Bayesian inference as a cross-linguistic word segmentation strategy: Always learning useful things

Lawrence Phillips and Lisa Pearl
Department of Cognitive Sciences
University of California, Irvine
{lawphill, lpearl}@uci.edu

Abstract

Statistical learning has been proposed as one of the earliest strategies infants could use to segment words out of their native language because it does not rely on language-specific cues that must be derived from existing knowledge of the words in the language. Statistical word segmentation strategies using Bayesian inference have been shown to be quite successful for English (Goldwater et al. 2009), even when cognitively inspired processing constraints are integrated into the inference process (Pearl et al. 2011, Phillips & Pearl 2012). Here we test this kind of strategy on child-directed speech from seven languages to evaluate its effectiveness cross-linguistically, with the idea that a viable strategy should succeed in each case. We demonstrate that Bayesian inference is indeed a viable cross-linguistic strategy, provided the goal is to identify useful units of the language, which can range from sub-word morphology to whole words to meaningful word combinations.

1 Introduction

Word segmentation is one of the first tasks children must complete when learning their native language, and infants are able to identify words in fluent speech by around 7.5 months (Jusczyk & Aslin 1995; Echols et al. 1997; Jusczyk et al., 1993)). Proposals for learning strategies that can accomplish this (Saffran et al. 1996) have centered on language-independent cues that are not derived from existing knowledge of words. Bayesian inference is a statistical strategy operating over transitional probability that has been shown to be successful for identifying words in English, whether the salient perceptual units are phonemes (Goldwater et al. 2009 [**GGJ**], Pearl et al. 2011 [**PGS**]) or syllables (Phillips & Pearl 2012 [**P&P**]), and whether the inference process is optimal (GGJ, PGS) or constrained by cognitive limitations that children may share (PGS, P&P). It may, however, be the case that these strategies work well for English, but not other languages (Fourtassi et al. 2013). Therefore, we evaluate this same learning strategy on seven languages with different linguistic profiles: English, German, Spanish, Italian, Farsi, Hungarian, and Japanese. If Bayesian inference is a viable strategy for word segmentation, it should succeed on all languages. While some attempts have been made to evaluate Bayesian word segmentation strategies on languages other than English (e.g., Sesotho: Johnson 2008, Blanchard et al. 2010), this is the first evaluation on a significant range of languages that we are aware of.

We assume the relevant perceptual units are syllables, following previous modeling work (Swingly 2005, Gambell & Yang 2006, Lignos & Yang 2010, Phillips & Pearl 2012) that draws from experimental evidence that infants younger than 7.5 months are able to perceive syllables but not phonemes (Werker & Tees 1984, Juszyck & Derrah 1987, Eimas 1999). We demonstrate that Bayesian word segmentation is a successful cross-linguistic learning strategy, provided we define success in a more practical way than previous word segmentation studies have done. We consider a segmentation strategy successful if it identifies units useful for subsequent language acquisition processes (e.g., meaning learning, structure learning). Thus, not only is the orthographic gold standard typically used in word segmentation tasks acceptable, but also productive morphology and coherent chunks made up of multiple words. This serves as a general methodological contribution about the definition of segmentation success, especially when considering that the meaningful units across the world's languages may vary.

2 The Bayesian learning strategy

Bayesian models are well suited to questions of language acquisition because they distinguish between the learner's pre-existing beliefs (prior) and how the learner evaluates incoming data (likelihood), using Bayes' theorem:

$$P(h|d) \propto P(d|h)P(h)$$

The Bayesian learners we evaluate are the optimal learners of GGJ and the constrained learners of PGS. All learners are based on the same underlying models from GGJ. The first of these models assumes independence between words (a *unigram* assumption) while the second assumes that a word depends on the word before it (a *bigram* assumption). To encode these assumptions into the model, GGJ use a Dirichlet Process (Ferguson, 1973), which supposes that the observed sequence of words $w_1 \ldots w_n$ is generated sequentially using a probabilistic generative process. In the unigram case, the identity of the *i*th word is chosen according to:

$$P(w_i = w|w_1 \ldots w_{i-1}) = \frac{n_{i-1}(w)+\alpha P_0(w)}{i-1+\alpha} \quad (1)$$

where $n_{i-1}(w)$ is the number of times w appears in the previous $i - 1$ words, α is a free parameter of the model, and P_0 is a base distribution specifying the probability that a novel word will consist of the perceptual units $x_1 \ldots x_m$:

$$P(w = x_1 \ldots x_m) = \prod_{j=1}^{m} P(x_j) \quad (2)$$

In the bigram case, a hierarchical Dirichlet Process (Teh et al. 2006) is used. This model additionally tracks the frequencies of two-word sequences and is defined as:

$$P(w_i = w|w_{i-1} = w', w_1 \ldots w_{i-2}) = \frac{n_{i-1}(w',w)+\beta P_1(w)}{n_{i-1}(w')+\beta} \quad (3)$$

$$P_1(w_i = w) = \frac{b_{i-1}(w)+\gamma P_0(w)}{b_{i-1}+\gamma} \quad (4)$$

where $n_{i-1}(w',w)$ is the number of times the bigram (w',w) has occurred in the first $i - 1$ words, $b_{i-1}(w)$ is the number of times w has occurred as the second word of a bigram, b_{i-1} is the total number of bigrams, and β and γ are free model parameters.[1]

In both the unigram and bigram case, the model implicitly incorporates preferences for smaller lexicons by preferring words that appear frequently (due to (1) and (3)) and preferring shorter words in the lexicon (due to (2) and (4)).

The **BatchOpt** learner for this model is taken from GGJ and uses Gibbs sampling (Geman & Geman 1984) to run over the entire input in a single batch, sampling every potential word boundary 20,000 times. We consider this learner "optimal" in that it is unconstrained by cognitive considerations. We also evaluate the constrained learners developed by PGS that incorporate processing and memory constraints into the learning process.

The **OnlineOpt** learner incorporates a basic processing limitation: linguistic processing occurs online rather than in batch after a period of data collection. Thus, the OnlineOpt learner processes one utterance at a time, rather than processing the entire input at once. This learner uses the Viterbi algorithm to converge on the local optimal word segmentation for the current utterance, conditioned on all utterances seen so far.

The **OnlineSubOpt** learner is similar to the OnlineOpt learner in processing utterances incrementally, but is motivated by the idea that infants are not optimal decision-makers. Infants may not *always* select the best segmentation, and instead sample segmentations based on their perceived probabilities. The OnlineSubOpt learners will often choose the best segmentation but will occasionally choose less likely alternatives, based on the probability associated with each segmentation. The Forward algorithm is used to compute the likelihood of all possible segmentations and then a segmentation is chosen based on the resulting distribution.

The **OnlineMem** learner also processes data incrementally, but uses a Decayed Markov Chain Monte Carlo algorithm (Marthi et al. 2002) to implement a kind of limited short-term memory. This learner is similar to the original GGJ ideal (BatchOpt) learner in that it uses something like Gibbs sampling. However, the OnlineMem learner does not sample all potential boundaries; instead, it samples some number s of previous boundaries using the decay function b^{-d} to select the boundary to sample; b is the number of potential boundary locations between the boundary under consideration b_c and the end of

[1] Parameters for the unigram and bigram models underlying all learners were chosen to maximize the performance of the BatchOpt learner, discussed below. English: α=1, β=1, γ=90; German: α=1, β=1, γ=100; Spanish: α=1, β=200, γ=50; Italian: α=1, β=20, γ=200; Farsi: α=1, β=200, γ=500; Hungarian: α=1, β=300, γ=500; Japanese: α=1, β=300, γ=100

the current utterance while d is the decay rate. Thus, the further b_c is from the end of the current utterance, the less likely it is to be sampled. Larger values of d indicate a stricter memory constraint. All our results here use a set, non-optimized value for d of 1.5, which was chosen to implement a heavy memory constraint (e.g., 90% of samples come from the current utterance, while 96% are in the current or previous utterances). Having sampled a set of boundaries[2], the learner can then update its beliefs about those boundaries and subsequently update its lexicon.

3 Cross-linguistic input

We evaluate the Bayesian learner on input derived from child-directed speech corpora in seven languages: English, German, Spanish, Italian, Farsi, Hungarian and Japanese. All corpora were taken from the CHILDES database (MacWhinney, 2000). When corpora were available only in orthographic form, they were first converted into the appropriate phonemic form. Afterwards, the corpora were syllabified. Where possible, we utilized adult syllabification judgments. All other words were syllabified using the Maximum-Onset principle, which states that the beginning of a syllable should be as large as possible, without violating the language's phonotactic constraints.

Our corpora vary in a number of important ways. Although we attempt to limit our corpora to early child-directed speech, some of our corpora contain speech directed to children as old as age five (e.g. Farsi). Many of our corpora do, however, consist entirely of early child-directed speech (e.g., English, Japanese). Similarly, the same amount of data is not always easily available for each language. Our shortest corpus (German) consists of 9,378 utterances, while the longest (Farsi) consists of 31,657.

The languages themselves also contain many differences that potentially affect syllable-based word segmentation. While our English and Hungarian corpora contain 2,330 and 3,029 unique syllables, respectively, Japanese and Spanish contain only 526 and 524, respectively. Some languages may be easier to segment than others based on distributional factors. Fourtassi et al. (2013) show, for example, that English has less ambiguous segmentation than Japanese. In addition, the languages also have differences in their syntax and morphology. For example, Hungarian and Japanese are both agglutinative languages that have more regular morphological systems, while English, German, Spanish, Italian and Farsi are all fusional languages to varying degrees. If a language has regular morphology, an infant might reasonably segment out morphemes rather than words. This highlights the need for a more flexible metric of segmentation performance: A segmentation strategy which identifies units useful for later linguistic analysis should not be penalized.

4 Learning results & discussion

We analyze our results in terms of word token F-scores, which is the harmonic mean of token precision and recall, where precision is the probability that a word segmented by the model is a true word (# identified true / # identified) and recall measures the probability that any true word was correctly identified (# identified true / total # true). F-scores range from 0 to 100, with higher values indicating better performance. Performance on all languages is presented in Table 1. An error analysis was conducted where we systematically counted the following "reasonable errors" as successful segmentation:

(i) Mis-segmentations resulting in real words. For example, the word "alright" might be oversegmented as "all right", resulting in two actual English words. Most languages show errors of this type, with more occurring for the bigram model, with the least in English (BatchOpt: 4.52%) and most in Spanish (BatchOpt: 23.97%). We restrict these errors to words which occur minimally ten times in the corpus in order to avoid accepting errors in the corpora or nonsense syllables as real words.

(ii) Productive morphology. Given the syllabic nature of our corpora, only syllabic morphology can be identified. Languages like English, Spanish and Italian have relatively few errors that produce morphemes (e.g., BatchOpt: 0.13%, 0.05%, and 1.13% respectively), while Japanese, with more syllabic morphology has many such errors (e.g., BatchOpt: 4.69%).

[2] All OnlineMem learners sample s=20,000 boundaries per utterance. For a syllable-based learner, this works out to approximately 74% less processing than the BatchOpt learner (P&P).

		English	German	Spanish	Italian	Farsi	Hungarian	Japanese
Unigram	BatchOpt	55.70	73.43	64.28	70.48	72.48	64.01	69.11
	OnlineOpt	60.71	58.41	74.98	65.05	75.66	56.77	71.56
	OnlineSubOpt	65.76	70.95	77.15	66.48	74.89	60.21	71.73
	OnlineMem	58.68	73.85	67.78	66.77	67.31	60.07	70.49
Bigram	BatchOpt	80.19	84.15	80.34	79.36	76.01	70.87	73.11
	OnlineOpt	78.09	82.08	82.71	75.78	79.23	69.67	73.36
	OnlineSubOpt	80.44	82.03	80.75	73.59	67.54	65.48	66.14
	OnlineMem	89.58	88.83	83.27	74.08	73.98	69.48	73.24

Table 1. Token F-scores (presented as percents, from 0 to 100) for each learner across every language. Higher Token F-scores indicate better performance.

(iii) Common sequences of function words. For example, a learner might identify "is that a" as a single word, "isthata". These errors tend to be more common for unigram learners than bigram learners, which makes sense from a statistical standpoint since the unigram learner is unable to account for commonly occurring sequences of words and must do so by positing the collocation as a single word. Still, function word sequence errors are relatively uncommon in every language except German (e.g., BatchOpt: 21.73%)

Table 2 presents common examples of each type of acceptable error in English.

	True Word(s)	Model Output
Real words	something	some thing
	alright	all right
Morphology	going	go ing
	really	rea lly
Function word	you can	youcan
	are you	areyou

Table 2. Example reasonable errors of each type from English that result in real words, morphology, or function word collocations.

Generally speaking, the bigram learners tend to outperform the unigram learners, suggesting that the knowledge that words depend on previous words continues to be a useful one (as GGJ, PGS, and P&P found for English), though this difference may be small for some languages (e.g., Farsi, Japanese). Overall, performance for English and German is very high (best score: ~90%), while for other languages the learners tend to fare less well (best score: 70-83%), though still quite good. These results match previous work which indicated that English is particularly easy to segment compared to other languages (Johnson 2008; Blanchard et al. 2010; Fourtassi et al. 2013)

Importantly, the goal of early word segmentation is not for the infant to entirely solve word segmentation, but to get the word segmentation process started. Given this goal, Bayesian word segmentation seems effective for all these languages. Moreover, because our learners are looking for useful units, which can be realized in different ways across languages, they can identify foundational aspects of a language that are both smaller and larger than orthographic words.

5 Conclusion

We have demonstrated that Bayesian word segmentation performs quite well as an initial learning strategy for many different languages, so long as the learner is measured by how useful the units are that it identifies. This not only supports Bayesian word segmentation as a viable cross-linguistic strategy, but also suggests that a useful methodological norm for word segmentation research should be how well it identifies units that can scaffold future language acquisition. By taking into account reasonable errors that identify such units, we bring our model evaluation into alignment with the actual goal of early word segmentation.

References

Blanchard, D., Heinz, J., & Golinkoff, R. 2010. Modeling the contribution of phonotactic cues to the problem of word segmentation. *Journal of child language, 37*(3), 487.

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.

Fourtassi, A., Börschinger, B., Johnson, M., & Dupoux, E. 2013. Whyisenglishsoeasytosegment? *Proceedings of the Fourth Annual Workshop on Cognitive Modeling and Computational Linguistics*, 1-10.

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

Geman S. & Geman D. 1984. Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 6, 721-741.

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

Johnson, M. 2008. Unsupervised word segmentation for Sesotho using adaptor grammars. *Proceedings of the Tenth Meeting of the ACL Special Interest Group on Computational Morphology and Phonology,* 20-27.

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. 1993. Infants' preference for the predominant stress pattern of English words. *Child Development*, 64(3), 675-687.

Jusczyk, P.W. & Aslin, R.N. 1995. Infants' detection of the sound patterns of words in fluent speech. *Cognitive Psychology*, 29, 1-23.

Lignos, C. & Yang, C. 2010. Recession segmentation: Simpler online word segmentation using limited resources. *Proceedings of the Fourteenth Conference on Computational Natural Language Learning,* 88-97.

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

Marthi, B., Pasula, H., Russell, S. & Peres, Y., et al. 2002. Decayed MCMC filtering. In *Proceedings of 18th UAI,* 319-326.

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.

Phillips, L. & Pearl, L. 2012. "Less is more" in Bayesian word segmentation: When cognitively plausible learners outperform the ideal. In *Proceedings of the 34th Annual Conference of the Cognitive Science Society*.

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

Swingley, D. 2005. Statistical clustering and the contents of the infant vocabulary. *Cognitive Psychology, 50,* 86-132.

Teh, Y., Jordan, M., Beal, M., & Blei, D. 2006. Heirarchical Dirichlet processes. *Journal of the American Statistical Association, 101*(476), 1566-1581.

Werker, J.F. & Tees, R.C. 1984. Cross-language speech perception: Evidence for perceptual reorganization during the first year of life. *Infant Behavior & Development, 7,* 49-63.