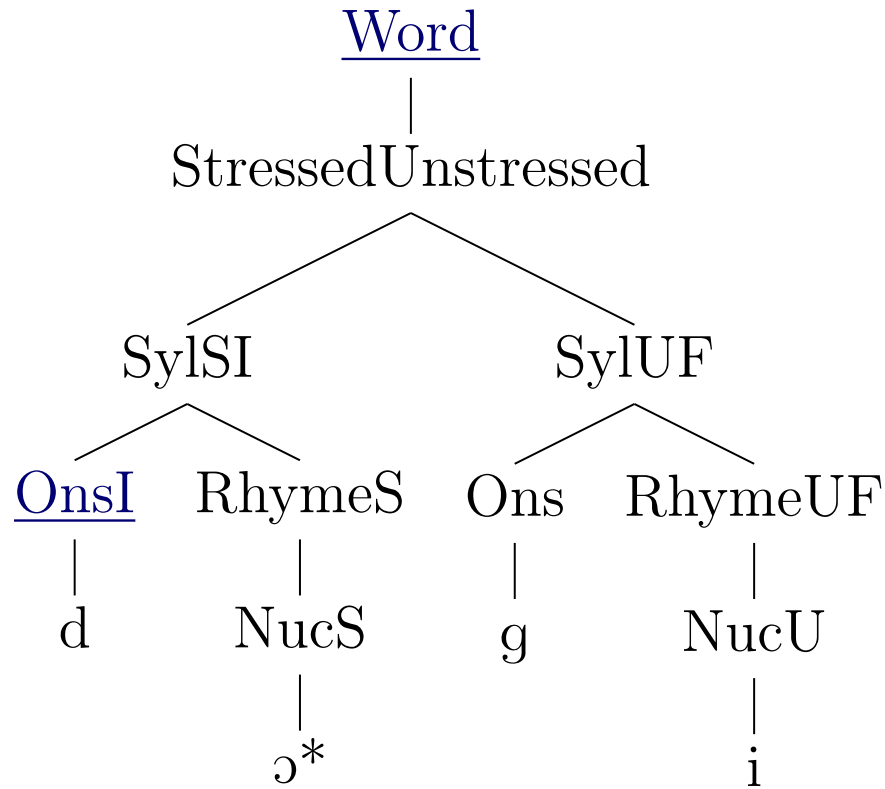
- Same as above, except that model distinguishes initial onsets OnsI and final codas CodaF
  - ⇒ model learns *word initial and word final clusters*
    - same as Johnson and Goldwater (2009)



Models that learn stress patterns

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





Outline

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*

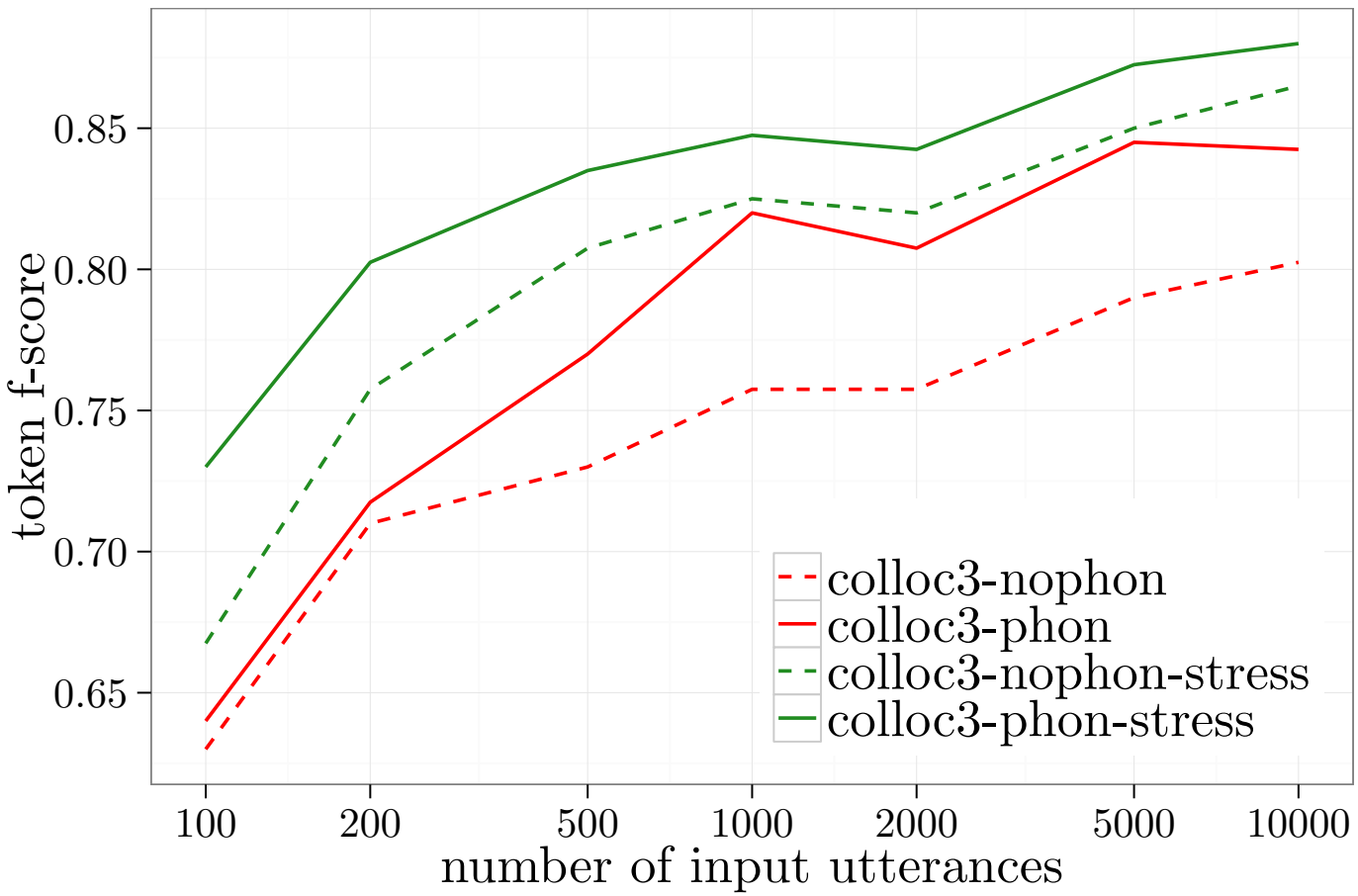
	phon	stress	train	held-out
baselines			.81	.81
	•		.85	.84
stress models		•	.86	.87
	•	•	**.88**	**.88**

⇒ 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

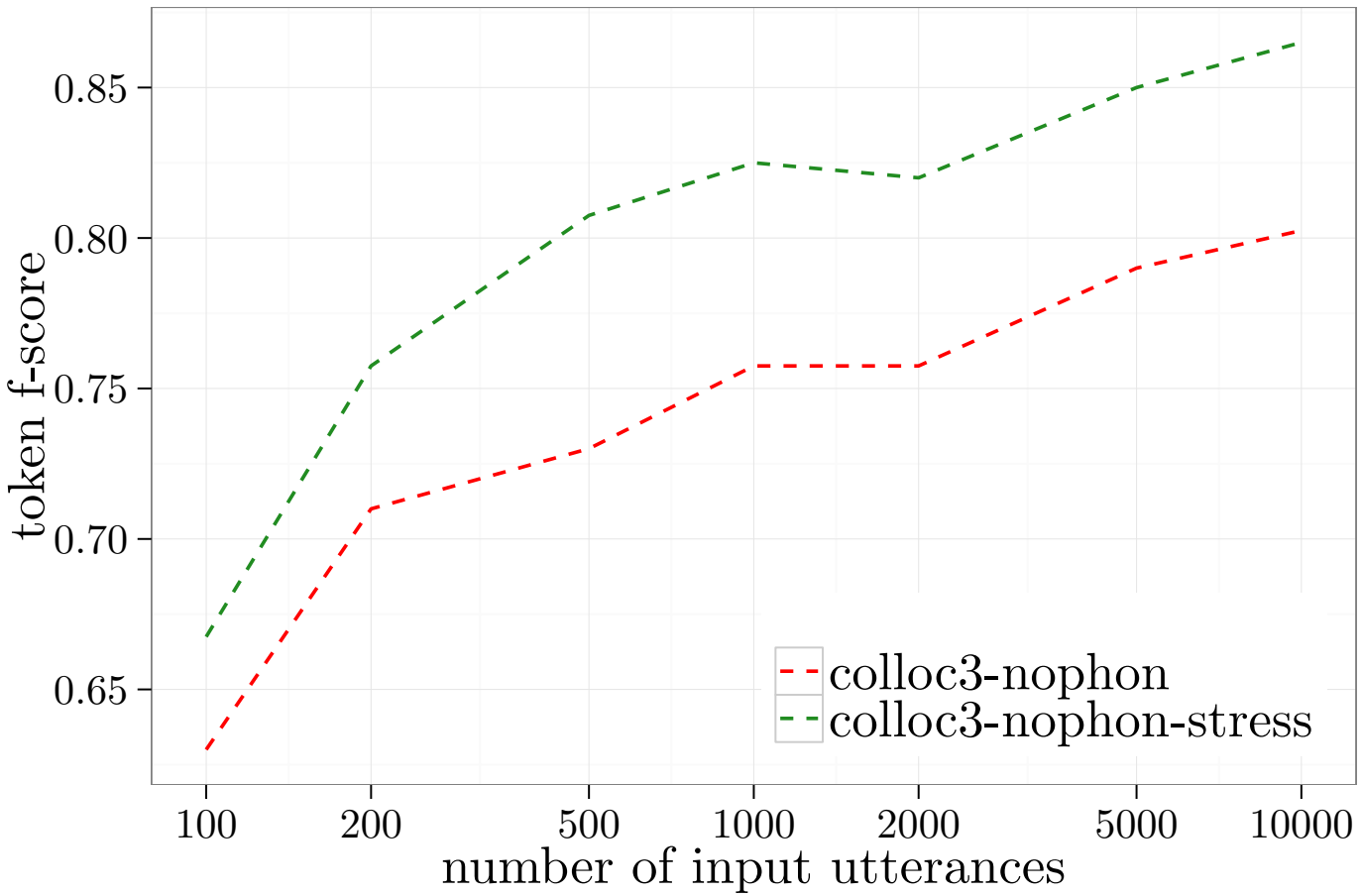


MACQUARIE UNIVERSITY

Stress without phonotactics

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

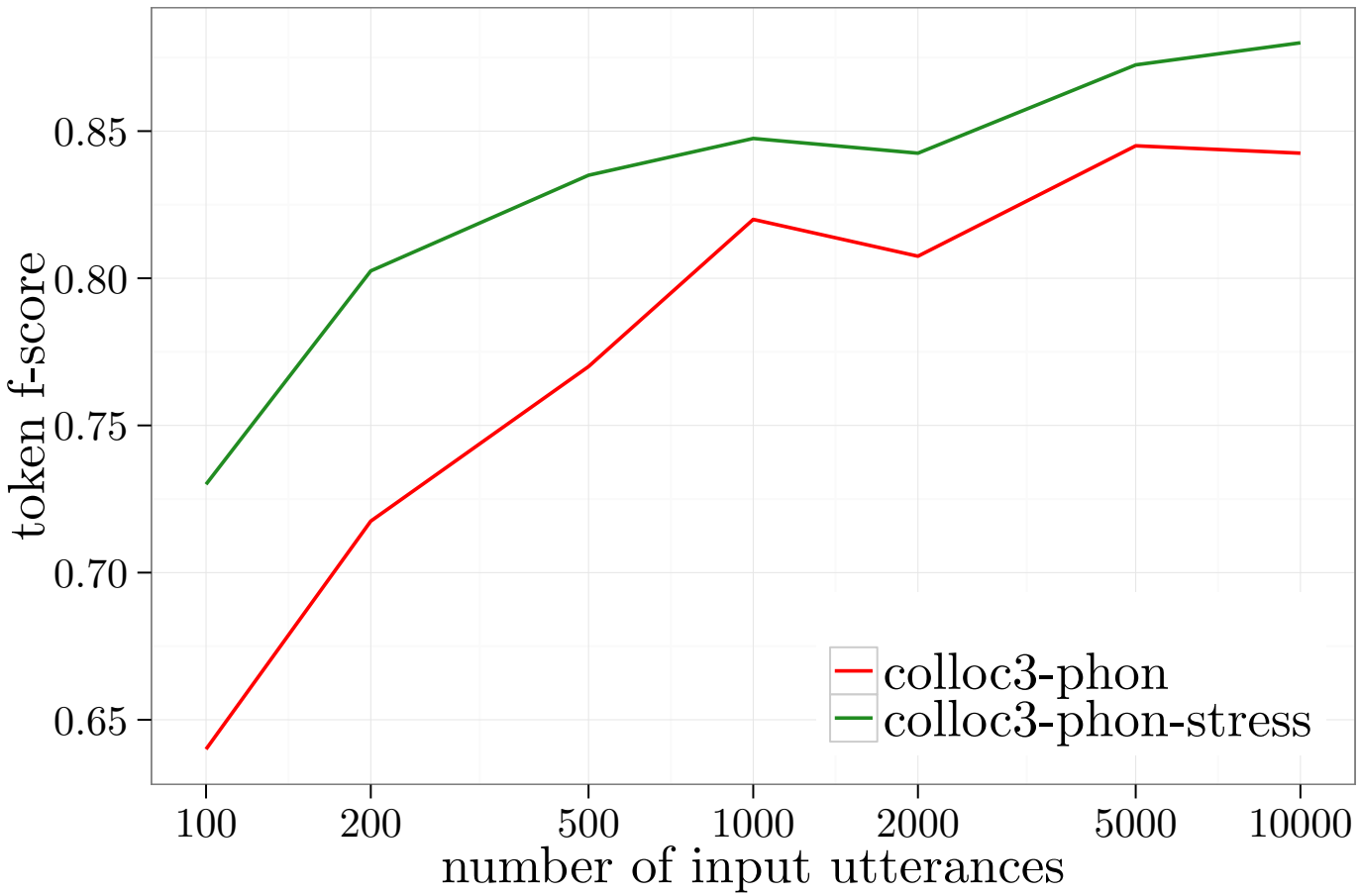
$\Rightarrow$ In absence of phonotactics, stress is a powerful cue across all data sizes



MACQUARIE UNIVERSITY

Interaction of stress and phonotactics

- On small data, stress and phonotactics interact synergistically
- Stress and phonotactics become more redundant as data grows
  - On full data, only 4% improvement (c.f., 7% without phonotactics)



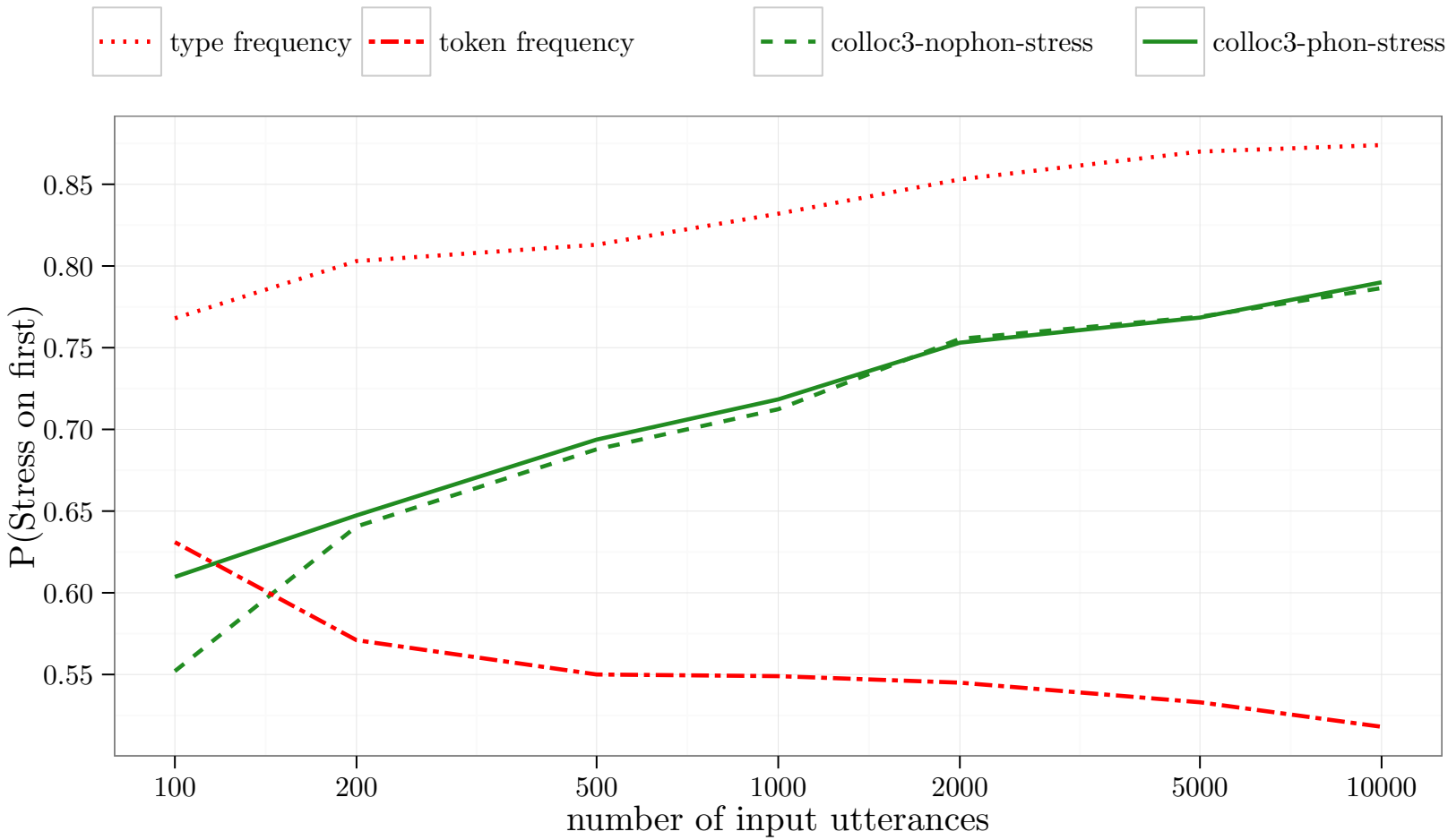
MACQUARIE UNIVERSITY

Identifying the stress patterns of a language

- Goal: *learn the stress generalisations of a language*
  - extract inferred posterior probabilities of $\underline{\text{Word}}$ expansions
    - e.g., $\text{P}(\underline{\text{Word}} \rightarrow \text{StressedUnstressed})$ 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.

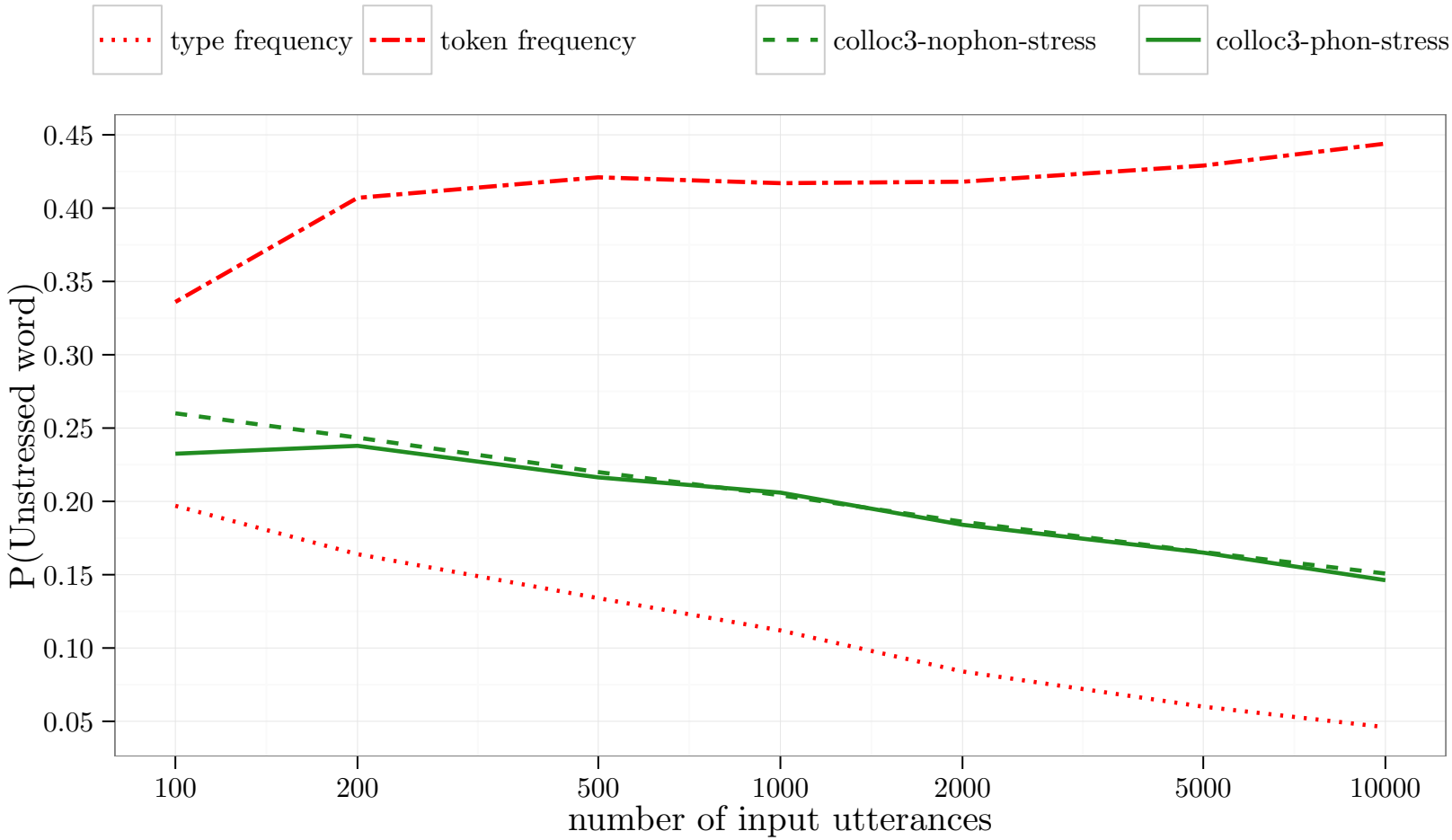
Model learns probability of initial stress

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



Model learns probability of unstressed words

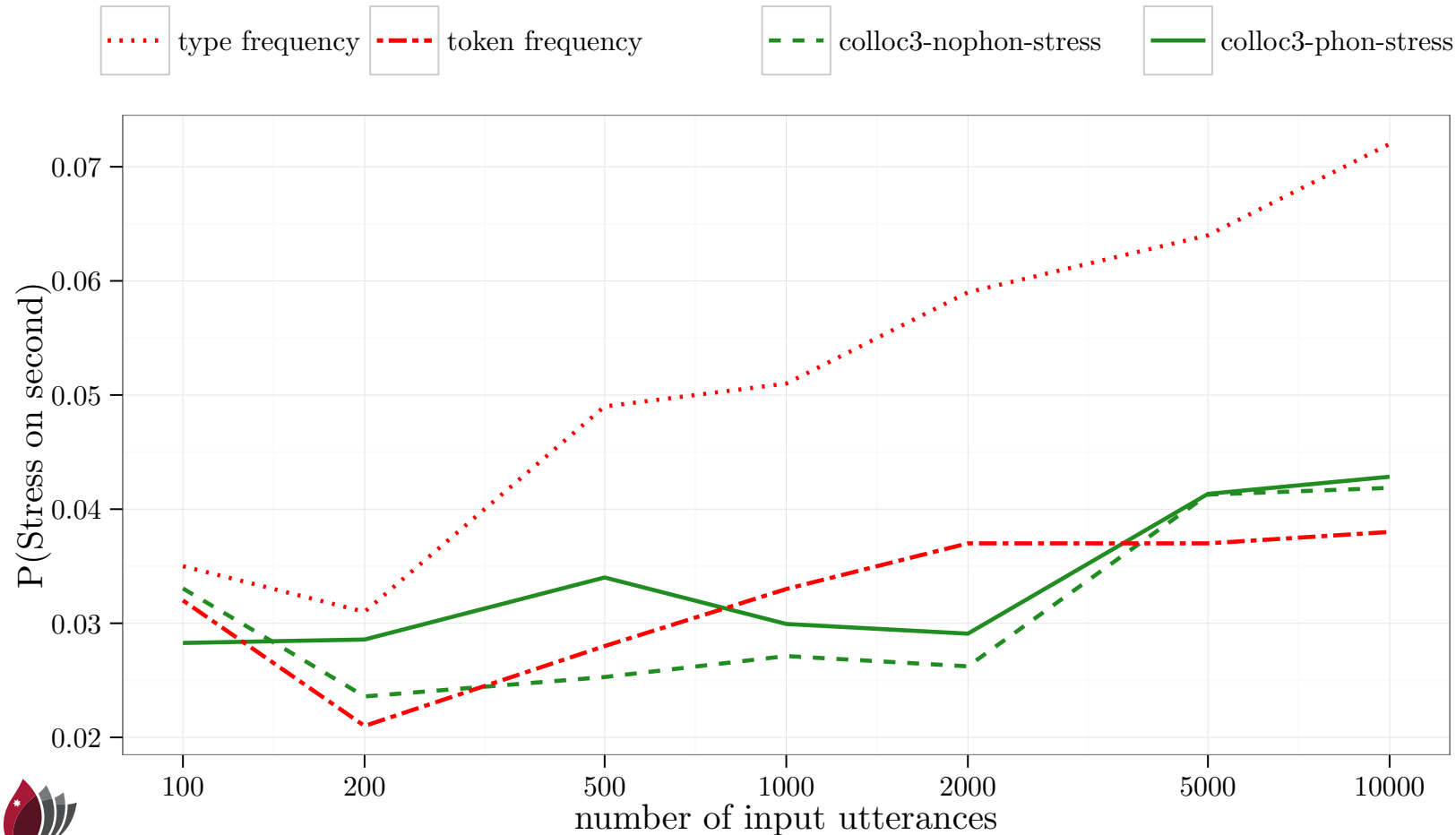
- Typically high token frequency function words
- Estimated probability tracks type frequency



MACQUARIE UNIVERSITY

Model does not accurately learn second syllable stress

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



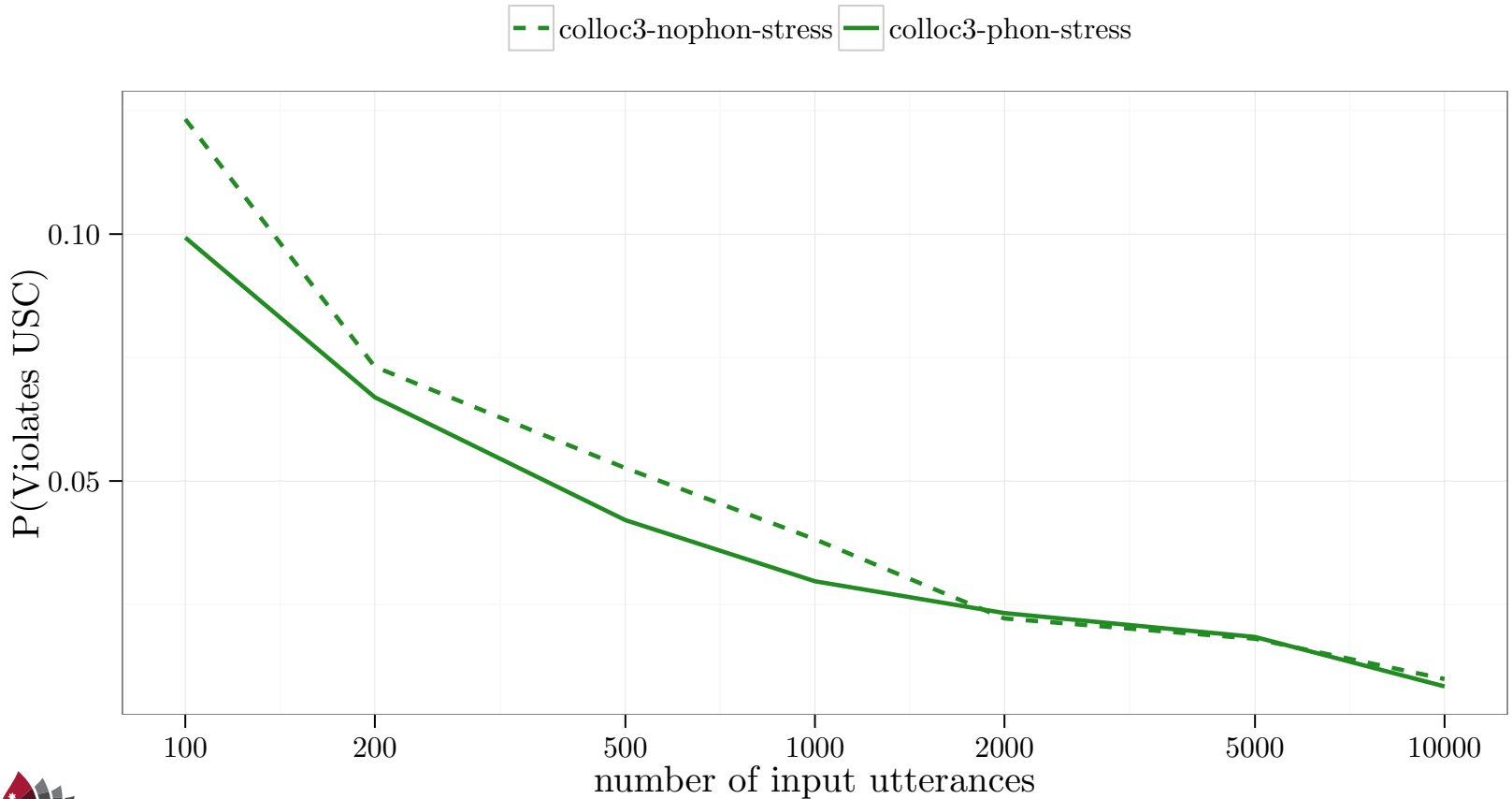
MACQUARIE UNIVERSITY

Unique stress constraint can be learnt

- Probability of words with multiple stressed syllables approaches 0

$\Rightarrow$ Model learns that there is at most one stressed syllable per word

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



MACQUARIE UNIVERSITY

Outline

MACQUARIE UNIVERSITY

Conclusions

- We performed a "controlled computational experiment" comparing models that can learn phonotactic and/or stress generalisations
- We showed that:
  - stress is a very useful cue, even if function words are unstressed
  - stress is more useful than phonotactics with little data, but the gap diminishes with more data
  - initially there are synergies jointly learning stress and phonotactics, but eventually they provide redundant information
  - the English preference for word-initial stress can be learnt
  - the "one primary stress per word" constraint can be learnt (i.e., does not have to be innate)

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?