The Role of Prosody and Speech Register in Word Segmentation: A Computational Modelling Perspective

Bogdan Ludusan
LSCP
EHESS/ENS/PSL/CNRS
29 rue d'Ulm, 75005 Paris, France
bogdan.ludusan@ens.fr

Reiko Mazuka
Language Development Lab
RIKEN Brain Science Institute
2-1 Hirosawa, Wako, 351-0198, Japan
mazuka@brain.riken.jp

Mathieu Bernard
Alejandrina Cristia
Emmanuel Dupoux
LSCP - EHESS/ENS/PSL/CNRS
29 rue d'Ulm, 75005 Paris, France
{mmathieubernardd, alecristia,
emmanuel.dupoux}@gmail.com

Abstract

This study explores the role of speech register and prosody for the task of word segmentation. Since these two factors are thought to play an important role in early language acquisition, we aim to quantify their contribution for this task. We study a Japanese corpus containing both infant- and adult-directed speech and we apply four different word segmentation models, with and without knowledge of prosodic boundaries. The results showed that the difference between registers is smaller than previously reported and that prosodic boundary information helps more adult- than infant-directed speech.

1 Introduction

Infants start learning their native language even before birth and, already during their first year of life, they succeed in acquiring linguistic structure at several levels, including phonetic and lexical knowledge. One extraordinary aspect of the learning process is infants' ability to segment continuous speech into words, while having little or no knowledge of the sounds of their native language.

Several hypotheses have been proposed in the experimental literature to explain how they achieve this feat. Among the main classes of cues put forward, prosodic cues (e.g. stress, prosodic boundaries) have been shown to be particularly useful in early-stage word segmentation (Christophe et al., 2003; Curtin et al., 2005; Seidl and Johnson, 2006). Previous work suggests that these cues may be emphasized in the speech register often used when addressing infants (infant-directed speech; IDS). This register is characterized by shorter utterances, repeated words and exaggerated prosody (see (Cristia, 2013) for a review). It has been shown that IDS can facilitate segmentation performance in infants (Thiessen et al., 2005), when compared to the register that parents use when talking to adults (adult-directed speech; ADS).

The process of word segmentation has received considerable attention also from the computational linguistics community, where various computational models have been proposed (e.g. (Brent and Cartwright, 1996; Goldwater et al., 2009)). Yet, despite the role that prosodic cues play in early word segmentation, only lexical stress has been addressed in detail, in the computational modelling literature (e.g. (Börschinger and Johnson, 2014; Doyle and Levy, 2013; Lignos, 2011)). As for prosodic boundary information, it was investigated in only one previous study (Ludusan et al., 2015). That study found that that an Adaptor Grammar model (Johnson et al., 2007) performed better on both English and Japanese corpora when prosodic boundary information was added to its grammar. These previous studies investigated the effect of prosodic cues while keeping register constant, investigating either IDS (e.g. (Börschinger and Johnson, 2014)) or ADS (Ludusan et al., 2015). Other work focuses on register only. For instance, (Fourtassi et al., 2013) used the Adaptor Grammar framework to examine English and Japanese corpora of infant- and adult-directed speech, concluding that IDS was easier to segment

Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Short Papers), pages 178–183
Vancouver, Canada, July 30 - August 4, 2017. ©2017 Association for Computational Linguistics
https://doi.org/10.18653/v1/P17-2028

than ADS. However, the corpora were not parallel or necessarily directly comparable, as, the ADS in Japanese was transcribed from academic presentation speeches, whereas the IDS came from spontaneous conversational speech.

We aim to put together these two lines of research, by conducting the first computational study of word segmentation that takes into account both variables: speech register and prosodic boundary information. This investigation extends the previously mentioned studies, by allowing us to observe not only the effect of each individual variable, but also any interaction between the two. More importantly, it is performed in a more controlled manner as it makes use of a large corpus of spontaneous verbal interactions, containing both IDS and ADS uttered by the same speakers. Furthermore, we do not limit ourselves to a specific model, but test several, different, unsupervised segmentation models in order to increase the generalizability of the findings.

2 Methods

Several unsupervised segmentation algorithms were employed. We selected 2 sub-lexical and 2 lexical models, all of which are made freely available through the CDSwordSeg package[1].

The first model performs transition-probability-based segmentation (TP) employing the *relative algorithm* of Saksida et al. (2016). It takes in input transcribed utterances, segmented at the syllable level and computes the forward transitional probabilities between every pair of syllables in the corpus. The transition probability between two syllables X and Y is defined as the frequency of the pair (X,Y) divided by the frequency of the syllable X. Once probabilities are computed, word boundaries are posited using local minima of the probability function. As this algorithm only attempts to posit boundaries based on phonological information it is called a 'sub-lexical' model.

Diphone-based segmentation (DiBS) is another sub-lexical model, which uses diphones instead of syllables pairs (Daland and Pierrehumbert, 2011). The input is represented as a sequence of phonemes and the model tries to place boundaries based on the identity of each consecutive sequence of two phonemes. The goal is accomplished by computing the probability of a word boundary falling within such a sequence, with the probability being rewritten using Bayes' rule. The information needed for the computation of the word boundary probability is estimated on a small subset of the corpus, using the gold word boundaries. Thereafter, a boundary is placed between every diphone whose probability is above a predetermined threshold.

Monaghan and Christiansen (2010)'s PUDDLE is a lexical model which utilizes previously seen utterances to extract lexical and phonotactic information knowledge later used to "chunk" sequences. In a nutshell, it is an incremental algorithm that initially memorizes whole utterances into its long-term lexical storage, from which possible word-final and word-initial diphones are extracted. The model continues to consider each utterance as a lexical unit, unless sub-sequences of the given utterance have already been stored in the word list. In that case, it cuts the utterance based on the words which it already knows and considers the newly segmented chunks as word candidates. In order for the word candidates to be added to the lexical list, they have to respect two rules: 1) the final diphones of the left chunk and the beginning diphones of the right chunk must be on the list of permissible final diphones; and 2) both chunks have to contain at least one vowel. Once a candidate is added to the lexical list, its beginning and final diphones are included into the list of permissible diphones.

The last model was a unigram implementation of Adaptor Grammar (AG) (Johnson et al., 2007). AG is a hierarchical Bayesian model based on an extension of probabilistic context free grammars. It alternates between using the previously learned grammar to parse an utterance into a hierarchical tree structure made up of words and phonemes, and updating the grammar by learning probabilities associated to rules and entire tree fragments, called adapted non-terminals. The unigram model is the simplest grammar, considering utterances as being composed of words, which are represented as a sequence of phonemes.

3 Materials

The RIKEN corpus (Mazuka et al., 2006) contains recordings of 22 Japanese mothers interacting with their 18 to 24-month old infants, while playing with toys or reading a book. The same mothers were then recorded while talking to an experimenter. Out of the total 14.5 hours of recordings,

[1] https://github.com/alecristia/CDSwordSeg

about 11 hours represent infant-directed speech, while the rest adult-directed speech.

The corpus was annotated at both segmental and prosodic levels. We made use in this study of the prosodic boundary annotation, labelled using the X-JToBI standard (Maekawa et al., 2002). X-JToBI defines prosodic breaks based on the degree of their perceived disjuncture, ranging from level 0 (the weakest) to level 3 (the strongest). We use here level 2 and level 3 prosodic breaks, which in the Japanese prosodic organization (Venditti, 2005) correspond, respectively, to accentual phrases and intonational phrases. Accentual phrases are sequences of words that carry at most one pitch accent; for instance, a noun with a postposition will typically only have one accent. Intonational phrases are made up of sequences of accentual phrases, and constitute the domain where pitch range is defined such that, for instance, the onset of an intonational phrase will be marked by a reset the pitch level.

An additional dataset, part of the Corpus of Spontaneous Japanese (CSJ) (Maekawa, 2003), was considered as control. It contains academic speech and was previously used to investigate either the effect of speech register (Fourtassi et al., 2013) or that of prosodic boundaries (Ludusan et al., 2015) on unsupervised word segmentation. The same levels of annotations are available as for the RIKEN corpus. Statistics about the number of utterances and word token and types, for all three corpora, can be found in Table 1.

4 Experimental settings

The transitional probabilities used by TP were computed on the entire input dataset, while the estimation of the probabilities needed by DiBS was performed on the first 200 utterances of the corpus. PUDDLE, being an incremental algorithm, was evaluated using a five-fold cross-validation. For AG, the process was repeated five times for each register and prosodic boundary condition, and the average across the five runs was reported. Each run had 2000 iterations and Minimum Bayes Risk (Johnson and Goldwater, 2009) decoding was used for the evaluation.

Dataset	#utts	#tokens	#types
CSJ	20,052	216,932	7,340
ADS	3,582	22,844	2,022
IDS	14,570	51,315	2,850

Table 1: Statistics regarding the utterances and words contained in the investigated corpora.

Each algorithm was run on the ADS, IDS and CSJ datasets for each of the 3 cases considered: no prosody (*base*), level 3 prosodic breaks (*brk3*) and level 2 and level 3 prosodic breaks (*brk23*). For the base case, the system had in input a file containing on each line an utterance, defined as being an intonational phrase or a filler phrase followed by a pause longer than 200 ms. In the brk3 and brk23 cases, each prosodic phrase was considered as a standalone utterance, and thus was written on a separate line. During the evaluation of the brk3 and brk23 cases, the original utterances were rebuilt by concatenating all the prosodic phrases contained in them, after which they were compared against the reference.

Additionally, we checked whether the size difference between the ADS and IDS datasets might have an effect on the results obtained. For this, we created two additional, balanced, subsets of the IDS data. The first one contained an equal number of words from each speaker as in their ADS data, while the second one an equal number of utterances, for each speaker, as in their ADS production. As there was no significant difference between the results with the two balanced subsets and the entire IDS corpus, we will present here only the latter results.

5 Results and discussion

The segmentation evaluation was performed against the gold word segmentation, provided with the corpus. A classical metric, the token F-score, was used as evaluation measure. It is defined as the harmonic average between the token precision (how many word tokens, out of the total number of segmented words, were correct) and token recall (how many word tokens, out of the total number of words in the reference data, were found).

Next, we illustrate the obtained token F-score for the three corpora (IDS, ADS and CSJ) in Figure 1, for the three cases (base, brk3 and brk23) and for the four algorithms investigated (TP, DiBS, PUDDLE and AG). We observe that the largest differences are between algorithms. It appears that models employing sub-lexical information fare worse than the ones working at the lexical level. DiBS gives the lowest performance (.132 token F-score for CSJ base), followed by TP, PUDDLE

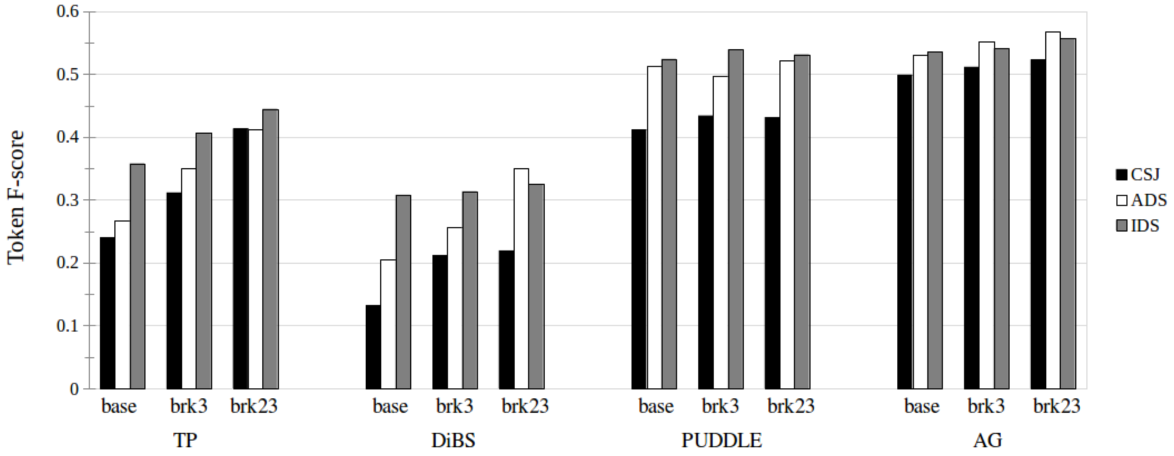

Figure 1: Segmentation results obtained using four algorithms: TP, DiBS, PUDDLE and AG on the IDS, ADS and CSJ datasets, when no prosodic information was provided (base), when utterances where additionally broken at boundaries of type 3 (brk3), and when utterances where additionally broken at boundaries of type 2 and 3 (brk23).

and AG giving the best performance (.567 token F-score for ADS brk23). The goal of the present study, however, is not to pit algorithms against each other, but rather to sample from plausible segmentation strategies that infants could potentially use so as to provide more representative and generalizable results.

Register effects found in the comparison between IDS and CSJ with the AG model replicate previous work (Fourtassi et al., 2013). We considerably extend knowledge by additionally including a casual ADS sample matched to the IDS, and investigating 3 additional algorithms. This allows us to conclude that differences between IDS and ADS are considerably smaller than previous work could have suggested. This is expected in view of previous reports that using un-matched materials leads to an overestimation of the differences between IDS and ADS (Batchelder, 2002). Interestingly, we also found that the size and direction of this difference was dependent on the algorithm used. An important advantage can be observed in the IDS-ADS comparison for the sub-lexical algorithms (maximally 9% for TP and 10.3% for DiBS), which decreases for PUDDLE and AG (maximally 1-1.1%), and can sometimes reverse when prosodic information is taken into account (DiBS brk23, AG brk3 and brk23).

Turning to prosodic boundaries, breaking utterances using internal prosodic breaks seems to help to a different degree the two classes of segmentation models and the three corpora, in ways that resemble a crossed interaction. The performance of sub-lexical models improves more with the use of prosodic information than that of lexical models, and this for all corpora. By and large, performance is boosted by additional prosodic breaks more for CSJ and ADS than IDS. This boost is, however, rather variable for PUDDLE, with apparent declines when, for instance, type 3 breaks are added for ADS. These results only partially replicate those reported in (Ludusan et al., 2015). Overall, the improvement brought by prosodic boundaries is smaller. TP brk23 brings an absolute improvement of 17.3% over TP base, for CSJ, but the improvement brought for AG (3.6%) is modest compared to what was previously reported (12.3%).[2]

Overall, we observe that some of our conclusions are dependent on the actual corpus being used. For this reason, we further analysed several measures which could play a role in the segmentation process. The first one, the average number of words per utterance was highest for CSJ, followed by ADS and the lowest for IDS. This would be expected taken into account the characteristics of IDS (Cristia, 2013). It is important to note that the smallest difference with respect to utterance length

[2] These differences might stem from the model used (we used here a unigram model, while a colloc3-syll model was previously used) or from the way in which the prosodic information was integrated (at the input level, in the current study, compared to at the grammar level, before). Indeed, a model that makes explicit in its grammar the prosodic boundaries and, thus, learns word boundaries jointly with prosodic boundaries could be more powerful. These aspects will have to be investigated in a future study.

Set	cond	phn	typ	wrd	ambig
CSJ	base	3.498	.584	10.82	.02918
	brk3			5.25	.01996
	brk23			2.75	.01195
ADS	base	3.089	.579	6.38	.02981
	brk3			3.57	.02217
	brk23			2.53	.01746
IDS	base	3.402	.522	3.52	.03099
	brk3			2.48	.02681
	brk23			2.06	.02425

Table 2: Detailed statistics on the three corpora used: average number of phonemes per word token (phn), average number of types per tokens (typ), average number of words per utterance (wrd), and segmentation ambiguity (ambig).

between the base and brk23 was obtained for IDS, the same register that seems to take advantage the least by the information on prosodic boundaries.

Besides the length of the utterance, the length of the words plays an important role in the segmentation task. Longer words would increase the possibility of having substrings which are words on their own, thus decreasing the segmentation performance. As expected, CSJ has the highest average word length, but IDS was found to have a very similar word length, followed by ADS. The unexpected value obtained for IDS might be due to the high number of long onomatopoeia present in the corpus. Thus, any IDS advantage due to having shorter utterances might be reversed by having longer words. We computed also the average number of types per token, which can give information about the distribution of the words in the corpora. In order not to have a measure biased by the size of the corpus, we computed it as a moving average over a window of 100 words. It shows a slightly higher vocabulary diversity for CSJ and ADS, than IDS, suggesting a more difficult segmentation.

The segmental ambiguity score (Fourtassi et al., 2013) measures the number of different parses of a sentence given the gold lexicon, by computing the average entropy in parses, taken into account the probability of each parse. Fourtassi and colleagues argue that this measure captures the intrinsic difficulty of the segmentation problem and predicts segmentation scores across languages (but see Phillips and Pearl (2014)). Here, we found that segmentation ambiguity decreases with the use of prosodic information (by preventing segmentations that would straddle a prosodic break). In contrast, there is not much difference between registers; if anything, IDS is more ambiguous than the two adult corpora; we speculate that this may be due to the presence of many onomatopoeia in IDS (over 8% of the total word tokens) some of which contain a lot of reduplications, which would increase segmentation ambiguity. This may explain why, when prosody equates sentence lengths, the advantage of IDS over ADS becomes small or even reverts to a detrimental effect.

6 Conclusions

We examined the performance of 4 different word segmentation algorithms on two matched corpora of spontaneous ADS and IDS, and a control corpus of more formal ADS, all of them with and without prosodic breaks. We found that, overall, sub-lexical algorithms perform less well than lexical algorithms, that IDS was overall slightly easier or equal to informal ADS, itself easier than formal ADS. In addition, across all algorithms and registers, we observed that prosody helped word segmentation. However, the impact of prosody was unequal and showed an interaction with register: It helped more ADS than IDS to the point that, with prosody taken into account, spontaneous ADS and IDS yield somewhat similar scores.

This has impact for theories of language acquisition, since IDS has been assumed to provide infants with 'hyperspeech', i.e. a simplified kind of input that facilitates language acquisition. If our observations are true, as far as word segmentation goes, it is not the case that IDS is massively easier to segment than ADS, at least at the stage when infants have acquired the ability to use prosodic breaks to constrain word segmentation. Of course, our observations would need to be confirmed and replicated with other languages and recording procedures. To conclude, our study illustrates the interest of testing theories of language acquisition using quantitative tools.

Acknowledgments

BL, MB and ED's research was funded by the European Research Council (ERC-2011-AdG-295810 BOOTPHON), and AC by the Agence Nationale pour la Recherche (ANR-14-CE30-0003 MechELex). It was also supported by the Canon Foundation in Europe and the Agence Nationale pour la Recherche (ANR-10-LABX-0087 IEC, ANR-10-IDEX-0001-02 PSL*).

References

Eleanor Olds Batchelder. 2002. Bootstrapping the lexicon: A computational model of infant speech segmentation. *Cognition* 83(2):167–206.

Benjamin Börschinger and Mark Johnson. 2014. Exploring the role of stress in bayesian word segmentation using adaptor grammars. *Transactions of the Association for Computational Linguistics* 2:93–104.

Michael R Brent and Timothy A Cartwright. 1996. Distributional regularity and phonotactic constraints are useful for segmentation. *Cognition* 61(1):93–125.

Anne Christophe, Ariel Gout, Sharon Peperkamp, and James Morgan. 2003. Discovering words in the continuous speech stream: the role of prosody. *Journal of phonetics* 31(3):585–598.

Alejandrina Cristia. 2013. Input to language: The phonetics and perception of infant-directed speech. *Language and Linguistics Compass* 7(3):157–170.

Suzanne Curtin, Toben H Mintz, and Morten H Christiansen. 2005. Stress changes the representational landscape: Evidence from word segmentation. *Cognition* 96(3):233–262.

Robert Daland and Janet B Pierrehumbert. 2011. Learning diphone-based segmentation. *Cognitive science* 35(1):119–155.

Gabriel Doyle and Roger Levy. 2013. Combining multiple information types in bayesian word segmentation. In *Proceedings of NAACL-HLT*. pages 117–126.

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

Sharon Goldwater, Thomas L Griffiths, and Mark Johnson. 2009. A bayesian framework for word segmentation: Exploring the effects of context. *Cognition* 112(1):21–54.

Mark Johnson and Sharon Goldwater. 2009. Improving nonparameteric bayesian inference: experiments on unsupervised word segmentation with adaptor grammars. In *Proceedings of NAACL-HLT*. pages 317–325.

Mark Johnson, Thomas Griffiths, and Sharon Goldwater. 2007. Adaptor grammars: A framework for specifying compositional nonparametric bayesian models. *Advances in neural information processing systems* 19:641.

Constantine Lignos. 2011. Modeling infant word segmentation. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning*. pages 29–38.

Bogdan Ludusan, Gabriel Synnaeve, and Emmanuel Dupoux. 2015. Prosodic boundary information helps unsupervised word segmentation. In *Proceedings of NAACL-HLT*. pages 953–963.

K. Maekawa. 2003. Corpus of Spontaneous Japanese: Its design and evaluation. In *ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition*.

Kikuo Maekawa, Hideaki Kikuchi, Yosuke Igarashi, and Jennifer Venditti. 2002. X-JToBI: an extended J-ToBI for spontaneous speech. In *Proceedings of INTERSPEECH*. pages 1545–1548.

Reiko Mazuka, Yosuke Igarashi, and Ken'ya Nishikawa. 2006. Input for learning Japanese: RIKEN Japanese mother-infant conversation corpus (COE Workshop session 2). *IEICE Technical Report* 106(165):11–15.

Padraic Monaghan and Morten H Christiansen. 2010. Words in puddles of sound: modelling psycholinguistic effects in speech segmentation. *Journal of child language* 37(03):545–564.

Lawrence Phillips and Lisa Pearl. 2014. Bayesian inference as a viable cross-linguistic word segmentation strategy: It's all about what's useful. In *Proceedings of CogSci*. pages 2775–2780.

Amanda Saksida, Alan Langus, and Marina Nespor. 2016. Co-occurrence statistics as a language-dependent cue for speech segmentation. *Developmental science* .

Amanda Seidl and Elizabeth K Johnson. 2006. Infant word segmentation revisited: Edge alignment facilitates target extraction. *Developmental science* 9(6):565–573.

Erik D Thiessen, Emily A Hill, and Jenny R Saffran. 2005. Infant-directed speech facilitates word segmentation. *Infancy* 7(1):53–71.

Jennifer Venditti. 2005. The J-ToBI model of Japanese intonation. In Sun-Ah Jun, editor, *Prosodic typology: The phonology of intonation and phrasing*, Oxford University Press, Oxford, pages 172–200.