Japanese word segmentation using Bayesian inference

Sarah Pieper
Computation of Language Laboratory
University of California, Irvine.

Abstract In studies of human cognition, Bayesian models are increasingly popular tools for understanding how children acquire language. Ideal learner Bayesian models investigate what learning strategies are optimal for solving the problem under consideration, while constrained learner models reduce memory and processing power in order to investigate what learning strategies might be optimal for humans. Both of these model types have been used to investigate an early problem in language acquisition: word segmentation. Use of constrained models for word segmentation with an English corpus shows a surprising "less is more" effect, where simulated learners with fewer memory and processing resources out-perform the ideal learners. However, it remains to be seen if this effect persists in other languages. This study examines the results of Bayesian modeling for word segmentation on a Japanese corpus. The results of the modeling suggest that a robust "less is more" effect is not present in Japanese, and thus does not hold across languages. Instead, the presence of this effect may depend on language-specific properties.

1. Introduction: Bayesian inference for word segmentation

Children are born without knowing any language; they must acquire it, learning words and their appropriate usage through exposure to language produced by others. Words spoken to infants by adults, with particular meaning and order, form a data set that children process to learn language. Bayesian inference, a sophisticated probabilistic reasoning procedure, has been proposed as a model of how children accomplish different tasks in language acquisition. This inference process has been used in models of word segmentation (Goldwater et al. 2009, Pearl et al. 2011), which is the task of determining where words are in fluent speech. In particular, it has been investigated as a learning strategy that would apply at the earliest stages of word segmentation, before infants know other useful language-dependent cues that indicate the beginnings and ends of words, such as stress cues. For example, stress patterns vary between English (stress-initial: ***LIS**/ten*) and French (stress-final: *écou/**TONS***), and can therefore reliably signal word beginnings or word ends in these languages. However, infants can only use this stress cue once they know what some of the words in the language are, so that they can tell whether they're learning an English-type or a French-type language. Statistical strategies of analysis such as Bayesian inference, on the other hand, are language-independent. No prior knowledge of words in the language is required, making the statistical strategies suitable to model the initial stages of word segmentation.

2. Previous studies of Bayesian word segmentation in English

2.1. Bayesian inference: A lexicon-building strategy

Bayesian inference is useful in word segmentation studies where a corpus (represented as a collection of utterances with no word boundaries in them, like "Ilikekitties") is processed by a model with no initial lexicon. The model segments the utterances into words (e.g., "I like kitties") , learning the words as it goes (e.g., "I", "like", "kitties"), so that it builds a vocabulary based only on words it encounters, as a human child would.

2.2. Algorithms for implementing Bayesian inference

Two types of algorithms that carry out Bayesian inference have been investigated for word segmentation—ideal learners and constrained learners. Ideal learners have unlimited processing and memory resources, and are often used to assess whether the data are sufficient for learning, and what learning biases might be useful. Constrained learners have limited processing and memory resources, and are often used to assess whether the data are sufficient for children, who have limited cognitive resources, to actually learn. Putting different restrictions on the models (for processing or memory, or both) can result in different hypotheses about what the correct segmentation of the data is, which can be compared to the true segmentation adults have for the data (this is the segmentation children are presumably trying to learn). Comparing the performance of these different algorithms can indicate how human infants, with their limited processing and memory resources, actually learn to segment words.

There are four algorithms implementing Bayesian inference that are used in this study. The ideal learning algorithm, first described in Goldwater et al. (2009) [**GGJ**] uses Gibbs sampling to choose the best segmentation for an utterance. Gibbs sampling is a type of Markov chain Monte Carlo procedure that is guaranteed to converge on the optimal segmentation, given the learning assumptions in the Bayesian model. It is in this sense that the learning algorithm is "ideal". This ideal learning algorithm operates by iterating over the entire corpus multiple times (for example GGJ's implementation iterated 20,000 times), which involves a large amount of processing and requires the child to hold the entire corpus in memory at all times. This ideal learning algorithm contrasts with the constrained learning algorithms discussed below, which were first implemented by Pearl et al. (2011).

The first of the constrained model types is Dynamic Programming with Maximization [**DPM**]. The DPM algorithm uses the Viterbi algorithm to process each utterance as a whole, computing the highest probability segmentation, given the current lexicon. Once the utterance is segmented, it is added to the lexicon, which affects decisions for subsequent segmentation. Thus, it differs from the ideal learning algorithm by processing utterances one at a time. This means it does not require as much memory as the ideal learner, as it does not need to keep the entire corpus in mind. However, it is similar in choosing the optimal segmentation at each point in time.

The Dynamic Programming with Sampling algorithm [**DPS**] relaxes this second behavior. Instead of choosing the optimal segmentation, it selects a segmentation probabilistically. Thus, low probability segmentations may be chosen some of the time, as it processes each utterance incrementally. Like the DPM algorithm, it processes utterances incrementally.

The Decayed Markov Chain Monte Carlo algorithm [**DMCMC**] also processes utterances incrementally, but uses a modified form of the Gibbs sampling procedure that the ideal learner uses. In particular, it uses only these utterances it has already encountered to inform segmentation decisions, rather than the entire corpus. This reduces the processing power required, as compared to the ideal algorithm. In addition, this algorithm implements a recency effect, where it tends to focus on utterances (and parts of utterances) it has encountered more recently. In this way, it is likely to require less memory than the ideal algorithm since it does not need to remember everything it has seen in precise detail.

2.3. Learner assumptions & previous results

Previous Bayesian learning studies have investigated the effect of certain kinds of knowledge about words (Goldwater et al. 2009, Pearl et al. 2011). The *unigram* learner assumes words are independent units, while the *bigram* learner thinks that words predict what words follow them. Previous studies have also assumed the basic unit of representation is the phoneme – that is, a learner gets a stream of phonemes as input and must put word boundaries in appropriately. Pearl et al. (2011) is one such study, and found a "less is more" effect for some of their unigram learners, where cognitive limitations help, rather than impede, language acquisition (Newport 1990). While this idea may seem counterintuitive, it is nonetheless true that children, who have greater cognitive limitations than adults, are in fact better at language acquisition than adults. The presence of a "less is more" effect in English when using a Bayesian inference learning strategy thus seems in line with what we would expect of children's learning strategies.

The "less is more" effect occurred in the Pearl et al. (2011) study because unigram learners assumed that commonly occurring sequences of words (e.g., "at the") were in fact one word (i.e., "atthe"), which is an undersegmentation error. Ideal learners with unrestricted memory tended to notice just how frequent these word sequences are and so undersegmented them. In contrast, the learners with limited memory were not always able to notice how frequent these sequences are, and so did not undersegment them nearly as often. In this way, restricted memory and processing resources produced better word segmentation results.

The assumption that the phoneme is the basic input unit for children is unlikely to be true, however. Evidence suggests that infants are aware of syllables before phonemes (Thiessen & Saffran 2003). Because of this, other studies (Phillips & Pearl 2012) have used the syllable as the basic unit for a Bayesian learning strategy. Again, a "less is more" effect was found for English data – in fact, a more robust effect where both unigram and bigram constrained learners outperformed the ideal learners.

The ideal unigram learners in the Phillips & Pearl (2012) study again made many more undersegmentation errors on frequent bigrams composed of short words such as "can you" and "do you" than did the constrained unigram learners. The ideal learners used the entire corpus to calculate the frequency of these bigrams and then uniformly undersegmented all instances. The constrained learners did not have the context of the entire corpus that led to these errors and segmented these bigrams correctly as they occurred. While constrained learners tended to make more oversegmentation errors, the ideal learners made a greater number of errors overall.

The bigram constrained learners in the study also out-performed the bigram ideal learners. The ideal learner correctly segmented 72.5% of the words in the corpus, accounting for 80% of total word types. The constrained learners segmented 85% of words in the corpus, but only 76.8% of the total word types. Phillips and Pearl (2012) interpreted this to mean that the constrained learners were more successful at segmenting more frequently occurring words than the ideal learners, leading to greater overall accuracy despite correctly segmenting fewer word types.

3. Cross-linguistic Bayesian segmentation

If we are interested in a successful language-independent word segmentation strategy, it is important to demonstrate that it works for multiple languages. Specifically, it must be shown to be successful on many different language types before generalizations about its utility for language acquisition can be made.

3.1. Japanese

Japanese is not only a different language from English, but belongs to a different language family and has several notable differences. First, there are fewer Japanese syllable types (4: V, CV, VC, CVC) than English syllable types (6: V, CV, VC, VCC, CVC, CVCC). Most Japanese syllables consist of either a single vowel (*a*), or an onset consonant and vowel (*ta*). Other grammatical syllables types allow for a nasal or geminate coda (*hon*, ***at/ta***). Japanese also has fewer phonemes (21) than English (40). In addition, Japanese has more standardized morphology, with very predictable verb conjugation. For example, there are very few verbs that conjugate irregularly, and Japanese verbs do not conjugate differently for different subjects, so that *I, he, she, they,* or *we run* all translate to *haSiru.* The root of the verb does not change with conjugation as some English verbs do; instead there is only a changed suffix to indicate case (haSi/ru→haSi/nai).

3.2. Japanese segmentation results and discussion

Table 1 shows the F-scores of word tokens as well as the precision and recall of word boundaries on a Japanese corpus of child-directed speech to children between the ages of 2 and 20 months that was derived from the CHILDES database. Tokens (T) refers to unique words, which may have multiple instances throughout the corpus. Word boundaries (B) refers to the edges of words (e.g. "at the" has four: at the beginning and end of both "at" and "the"). F-score (F) is the harmonic mean of precision (p) and recall (r):

$$\mathbf{F = 2pr/(p+r)}$$

where precision is the fraction of retrieved instances that are correct and recall is the fraction of correct instances that are retrieved. For example, a learner analyzing a corpus of 10 words that identifies a total of 9 words, and correctly identifies 6 of those will have a precision of 6/9 and a recall of 6/10. The F-score is one way to represent this information in a single number, with a higher number representing greater segmentation accuracy. F-score over tokens therefore gives one concise measure of segmentation performance. Word boundary precision and recall, when compared against each other, can indicate whether the simulated learner is undersegmenting (boundary precision > boundary recall) or oversegmenting (boundary precision < boundary recall).

Table 1: Token F-scores (TF) and word boundary precision (BP) and recall (BR) of different learners from the Japanese corpus

	TF	BP	BR
Unigram			
Ideal	65.372	77.912	73.212
DPM	64.758	71.358	78.718
DPS	64.564	70.046	82.182
DMCMC	64.318	74.914	75.344
Bigram			
GGJ-Ideal	65.468	69.56	86.412
DPM	59.772	63.002	88.946
DPS	56.246	59.808	91.338
DMCMC	50.296	54.246	96.32

The unigram learners, ideal and constrained, all performed roughly the same with regards to correctly identifying unique tokens, as seen by their F-scores. While the unigram ideal learner has the highest F-score (65.372), the learner with the lowest F-score, the DMCMC learner, was not much lower (64.318). When we examine the boundary precision and recall scores, only the ideal learner appears to be undersegmenting – all the constrained learners appear to be oversegmenting (though the DMCMC learner is doing so only very slightly).

While the unigram constrained learners had lower token F-scores than did the unigram ideal learner, the bigram constrained learners more significantly under-performed when compared to the bigram ideal learner. This was somewhat surprising, as the bigram learners did better than the unigram learners on the English corpora investigated in GGJ and Pearl et al (2011) (see Table 2). However, for Japanese, all the constrained bigram learners have F-scores at least five points lower than the ideal learner F-score of 65.468, and the lowest scorer, the DMCMC learner, has a score roughly 15 points lower (50.296). Notably, when we examine the boundary precision and recall, *all* learners (both ideal and constrained) are oversegmenting, with the DMCMC learner doing this the most – the difference between its boundary precision and recall is over 40 points (precision: 54.246, recall: 96.32).

Table 2: Token F-scores (TF) and word boundary precision (BP) and recall (BR) of different learners from Pearl-Brent derived English corpus

	TF	BP	BR
Unigram			

Ideal	53.124	92.0	62.1
DPM	58.762	66.7	88.5
DPS	63.682	60.9	89.5
DMCMC	55.116	86.3	74.5
Bigram			
GGJ-Ideal	77.062	85.6	82.0
DPM	75.076	75.2	89.6
DPS	77.768	52.8	90.5
DMCMC	86.264	81.1	87.6

As seen in Table 2, the constrained learners, both unigram and bigram, performed better than the ideal learners when segmenting words in the English corpus. Importantly, the improved performance of the various constrained learners using Bayesian inference compared to the ideal learners suggests a "less is more" effect that increases segmentation accuracy. This is at odds with the data obtained from the Japanese corpus, shown in Table 1. In particular, the ideal learner always does the best, presumably because it does not make as many oversegmentation errors.This suggests that Bayesian inference does not automatically produce a "less is more" effect across all languages, where constrained learners outperform ideal learners.

We turn now to the specific kinds of errors the simulated learners made in Japanese. As mentioned above, there were many more oversegmentation errors, as indicated by the boundary precision and recall scores. Table 3 shows a detailed description of these kinds of errors, dividing them by linguistic category (noun, verb, adjective, nonsense word, expression).

Table 3: Types and tokens (# of unique tokens / # of instances) of oversegmentation errors made by different learners from Japanese corpus

Unigram	**Noun** kore→ ko re	**Verb** tabete→tabe te	**Nonsense** bUbU→bU bU	**Expression** dOzo→dO zo	**Adjective** GOzu→GO zu
GGJ-Ideal	9 / 58	11 / 81	2 / 11	0 / 0	3 / 21
DPM	9 / 186	9 / 171	2 / 31	4 / 103	5 / 93
DPS	12 / 314	6 / 125	6 / 104	3 / 68	3 / 42
DMCMC	7 / 73	9 / 98	5 / 56	2 / 18	4 / 66
Bigram					
GGJ-Ideal	8 / 415	12 / 427	3 / 82	0 / 0	3 / 61
DPM	8 / 508	9 / 277	3 / 59	0 / 0	3 / 127
DPS	10 / 480	7 / 274	2 / 50	1 / 20	4 / 127
DMCMC	10 / 747	12 / 585	2 / 64	4 / 179	2 / 86

The pattern of oversegmentation errors is consistent across different learners, with the main difference being the number of mistakes. For example, most of these oversegmented words, shown in Table 3 above, had two syllables, with only occasional errors for words with three syllables. Oversegmented verbs were segmented between the root and conjugated morphemes (*tabete* → *tabe te*). This occurred exclusively with short verbs of two to three syllables; verbs with longer conjugated endings were correctly segmented. This is likely due to the increased predictability of these longer morphemes, as opposed to the less predictable single-syllable morphemes.

Table 4: Types and tokens (# of unique tokens / # of instances) of undersegmentation errors made by different learners from Japanese corpus

Unigram	**Noun** hai kore→ haikore	**Verb** koko irete→ kokoirete	**Nonsense** A A→ AA	**Particle** kore wa→ korewa	**Repetition** takai takai→ takaitakai
GGJ-Ideal	5 / 94	3 / 59	2 / 60	19 / 444	2 / 32
DPM	3 / 37	8 / 82	0 / 0	15 / 163	5 / 70
DPS	1 / 9	1 / 12	0 / 0	18 / 231	7 / 118
DMCMC	1 / 15	2 / 36	1 / 68	21 / 420	2 / 32
Bigram					
GGJ-Ideal	2 / 27	0 / 0	3 / 146	17 / 152	5 / 24
DPM	1 / 8	3/ 36	1 / 8	16 / 244	8 / 102
DPS	2 / 13	4 / 57	1 /8	11 / 150	11 / 95
DMCMC	4 / 6	2 / 13	0 / 0	13 / 53	11 / 69

While the most common oversegmentation errors were made with verbs and nouns, as seen in Table 4 above, the most common undersegmentation errors were made for phrases with particles (*kore wa* → *korewa*) and repetitive words (*takai takai* → *takaitakai*). Since particles are frequently occurring parts of speech in Japanese, they comprised both the greatest number of error types and the greatest total number of errors for unigram learners. Similarly to the previously discussed English undersegmentation of frequently occurring bigrams (e.g. "at the"), the learners segment the Japanese phrase as if it were one word. The unigram learners, which do not treat previously occurring words as predictive, are more prone to this mistake. The unigram learners ranged from 15 (DPM) to 21 (DMCMC) errors segmenting unique tokens, with a range of 163 (DPM) to 444 (Ideal) total error instances. The bigram learners had a lower range of 11 (DPS) to 17 (Ideal) unique token errors, and a range of only 53 (DMCMC) to 244 (DPM) total error instances.

4. Conclusion and Future Research

It remains to be seen how well learner strategies using Bayesian inference do for other languages besides Japanese and English. It may only produce this "less is more" effect for languages with specific properties. Japanese differs from English in a number of ways – it is currently unclear which properties are responsible for the difference in behavior, though the predicatable morphology seems to have caused more oversegmentation errors. Comparison to other languages may suggest which properties are required to produce a robust "less is more" effect when modeling word segmentation.

References

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

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.

Phillips, L. & Pearl, L. (2012). ‘Less is More’ in Bayesian word segmentation: When cognitively plausible learners outperform the ideal, In N. Miyake, D. Peebles, & R. Cooper (eds), *Proceedings of the 34th Annual Conference of the Cognitive Science Society*. Austin, TX: Cognitive Science Society.

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.