How Ideal Are We?
Incorporating Human Limitations into Bayesian Models of Word Segmentation

Lisa Pearl, Sharon Goldwater, & Mark Steyvers

Human behavior is often consistent with the predictions of Bayesian ideal learners (e.g., Xu &Tenenbaum, 2007; Griffiths &Tenenbaum, 2005), which aim to explain why humans behave as they do, given the task and data they encounter. However, these models typically avoid answering *how* the observed behavior is produced, given human limitations on memory and processing. Here, we ask how such limitations might affect the results of identifying words in continuous speech, using a corpus of English child-directed speech (Bernstein-Ratner, 1984). Simulations with different algorithms suggest that results depend non-trivially on how the learner's limitations are implemented. Also, though these learners do not segment realistic speech as well as the most successful ideal learner, they outperform other purely statistical learning strategies, such as syllable transitional probability (Saffran et al, 1996; see Gambell &Yang (2006)).

We begin with the Bayesian model of word segmentation in Goldwater, Griffiths, and Johnson (2006) (GGJ), which provides an ideal learning analysis of how statistical information, a language-independent cue preferred by infants early in development (Thiessen & Saffran, 2003), could be used to begin identifying words in continuous speech. GGJ develop two model variants: in the *unigram* model, the learner assumes word context is not important; in the *bigram* model, context is used to guide segmentation decisions. GGJ demonstrate that ideal learners biased to heed context are more successful at word segmentation, since learners ignoring context tend to identify words that often occur together as one word, thereby committing undersegmentation.

The ideal learner can access the entire corpus simultaneously, equivalent to an infant remembering several weeks' worth of utterances in detail. To simulate limited memory resources, we present three *online* algorithms, where the learner segments one utterance and then moves on to the next one: Dynamic Programming Maximization (DPM), Dynamic Programming Sampling (DPS), and Decayed Markov Chain Monte Carlo (DMCMC). DPM and DPS also limit the hypotheses the learner can keep in mind, while DMCMC implements a recency effect.

Results of our simulations indicate that the performance of all the learners stabilizes rapidly, within about 2000 utterances (see Fig. 1 for examples). Table 1 gives the final performance of all learners. We find that adding limitations to a learner who assumes context is unimportant (a unigram learner) can actually improve performance by reducing the undersegmentation found in the ideal unigram learner (in fact, DPS and DPM exhibit slight oversegmentation). The different online learners show varying behavior when context is taken into account: DPM shows improved segmentation (like the ideal learner), while DPS and DMCMC fare worse than when ignoring context. Still, while no online models outperform the ideal bigram learner, all outperform the syllable transitional probability learner.

Our results show that a simple intuition (infants have memory limitations) can be cashed out multiple ways, and what learning biases are most successful depends on how these limitations are implemented. More practically, if infants require a seed pool of words to identify language-dependent strategies, these online language-independent strategies may provide a pool reliable enough to do so.

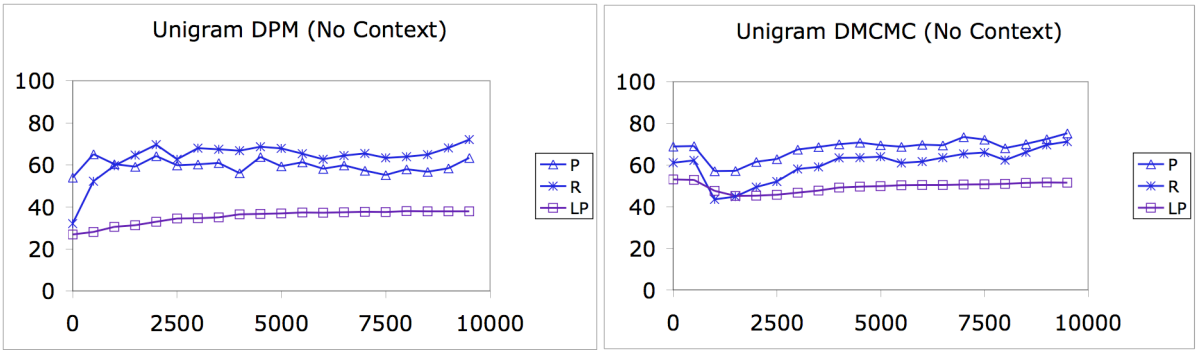

Figure 1. Performance of DPM and DMCMC unigram models over the corpus as a whole, divided into groups of 500 utterances (x-axis = utterance number, y-axis = score percentage). Precision (P) represents how often the learner is correct when believing something is a word; recall (R) represents how often the learner finds a word it should have found. Lexicon precision (LP) is measured over the lexicon inferred from the segmentation, which represents the vocabulary items identified by the learner.

	Precision	Recall	F-Score	Lex-Precision
Gambell & Yang (2006), testing Saffran et al. (1996)				
Syllable Transitional Probability	41.6	23.3	29.9	
Unigram Models (No Context)				
GGJ – Ideal	61.7	47.1	53.4	55.1
DPM	64.5	69.3	66.8	59.5
DPS	58.6	65.5	61.9	51.8
DMCMC	70.7	64.7	67.6	56.7
Bigram Models (Context)				
GGJ – Ideal	74.6	68.4	71.4	63.3
DPM	66.0	70.8	68.3	64.4
DPS	32.7	48.4	39.0	34.1
DMCMC	52.7	44.5	48.3	22.5

Table 1. Performance of different learning models on the second half of the corpus (to factor out learning curve differences). Precision and recall over word tokens are shown, as well as the F-score which combines these two scores into one score for easy comparison. Lexicon precision is included where available to indicate the accuracy of the seed pool of words identified.

References:

Bernstein-Ratner, N. (1984). Patterns of vowel Modification in motherese. *Journal of Child Language. 11,* 557-578.

Gambell, T. & Yang, C. (2006). Word Segmentation: Quick but not Dirty. Manuscript. Yale University.

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

Griffiths, T. L., & Tenenbaum, J. B. (2005). Structure and strength in causal induction. *Cognitive Psychology, 51,* 354-384.

Saffran, J., Aslin, R., & Newport, E. (1996). Statistical learning by 8-month-olds. *Science*, *274*, 1926-928.

Thiessen, E., & Saffran, J. R. (2003). When cues collide: Use of stress and statistical cues to word boundaries by 7- to 9-month-old infants. *Developmental Psychology, 39*, 706–716.

Xu, F. & Tenenbaum, J. B. (2007). Word learning as Bayesian inference. *Psychological Review*, 114:245-272.