Modeling Representational Constraints in Word Segmentation

Jordan Kodner

Word segmentation is an important and non-trivial part of early child language acquisition. The experimental literature has proposed a wide range of input cues that make the process possible, and variety of computational models have been implemented to make use of them. Most reasonable models perform well on English, but performance varies widely on other languages, notably struggling with Japanese (Fourtassi et al., 2013; Ludusan et al., 2017). For example, Boruta et al. (2011) finds that the NGS-u model (Venkataraman, 2001) achieves an F-score of 0.69 on English, 0.54 on French, and only 0.41 on Japanese (MBDP-1 (Brent, 1999) achieves the same results on Japanese). This is not just an artifact of the model. For comparison, more sophisticated Adaptor Grammar (Johnson and Goldwater, 2009) models achieves at best 0.89 on English but still only manages 0.70 on Japanese, subject to language-specific configuration (Fourtassi et al., 2013). This is primarily due to a strong tendency for models to over-segment Japanese text.

It is clear why this is, but it is less clear how to solve the problem. Segmentation is harder for languages with fewer phone and syllable types because words become more likely to contain other valid words as substrings, and this leads algorithms to over-segment. The same tendency for over-segmentation arises in languages with longer average word lengths both because of the subset problem and also because fewer of the potential segmentation points are true segmentation points. Models perform well on English child-directed speech (CDS) because it happens to have short words (just over one syllable on average), as well as high syllable diversity. Japanese CDS, on the other hand, poses a challenge because it has longer words on average and an order of magnitude fewer syllable types.

Clearly, Japanese-learning children are just as capable at segmenting input utterances as English-learning children are, so they must be using some additional cues to get around these problems. To the extent that the processes of language acquisition are universal, these cues should be available to English-learning children as well, so a major challenge for children working on word segmentation is the step of determining which of the available cues are the most useful for their own languages. We explore the impact that two of these potential sources of

* Jordan Kodner, University of Pennsylvania, jkodner@sas.upenn.edu. We thank Charles Yang for his helpful comments and Constantine Lignos for his helpful code. This work was funded by an NDSEG fellowship awarded to the author by the U.S. Department of Defense.

 Proceedings of the 42nd annual Boston University Conference on Language Development, ed. Anne B. Bertolini and Maxwell J. Kaplan, 391-399. Somerville, MA: Cascadilla Press.

information, word minimality constraints and prosodic cues, have on segmentation by augmenting a subtractive computational model of word segmentation on Japanese with English and French for comparison. The results indicate that these additional cues have no substantial impact on already high-performing English segmentation, harm French performance as predicted by French phonology, and yield substantial improvements for Japanese when implemented in tandem. The following two sections review a range of segmentation cues proposed in the literature followed by a selection of previous computational segmentation models, Section 4 describes this work's methodology, Section 5 lays out the results, and Section 6 discusses the implications of these findings.

1. Cues for Word Segmentation

There must be something about child-directed utterances that provides children with the information they need to carry out word segmentation. A huge range of potentially informative input cues have been proposed in the literature, and in practice, most of them are probably somehow useful. The most famous of these are the *transitional probabilities* between phones or syllables. The concept of phoneme transitional probabilities was first suggested in Harris (1955) for the segmentation of morphemes, and more recently, numerous artificial language learning experiments have demonstrated that children are sensitive to syllable transitional probabilities when segmenting strings of nonce words (Saffran et al., 1996; Thiessen and Saffran, 2003, etc.).

However, both experimental and computational studies show the limits of transitional probabilities. For example, Johnson and Tyler (2010) find that while Dutch children and adults are adept at segmenting nonce words of a fixed length with transitional probabilities, both age groups fail to segment nonce words of more naturalistic variable length. And when Yang (2004) calculates the syllable transitional probabilities of running text from a corpus, he find that segmenting words directly according those values would yield a segmentation accuracy score below chance.

A wide range of other sources of evidence have been presented either as alternatives or complements to transitional probabilities. One such cue is the presence of single-word utterances. Words heard in isolation in caregivers' speech tend to be produced with high relative frequency among children's early productions (Brent and Siskind, 2001), indicating that children are attuned to them. Single-word utterances are useful because they provide ready-made phonetic boundaries for words, and according to Seidl and Johnson (2006)'s Edge Hypothesis, prosodic boundaries provide the same kind of evidence. They and Shukla et al. (2011) both show that children segment words with higher accuracy when their stresses are aligned with prosodic boundaries than otherwise. But single-word phrases cannot solve segmentation on their own. Putting aside the question of how children recognize which utterances are single-word in the first place, there are simply not

enough of them in CDS to provide all the necessary evidence. That is not to say, however, that they are not useful in conjunction with other cues.

A whole set of additional cues can be thought of as sensitivities to phonological well-formedness. Hard constraints limit the phonological shape of words, thus rendering some incorrect segmentations impossible. For example, every word must contain at least one vowel or sonorant in most languages, so segmentations which create words with no vowel or other valid syllable nucleus can be dismissed (Brent and Cartwright, 1996; Johnson, 2003). A softer version of this constraint might be one that limits words to at least two morae. This pattern of word minimality has been observed as part of English phonology (why words like /kiː/ and /kɪt/ are valid, but /kɪ/ is not), and to a lesser extent, a tendency for Japanese. The constraint must be soft however, because there are notable violations in both languages including *the* and *a* in English, and the many final particles in Japanese. Similar kinds of minimality constraints exist in many other languages, but French learners are notable for violating them (Demuth and Johnson, 2003).

2. Computational Models

Transitional probabilities have been at the forefront of thought on word segmentation for decades, and transitional probabilities (on phones rather than syllables) have served as the foundation for a series of Bayesian segmentation models (Johnson and Goldwater, 2009; Goldwater et al., 2009; Boruta et al., 2011; Venkataraman, 2001, etc.). Bayesian inference presents a sophisticated way to make sense of patterns in the transitional probabilities of entire corpora in order to uncover how an ideal learner might optimally use them. The most successful of these models induce additional structure during the inference process. Goldwater et al. (2009) compare their performance on English over unigrams and bigrams, showing favorable performance for bigrams, and Johnson and Goldwater (2009)'s Adaptor Grammars are multilevel models can induce syllable-like units from phones outperform bigram models but must be configured according to a given language's phonotactics.

Phone-based models have proven popular, but they are both practically and empirically problematic. Empirically, children show sensitivity to syllable phonotactic constraints early on (Onishi et al., 2002), and word segmentation errors have been shown to align strongly with syllable boundaries as opposed to arbitrary phone boundaries (Peters, 1983), implying that syllables are the unit of segmentation. Practically, syllable-based segmentation is an easier problem, both for the researcher and for the child, since there are fewer syllable boundaries and by extension fewer candidate segmentation points per utterance. If syllable-based segmentation is more cognitively plausible, than phone-based models are making the task unnecessarily difficult.

The main alternative to optimization-based Bayesian models are algorithmic *subtractive* or *incremental* models. These largely eschew transitional probabilities and instead focus on incrementally removing or "subtracting" chunks of known

vocabulary from unsegmented utterances. The remaining pieces are added to a tentative lexicon and further chunked as needed later. Single-word automatically play an important role in such models because they can be learned correctly as soon as they are added to the lexicon. The subtraction can be approached multi-directionally (Monaghan and Christiansen, 2010), but these algorithms are also amenable to online implementation and thus easily model the course of child development as more and more utterances are presented. If they process utterances only once each from left-to-right, then they are processed in the same way that they are heard (Lignos, 2012; Gambell and Yang, 2004; Lignos, 2011).

3. Methodology

We investigate what impact word minimality constraints and prosodic cues have cross-linguistically by augmenting the Lignos (2012) online syllable-based subtractive model. Three CDS corpora were studied in order to facilitate direct comparison across languages to the extent possible: Brent (Brent and Siskind, 2001) for English, York (Plunkett and De Cat, 2001) for French, and Hamasaki (Oshima-Takane et al., 1995) for Japanese. The English corpus was converted into phonemes according to the CMU Pronouncing Dictionary (Weide, 1998) and then mechanically syllabified by applying the Maximum Onset Principle following Lignos (2012), and the French corpus was transcribed and syllabified according to the Lexique dictionary (New et al., 2004). The Japanese corpus is presented in romaji, so transcription and syllabification were possible without a dictionary. Because of the privileged position that morae have in Japanese, a second version of that corpus was prepared by segmenting utterances into single-mora units rather than syllables. The base Lignos algorithm (*Subtractive*) was augmented to produce three additional models: a version which obeys word minimality constraints (*Subtractive+WM*), a version augmented to respond to simulated prosodic cues (*Subtractive+PB*), and a version that obeys both (*Subtractive+PB+WM*).

Subtractive is a straightforward online left-to-right implementation, and like all such models, it is expected to perform well on English and poorly on Japanese. A soft word minimality constraint was added to *Subtractive+WM* by replacing the algorithm's uniform reward with one that down-weights atypically sized words according to $1/(|\texttt{seg\_len} - \texttt{mean\_word\_wlen}| + \epsilon)$. This punishes unusually long words as well, but since the model tends to over-segment, this has little impact on the results. A soft constraint was chosen because Japanese and English occasionally violate word minimality.

Annotated prosodic information was unavailable to us, but Japanese prosody lends itself well to automatic annotation. It has been noted that Japanese has *edge-prominent* prosody in which prosodic peaks correspond with phrase boundaries (Mazuka, 2010), and since Japanese syntax is strongly head-final, it is possible to identify many phrase boundaries automatically. In order to mark prosodic peaks in Japanese, the corpus was searched for phrase-final particles (*wa*, *ga*, *wo*, *ni*, etc.), postpositions, and verb inflectional suffixes, and prosodic peaks were annotated

at those points. An equivalent treatment was not possible for the other languages under study.

Subtractive+PB uses the resulting prosodic peaks/phrase boundaries to pre-segment the text in line with the Edge Hypothesis. These segmentation points are guaranteed to be correct, so each of the phrases can be treated as independent utterances. This approach of using phrase boundaries to pre-segment utterances was also taken in Ludusan et al. (2017), which tested the impact that prosodic cues had when segmenting the Japanese RIKEN corpus (Mazuka et al., 2006). A more novel approach is taken for *Subtractive+PB+WM*. Instead of just using prosody to indicate where segmentation should occur, it dictates how seriously the model should take other cues by enforcing a hard word minimality constraint in general but relaxing it exactly at phrase boundaries. This was inspired by the fact that word minimality-violating single-mora words are more likely to occur at these boundaries in Japanese (phrase-final particles and some postpositions), and in English (*the* and *a*).

A number of other models were tested for comparison. First, syllable (*Unit (Syll)*) and phone (*Unit (Phone)*) baselines were defined which inserted word boundaries after every segmentation unit. These yield 100% recall but low precision. Next, four phone-based models were compared: the unigram and bigram models (*G&J (Unigram)* and *G&J (Bigram)*) from Goldwater et al. (2009), the *NGS-u* model of Venkataraman (2001), and three adaptor grammar models (*Best AG*). The best AG model for English (Johnson and Goldwater, 2009) and Japanese (Fourtassi et al., 2013) are reported, and results for French were calculated with the same $\texttt{colloc3}_{S}\texttt{yllableIF}$ configuration from the original paper. The phone and syllable-based models are not directly comparable because they differ in the number of possible segmentation points, but they do serve to demonstrate how much simpler syllable-based segmentation is.

4. Results

Table 1 presents results for English and French. The English data contains 3,466 unique syllables and has an average gold-standard word length of 1.20 syllables, while the French data has only 1,329 unique syllables types and an almost identical average word length of 1.21. As expected, performance on English is very good. Since average word length is just over one, the syllable baseline alone achieves an F-score of over 0.87. There is little room for improvement here, so base *Subtractive* only achieves a few points improvement over unit segmentation, and the inclusion of word minimality has almost no impact. Of the phone-based models, the *AG* configured to construct syllable-like units according to English phonotactics does best, but it only slightly outperforms the syllable baseline.

The French results are more interesting. As pointed out by Boruta et al. (2011), *NGS-u* performs significantly worse on French than for English, and a similar pattern arises with the *AG* model. However, the simpler *G&J* models perform comparably with English, and the baselines perform even better. This suggests

that the previously reported performance difference between English and French is primarily an artifact of the specfic models tested. Most interestingly, *Subtractive+WM* yields a three-point decrease in French performance. This is reassuring because the language lack such a constraint. Attending to non-existent patterns should hurt performance.

Table 1: **English and French results.**

	English			**French**		
Model	Prec	Rec	F1	Prec	Rec	F1
Unit (Phone)	0.261	1.000	0.414	0.335	1.000	0.502
G&J (Uni)	0.619	0.476	0.538	0.603	0.474	0.531
G&J (Bi)	0.752	0.686	0.638	0.669	0.638	0.653
NGS-u	-	-	0.69-	-	-	0.54-
Best AG	-	-	**0.89-**	0.776	0.768	**0.772**
Unit (Syll)	0.776	1.000	0.874	0.790	1.000	0.883
Subtractive	0.837	0.996	**0.910**	0.805	0.997	**0.891**
Subtr+WM	0.841	0.992	**0.910**	0.842	0.883	0.862

The syllabified Japanese corpus contains 581 unique syllable types and an average word length of 2.06, and the "morafied" corpus contains only 168 unique mora types and an average word length of 2.46. Unsurprisingly given these numbers, all models reported in Table 2 perform poorly on Japanese relative to English and French.[1] Once again, the syllable baseline outperforms all phone models except for *Best AG*, and the mora baseline is about 8 points behind. *Subtractive* improves on the syllable baseline, and *Subtractive+WM* improves on *Subtractive* for both syllable and mora segmentations, but forcing phrase boundaries segmentations in *Subtractive+PB* yields almost no additional improvement. The only model that solidly outperforms *Best AG* model is *Subtractive+PB+WM* on syllables, which achieves a 19 point improvement over *Subtractive*, while *Subtractive+PB+WM* based on mora units is within rounding error of *Best AG*.

[1] Phone-based results are only reported once because they do not attend to the difference between syllable and mora units.

Table 2: **Japanese results.**

	Japanese (Syllables)			Japanese (Morae)		
Model	Prec	Rec	F-Score	Prec	Rec	F-Score
Unit (Phones)	0.173	1.000	0.295			
G&J (Unigrams)	0.500	0.576	0.535			
G&J (Bigrams)	0.447	0.587	0.508			
NGS-u	-	-	0.41-			
Best AG	0.70-	0.70-	**0.70-**			
Unit (Syll/Mora)	0.386	1.000	0.557	0.315	1.000	0.479
Subtractive	0.404	0.995	0.575	0.327	0.995	0.492
Subtr+WM	0.421	0.988	0.591	0.3450	0.984	0.516
Subtr+PB	0.405	0.995	0.576	0.326	0.995	0.492
Subtr+PB+WM	0.699	0.840	**0.763**	0.583	0.863	**0.696**

5. Discussion

By demonstrating the varying potential impact of word minimality on English, French, and Japanese, together with prosodic cues on Japanese, this work suggests a scenario where young language learners are sensitive to many cues early on and select which of them provide the best information for segmenting their languages. As the results on French show, giving attention to a constraint that the language lacks harms performance even though it helps on other languages, so simply folding in all possible sources of information is likely not the best option.

Similarly, the Japanese model which takes advantage of the language-specific alignment between single-mora words and prosodic peaks greatly improves results over models which do not take both into account together. This implies that early attention to language-specific phonological and phonetic patterns may be critical to segmentation in languages for which straightforward syllable unit segmentation does not yield a strong baseline. This model is language-specific, but that it not unique. It should be noted that the best performing AG model is also language-specific in that it requires information about the language's phonotactics in order to achieve its reported performance. In that sense, the AG results also suggest that language-specific attention is a worthwhile approach.

Future research will be improved by using the manual prosodic annotations from the RIKEN corpus rather than automatic annotations. The automatic scheme utilized here is necessarily an underestimation of the number of prosodic peaks in the language, which likely explains why pre-segmentation at phrase boundaries had so little effect here as compared to Ludusan et al. (2017). Nevertheless, this work represents a solid first step. The use of prosodic cues to not only indicate segmentation points directly but also influence the application of heuristic constraints like word minimality represents a path for more research. This work also highlights the practical benefits of syllable-based segmentation, something that has been argued for elsewhere (Lignos, 2011). Given that the notion has some psychological backing and that even the no-effort syllable unit baselines often

outperformed much more sophisticated models, syllable-based segmentation may be the most productive way to move forward.

References

Boruta, Luc, Sharon Peperkamp, Benoît Crabbé, and Emmanuel Dupoux. 2011. Testing the robustness of online word segmentation: Effects of linguistic diversity and phonetic variation. In *Proceedings of the 2nd Workshop on Cognitive Modeling and Computational Linguistics*, 1–9. Association for Computational Linguistics.

Brent, Michael R. 1999. An efficient, probabilistically sound algorithm for segmentation and word discovery. *Machine Learning* 34:71–105.

Brent, Michael R, and Timothy A Cartwright. 1996. Distributional regularity and phonotactic constraints are useful for segmentation. *Cognition* 61:93–125.

Brent, Michael R, and Jeffrey Mark Siskind. 2001. The role of exposure to isolated words in early vocabulary development. *Cognition* 81:B33–B44.

Demuth, Katherine, and Mark Johnson. 2003. Truncation to subminimal words in early french. *Canadian Journal of Linguistics/Revue canadienne de linguistique* 48:211–241.

Fourtassi, Abdellah, Benjamin Börschinger, Mark Johnson, and Emmanuel Dupoux. 2013. Whyisenglishsoeasytosegment. In *Proceedings of the Fourth Annual Workshop on Cognitive Modeling and Computational Linguistics*, 1–10.

Gambell, Timothy, and Charles Yang. 2004. Statistics learning and universal grammar: Modeling word segmentation. In *First Workshop on Psycho-computational Models of Human Language Acquisition*, 49.

Goldwater, Sharon, Thomas L Griffiths, and Mark Johnson. 2009. A bayesian framework for word segmentation: Exploring the effects of context. *Cognition* 112:21–54.

Harris, Zellig S. 1955. From phoneme to morpheme. *Language* 31:190–222.

Johnson, Elizabeth K. 2003. Word segmentation during infancy: The role of subphonemic cues to word boundaries. *Unpublished doctoral dissertation, The Johns Hopkins University*.

Johnson, Elizabeth K, and Michael D Tyler. 2010. Testing the limits of statistical learning for word segmentation. *Developmental science* 13:339–345.

Johnson, Mark, and Sharon Goldwater. 2009. Improving nonparameteric bayesian inference: experiments on unsupervised word segmentation with adaptor grammars. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 317–325. Association for Computational Linguistics.

Lignos, Constantine. 2011. Modeling infant word segmentation. In *Proceedings of the fifteenth conference on computational natural language learning*, 29–38. Association for Computational Linguistics.

Lignos, Constantine. 2012. Infant word segmentation: An incremental, integrated model. In *Proceedings of the West Coast Conference on Formal Linguistics*, volume 30, 13–15.

Ludusan, Bogdan, Reiko Mazuka, Mathieu Bernard, Alejandrina Cristia, and Emmanuel Dupoux. 2017. The role of prosody and speech register in word segmentation: A computational modelling perspective. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, 178–183.

Mazuka, Reiko. 2010. Learning the sound system of japanese: What does it tell us about language acquisition? In *Proceedings of the International Congress on Acoustics*, volume 20, 4186–4193.

Mazuka, Reiko, Yosuke Igarashi, and Ken'ya Nishikawa. 2006. Input for learning japanese: Riken japanese mother-infant conversation corpus (coe workshop session 2). *ŁŁ. TL, ł* 106:11–15.

Monaghan, Padraic, and Morten H Christiansen. 2010. Words in puddles of sound: Modelling psycholinguistic effects in speech segmentation. *Journal of child language* 37:545–564.

New, Boris, Christophe Pallier, Marc Brysbaert, and Ludovic Ferrand. 2004. Lexique 2: A new french lexical database. *Behavior Research Methods, Instruments, & Computers* 36:516–524.

Onishi, Kristine H, Kyle E Chambers, and Cynthia Fisher. 2002. Learning phonotactic constraints from brief auditory experience. *Cognition* 83:B13–B23.

Oshima-Takane, Yuriko, Brian MacWhinney, Hidetoshi Shirai, Susanne Miyata, and Norio Naka. 1995. Childes manual for japanese. *Montreal: McGill University*.

Peters, Ann M. 1983. *The units of language acquisition*, volume 1. CUP Archive.

Plunkett, Bernadette, and Cécile De Cat. 2001. Root specifiers and null subjects revisited. In *Proceedings of the 25th Annual Boston University Conference on Language Development*, 611–622.

Saffran, Jenny R, Richard N Aslin, and Elissa L Newport. 1996. Statistical learning by 8-month-old infants. *Science* 1926–1928.

Seidl, Amanda, and Elizabeth K Johnson. 2006. Infant word segmentation revisited: Edge alignment facilitates target extraction. *Developmental science* 9:565–573.

Shukla, Mohinish, Katherine S. White, and Richard N. Aslin. 2011. Prosody guides the rapid mapping of auditory word forms onto visual objects in 6-mo-old infants. *Proceedings of the National Academy of the Sciences* 108:6038–6043.

Thiessen, Erik D, and Jenny R Saffran. 2003. When cues collide: use of stress and statistical cues to word boundaries by 7-to 9-month-old infants. *Developmental psychology* 39:706.

Venkataraman, Anand. 2001. A statistical model for word discovery in transcribed speech. *Computational Linguistics* 27:351–372.

Weide, Robert. 1998. The cmu pronunciation dictionary, release 0.6. *Carnegie Mellon University*.

Yang, Charles D. 2004. Universal grammar, statistics or both? *Trends in cognitive sciences* 8:451–456.

Proceedings of the 42nd annual Boston University Conference on Language Development

edited by Anne B. Bertolini
and Maxwell J. Kaplan

Cascadilla Press Somerville, MA 2018

ISSN 1080-692X
ISBN 978-1-57473-086-9 (2 volume set, paperback)
ISBN 978-1-57473-186-6 (2 volume set, library binding)

Ordering information

To order a copy of the proceedings or to place a standing order, contact:

Cascadilla Press, P.O. Box 440355, Somerville, MA 02144, USA
phone: 1-617-776-2370, sales@cascadilla.com, www.cascadilla.com