Taking the child's view:
Syllable-based Bayesian inference as a (more) plausible statistical word segmentation strategy

Lawrence Phillips and Lisa Pearl
(lawphill@uci.edu, lpearl@uci.edu)
University of California: Irvine

Because knowledge of words plays a crucial role in acquisition and children seem to accomplish word segmentation very early (~7.5 months (Jusczyk et al., 1999; Echols et al., 1997; Jusczyk et al., 1993a)), many strategies have been proposed for how children learn to identify words in their native language. Because of experimental evidence that infants are sensitive to statistical information in their environment (e.g. Saffran, Aslin & Newport, 1996) statistical strategies have seen a rise in popularity. Many recent statistical models of word segmentation have assumed one basic unit of representation available to the learner is the phoneme (e.g., Goldwater, Griffiths, Johnson 2009 (***GGJ2009***); Brent & Siskind, 1999; Johnson & Goldwater, 2009; Pearl, Goldwater & Steyvers, 2011 (***PGS2011***)). However, this supposes that learners have already discovered their language's phonemic inventory, while experimental evidence suggests that this is not true for young infants (Jusczyk et al., 1993b; Jusczyk et al., 1994). Instead, syllables seem to be stronger representational units at this stage (Jusczyk & Derrah, 1987; Eimas, 1999; Saffran, Aslin & Newport, 1996). While the success of previous statistical word segmentation models is heartening, how dependent is that success on the assumption of the phoneme as a representational unit? With this question in mind, we modify existing, highly successful, phoneme-based statistical models of word segmentation that use Bayesian inference (GGJ2009; PGS2011) to operate over syllables and so create a more psychologically faithful model of word segmentation. Interestingly, we find that these Bayesian word segmentation approaches are even *more* successful when operating over syllables, but only when learners assume words depend on the words before them (a bigram assumption). This demonstrates the robustness of this purely statistical approach, though only when certain assumptions hold. In addition, we replicate and extend results from PGS2011 concerning the surprising utility of processing constraints for Bayesian word segmentation strategies.

All our modified Bayesian learners treat syllables as atomic units, with the idea that this is more psychologically faithful. In particular, this mimics the performance of infants who are able to discriminate between the three syllables /ba/, /bu/, and /lu/, but who are unable to recognize that /ba/ and /bu/ are more similar than /ba/ and /lu/ (Jusczyk & Derrah, 1987). While this alleviates the learning problem to some extent, because it reduces the number of potential segmentations, a potential sparse data problem then surfaces. In particular, while a model operating over English phonemes must track statistics between approximately 40 units of representation, a model operating over English syllables must deal with statistics between approximately 4000 units, while still using the same quantity of data as a phoneme-based model. Additionally, almost all phonotactic information about English, which can be helpful for word segmentation (Blanchard, Golinkoff & Heinz, 2010), is lost to the model.

We test our syllable-based models using English child-directed speech from the Pearl-Brent corpus (CHILDES: MacWhinney, 2000). We restrict ourselves to child-directed utterances before 9 months, approximately 28,000 utterances. We compare ideal learners, which have no processing constraints, to more psychologically plausible constrained learners that segment utterances as they are encountered and sometimes perform non-optimal statistical inference. Additionally, we compare modeled learners that assume words are produced independent of all other words (a *unigram* assumption (GGJ, 2009)) with modeled learners that assume a word depends on the word that occurred directly before it (a *bigram* assumption (GGJ, 2009)).

All modeled learners use the Bayesian generative model framework described in GGJ (2009), which implicitly incorporates preferences for smaller lexicons and shorter words in the lexicon. The ideal learner uses Gibbs sampling to converge on the optimal word segmentation, and sees all utterances at once when making decisions about where boundaries should be placed. We compare this learner with a number of constrained learners implemented in PGS (2011). The Dynamic Programming Maximization (DPM) learner is motivated by the insight that linguistic processing is incremental, and so processes data as they appear, rather than in a batch. The DPM learner additionally uses the Viterbi algorithm to converge on the optimal word segmentation, based on the data already encountered. The Dynamic Programming Sampling (DPS) learner also makes decisions incrementally, but rather than necessarily choosing the optimal segmentation, it samples a potential segmentation probabilistically. Therefore, the optimal segmentation is most likely to be chosen, but occasionally the learner will select unlikely segmentations as well. The Decayed Markov Chain Monte Carlo (DMCMC) learner also processes data incrementally, but uses a modified form of Gibbs sampling to implement a recency effect, where more recent word boundaries are more likely to be sampled, mimicking memory constraints.

Table 1 shows the results of these syllable-based learners, as compared with their previous phoneme-based counterparts. We measure our results in terms of precision, recall, and F-score (the harmonic mean) of individual word tokens, word boundaries, and lexicon items. We find that the phoneme-based learners perform better than the syllable-based learners when using the unigram assumption, with some of the syllable-based learners performing barely better than a simple syllable-based transitional probability learner. However, this trend is notably reversed when learners have the bigram assumption: all syllable-based learners significantly outperform the transitional probability learner, and most significantly outperform their phoneme-based counterparts. This suggests that syllable-based Bayesian inference is only a useful word segmentation strategy if children know that words depend on one another. Table 2 compares our results with those of Gambell & Yang (2006) a previous syllable-based model which incorporates stress information as well as the knowledge that words can contain only a single primary stress. By adding this linguistic knowledge they achieve higher performance than even our DMCMC learner. Their model, however, takes stress information from a pronunciation dictionary which does not reflect how words are stressed in spoken language. The model is therefore able to segment strings of monosyllabic words effortlessly, when in reality these strings may be quite difficult for children.

In addition, there is empirical evidence that children undersegment the utterances they hear, grouping together commonly occurring words (Peters, 1983). Based on the boundary precision and recall scores, we can tell whether a particular modeled learner is undersegmenting. High boundary precision and low recall indicates that the learner is highly accurate when placing a boundary, but does not insert enough boundaries in general, thereby undersegmenting the corpus. As Table 1 shows, while only some of the phoneme-based Bayesian learners undersegment the data, we find that nearly *all* syllable-based learners – ideal and constrained – show undersegmentation behavior. Thus, syllable-based Bayesian learners match this empirical behavior better.

Moreover, we also find support for the somewhat counter-intuitive “Less is More” hypothesis (Newport, 1990), where processing constraints placed on children are hypothesized to help, rather than hinder, language acquisition. In particular, while we do find that an ideal Bayesian learner with unlimited memory and processing resources can succeed, we crucially find results similar to PGS (2011) that constrained learners who learn incrementally and with limited memory, as in the case of actual children, *outperform* the ideal learner. While there is a literature in computer science on the benefits unsupervised models gain from online learning (Liang & Klein 2009), our model benefits not from online learning, but from suboptimal sampling (DPS learner) and memory constraints (DMCMC learner). This indicates that non-optimal segmentation strategies may be useful in acquiring word segmentation, although the reason for this behavior is still poorly understood. By examining the specific constrained strategies learners could use, and their resulting segmentation effects, we may be able to offer an explanation for *why* processing constraints could help language acquisition. A literature exists on “Less is More” findings in artificial language learning (Chin & Kersten 2010; Kersten & Earles 2001; Cochran et al. 1999), but to the author’s knowledge there are no experiments that show why constrained processing helps in acquisition. This highlights one very major contribution computational modeling can make to developmental linguistics.

In the broader picture, this study highlights the benefits of using empirical research from psychology to inform decisions on how to model language acquisition: not only can we identify the strategies that are likely to be used by children, but we may also discover potential explanations for existing, sometimes puzzling, observations about child language acquisition, as with the “Less is More” hypothesis. While our preliminary work focuses on a single English corpus, we plan on extending these models to multiple corpora across languages varied in their manner of syllabification.

Unigram Models (words are independent)									
	TP	TR	TF	BP	BR	BF	LP	LR	LF
Ideal-Pho	63.2	48.4	54.8	92.8	62.1	74.4	54.0	73.6	62.3
Ideal-Syl	**65.34**	**45.85**	**53.89**	**92.20**	**56.38**	**71.63**	**45.59**	**71.78**	**55.75**
DPM-Pho	63.7	68.4	65.9	77.2	85.3	81.0	61.9	56.9	59.3
DPM-Syl	**71.97**	**48.58**	**57.96**	**98.07**	**52.50**	**68.32**	**37.35**	**53.14**	**43.86**
DPS-Pho	55.0	62.6	58.5	70.4	84.2	76.7	54.8	49.2	51.8
DPS-Syl	**74.33**	**53.27**	**62.03**	**97.20**	**57.9**	**72.51**	**41.17**	**57.21**	**47.87**
DMCMC-Pho	71.2	64.7	67.8	88.8	77.2	82.6	61.0	69.6	65.0
DMCMC-Syl	**67.31**	**49.67**	**57.16**	**96.82**	**60.55**	**74.48**	**48.74**	**72.79**	**58.38**
Bigram Models (words depend on previous words)									
Ideal-Pho	74.5	68.8	71.5	90.1	80.4	85.0	65.0	73.5	69.1
Ideal-Syl	**81.84**	**72.08**	**76.65**	**96.05**	**79.67**	**87.09**	**65.27**	**79.06**	**71.50**
DPM-Pho	67.5	71.3	69.4	80.4	86.8	83.5	66.0	63.2	64.6
DPM-Syl	**81.49**	**68.57**	**74.46**	**96.67**	**74.84**	**84.35**	**56.96**	**70.46**	**62.99**
DPS-Pho	34.2	47.6	39.8	54.9	85.3	66.8	39.0	34.4	36.5
DPS-Syl	**82.96**	**71.34**	**76.70**	**96.48**	**77.20**	**85.75**	**57.83**	**71.23**	**63.83**
DMCMC-Pho	72.0	74.0	73.0	84.1	87.4	85.7	61.1	64.2	62.6
DMCMC-Syl	**87.19**	**85.23**	**86.19**	**94.01**	**91.05**	**92.49**	**74.18**	**77.28**	**75.70**
Comparison Models									
TransProb-Pho	34.3	42.7	38.0	52.8	71.1	60.6	24.3	39.7	30.1
TransProb-Syl	**53.03**	**37.57**	**43.98**	**90.00**	**53.14**	**66.82**	**11.72**	**63.08**	**19.77**

Table 1. Comparison of phoneme-based (-Pho) and syllable-based (-Syl) learners on three sets of measures: word tokens (T), word boundaries (B), and lexicon items (L). For each, precision (P), recall (R), and F-score (F) are shown. P = # correctly identified / # identified, R= # correctly identified / # should have been identified, F = harmonic mean of P and R ((2*P*R)/(P+R)). Ideal = optimal segmentation, batch learning. DPM = optimal segmentation, incremental learning. DPS = non-optimal segmentation, incremental learning. DMCMC = non-optimal segmentation, incremental learning, recency effect. A syllabic transitional probability (TransProb) learner (based on Saffran, Aslin, & Newport (1996)) is provided as a baseline.

Unigram Models (words are independent)			
	TP	TR	TF
Ideal	65.34	45.85	53.89
DPM	71.97	48.58	57.96
DPS	74.33	53.27	62.03
DMCMC	67.31	49.67	57.16
Bigram Models (words depend on previous words)			
Ideal	81.84	72.08	76.65
DPM	81.49	68.57	74.46
DPS	82.96	71.34	76.70
DMCMC	87.19	85.23	86.19
Comparison Models			
TransProb	53.03	37.57	43.98
TP (Brown)	41.6	23.3	29.8
TP+USC (Brown)	73.5	71.2	72.3
Algebraic agnostic (Brown)	85.9	89.9	87.9
Algebraic random (Brown)	95.9	93.4	94.6

Table 2. Comparison of our results against other syllable-based learners.

References

Brent, M.R. & Siskind, J.M. 2001. The role of exposure to isolated words in early vocabulary development. *Cognition*, *81*, 31-44.

Chin, S.L. & Kersten, A.W. In Press. The application of the less is more hypothesis in foreign language learning. *Proceedings of the 32nd Annual Conference of the Cognitive Science Society*.

Cochran, B., McDonald, J. & Parault, S. 1999. Too smart for their own good: The disadvantage of superior processing capacity for adult language learners. *Journal of Memory and Language*, 41, 30-58.

Echols, C.H., Crowhurst, M.J. & Childers, J.B. 1997. The perception of rhythmic units in speech by infants and adults. *Journal of Memory and Language*, 36, 202-225.

Eimas, P.D. 1999. Segmental and syllabic representations in the perception of speech by young infants. *Journal of the Acoustical Society of America*, 105(3), 1901-1911.

Frank, M. C., Goodman, N. D., & Tenenbaum, J. 2009. Using speakers' referential intentions to model early cross-situational word learning. *Psychological Science, 20*, 579–585.

Gambell, T. & Yang, C. 2006. Word Segmentation: Quick but not dirty. Manuscript. New Haven: Yale University

Goldwater, S., Griffiths, T. & Johnson, M. 2009. A Bayesian framework for word segmentation: Exploring the effects of context. *Cognition 112*(1), 21-54.

Jusczyk, P.W. & Derrah, C. 1987. Representation of speech sounds by young infants. *Developmental Psychology*, 23(5), 648-654.

Jusczyk, P.W., Cutler, A. & Redanz, N.J. 1993a. Infants' preference for the predominant stress patterns of English words. *Child Development*, 64(3), 675-687.

Jusczyk, P.W., Friederici, A.D., Wessels, J.M.I., Svenkerud, V.Y. & Jusczyk, A.M. 1993b. Infants's sensitivity to the sound patterns of native language words. *Journal of Memory and Language*, 32, 402-420.

Jusczyk, P.W., Luce, P.A. & Charles-Luce, J. 1994. Infants' sensitivity to phonotactic patterns in the native language. *Journal of Memory and Language*, 33, 630-645.

Jusczyk, P.W., Houston, D.M. & Newsome, M. 1999. The beginnings of word segmentation in English learning infants. *Cognitive Psychology*, 39, 159-207.

Kersten, A.W. & Earles, J.L. 2001. Less really is more for adults learning a miniature artificial language. *Journal of Memory and Language*, 44, 250-273.

Liang, P. & Klein, D. 2009. Online EM for unsupervised models. *Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the ACL*, 611-619.

MacWhinney, B. 2000. *The CHILDES project: Tools for analyzing talk*. Mahwah, NJ: Lawrence Erlbaum Associates.

Newport, E. 1990. Maturational constraints on language learning. *Cognitive Science*, *14*, 11-28.

Pearl, L., Goldwater, S., & Steyvers, M. 2011. Online Learning Mechanisms for Bayesian Models of Word Segmentation, *Research on Language and Computation*, special issue on computational models of language acquisition. DOI 10.1007/s11168-011-9074-5.

Peters, A. (1983). *The Units of Language Acquisition, Monographs in Applied Psycholinguistics*, New York: Cambridge University Press.

Saffran, J.R., Aslin, R.N. & Newport, E.L. 1996. Statistical learning by 8-Month-Old Infants. *Science, 274*, 1926-1928.

Xu, F. & Tenenbaum, J.B. 2007. Word learning as Bayesian inference. *Psychological Review, 114*(2), 245-272.