Cognition 219 (2022) 104961

Contents lists available at ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/cognit

How much does prosody help word segmentation? A simulation study on infant-directed speech

Bogdan Ludusan [a,b,1,*], Alejandrina Cristia [b], Reiko Mazuka [a,c], Emmanuel Dupoux [d]

[a] Laboratory for Language Development, RIKEN Center for Brain Science, Japan
[b] Laboratoire de Sciences Cognitives et Psycholinguistique, ENS Paris Sciences Lettres, EHESS, CNRS, France
[c] Department of Psychology and Neuroscience, Duke University, USA
[d] Laboratoire de Sciences Cognitives et Psycholinguistique, ENS Paris Sciences Lettres, EHESS, CNRS, INRIA, France

ARTICLE INFO

Keywords:
Prosody
Word segmentation
Computational model
Infant language acquisition
Infant-directed speech

ABSTRACT

Infants come to learn several hundreds of word forms by two years of age, and it is possible this involves carving these forms out from continuous speech. It has been proposed that the task is facilitated by the presence of prosodic boundaries. We revisit this claim by running computational models of word segmentation, with and without prosodic information, on a corpus of infant-directed speech. We use five cognitively-based algorithms, which vary in whether they employ a sub-lexical or a lexical segmentation strategy and whether they are simple heuristics or embody an ideal learner. Results show that providing expert-annotated prosodic breaks does not uniformly help all segmentation models. The sub-lexical algorithms, which perform more poorly, benefit most, while the lexical ones show a very small gain. Moreover, when prosodic information is derived automatically from the acoustic cues infants are known to be sensitive to, errors in the detection of the boundaries lead to smaller positive effects, and even negative ones for some algorithms. This shows that even though infants could potentially use prosodic breaks, it does not necessarily follow that they should incorporate prosody into their segmentation strategies, when confronted with realistic signals.

1. Introduction

Laboratory studies suggest that infants are sensitive to prosodic cues and use them to segment utterances (Hirsh-Pasek et al., 1987; Wellmann, Holzgrefe, Truckenbrodt, Wartenburger, & Höhle, 2012). Large prosodic breaks (akin to sentence boundaries) facilitate word segmentation: Infants recognized a word better when it was aligned with such boundaries (e.g., Shukla, White, & Aslin, 2011; E. Johnson, Seidl, & Tyler, 2014; Seidl & Johnson, 2008). Based on these results, one could conclude that prosodic breaks help infants discover words in real life. What is missing to draw this conclusion, however, is an estimate of the *prevalence*, *detectability* and the *added value* of prosodic boundaries in natural conditions. We review these three factors in turn.

First, prosodic breaks may not be *prevalent* in natural infant-directed speech (IDS), because the utterances given to infants are already short (see Cristia, 2013 for a review). Short utterances, to the extent that they are separated by easy-to-detect pauses, provide word boundaries for free and reduce segmentation ambiguity. The longer the utterance, the greater the number of possible parses, and the more prosodic breaks can provide additional segmentation information. To assess the potential effectiveness of prosody for segmentation, it is therefore important to study it in actual IDS and not in artificially constructed stimuli.

Second, within-utterance prosodic breaks may be difficult to *detect* in the raw signal, yielding segmentation errors. Laboratory evidence suggests infants typically require multiple acoustic cues (pitch reset, pre-boundary lengthening, and pause) to detect a break (Seidl, 2007; Soderstrom, Seidl, Kemler Nelson, & Jusczyk, 2003; Wellmann et al., 2012), with the possible exception of long pauses (E. Johnson & Seidl, 2008). Also adults are better at discovering breaks when these are more strongly marked (e.g., discussions in E. Johnson & Seidl, 2008 and Wellmann et al., 2012). Speech technology research suggests that extracting prosodic cues from phonetic information is challenging (e.g., Ananthakrishnan & Narayanan, 2007), and several systems use lexical or syntactic information to help improve the results. To assess effectiveness, it is therefore important to use not only gold-standard prosodic breaks, but also prosodic breaks that can be reasonably extracted by pre-

* Corresponding author.
E-mail address: bogdan.ludusan@brain.riken.jp (B. Ludusan).
[1] Currently affiliated with the Phonetics Workgroup, Faculty of Linguistics and Literary Studies, Bielefeld University.

https://doi.org/10.1016/j.cognition.2021.104961
Received 22 October 2020; Received in revised form 1 July 2021; Accepted 15 November 2021
Available online 29 November 2021

linguistic infants. While speech technology research showing prosodic break discovery is difficult was based on adult-directed speech corpora, utterance-internal breaks are easier to detect in IDS (Ludusan, Cristia, Martin, Mazuka, & Dupoux, 2016).

Third, regarding *added value*, perhaps prosody, although prevalent and detectable, is redundant with other cues available for word segmentation (e.g., phonotactic or lexical cues), so that prosodic breaks do not contribute additional information. This possibility is difficult to study with human infants, because we cannot require them to use only specific cues when all cues are present. Computational models programmed to use some of these cues, thus, enable us to evaluate the effective role of prosodic breaks when segmenting realistic corpora.

1.1. Modeling word segmentation

The process of infant word segmentation has received considerable attention from computational linguists. Various symbolic segmentation models have been proposed (e.g., Brent & Cartwright, 1996; Goldwater, Griffiths, & Johnson, 2009) that use as input a phonemic transcription and return a hypothesized segmentation, to be compared against (orthographic) word boundaries. Before summarizing this literature further, we acknowledge two limitations. First, strings of phonemes may not be the most accurate input representation for young infants' perception. However, algorithmic development is more developed for text than for acoustic input (Ludusan et al., 2014). Additionally, Daland and Pierrehumbert (2011) argue that it is not an unreasonable representation format for older infants. Second, the ideal evaluation is against infants' segmentation of the exact same phrases. Such results are currently unavailable and are difficult to gather at scale given the low power of infant experiments. Therefore, performance against adult-defined gold standards is our best proxy at present.

Two classes of algorithms are used in this study, sub-lexical and lexical. *Sub-lexical* algorithms attempt to detect possible word boundaries by relying on statistics over sequences of phonemes. Here, we use two such algorithms, one relying on transition probabilities (TP), and another on diphone frequencies (DiBS). TP was inspired by the intuition that transition probabilities within words are typically larger than between words, a property that infants are sensitive to (e.g., Saffran, Aslin, & Newport, 1996; see Saksida, Langus, & Nespor, 2017 for a computational approach). We employ two versions of the TP algorithm, one taking a decision whether a word boundary should be posited between two syllables locally (comparing the current TP to that of neighbouring disyllables; TP_R) and another based on global statistics (comparing the TP against an absolute value; TP_A). DiBS operates optimal Bayesian inference to learn the parameters of a mixture of within- and between-word diphones. Daland and Pierrehumbert (2011) summarizes the cognitive relevance of this approach, notably infants' sensitivity to phonotactics (Mattys, Jusczyk, Luce, & Morgan, 1999).

Lexical algorithms attempt to build a lexicon and use it to break utterances up into words. In these models, the assignment of word boundaries is a by-product of word recognition. We use two algorithms, PUDDLE (PUD; Monaghan & Christiansen, 2010) and Adaptor Grammar (AG; M. Johnson, Griffiths, & Goldwater, 2006). PUDDLE employs both lexical strategies and phonotactics. It starts by storing whole utterances it comes across in a lexicon component. Then, it tries to break new input utterances using the memorized chunks and, if they conform to the phonotactics derived from the lexicon, it adds the new chunks to the lexicon and updates the phonotactics. Using known chunks to segment incoming input is a behavior found even among 6-month-olds (Bortfeld, Morgan, Golinkoff, & Rathbun, 2005). Finally, Adaptor Grammar represents the optimal learner within the lexical sub-class, and its cognitive relevance of this approach for infant word segmentation has been argued for in Goldwater et al. (2009).

Although there are many other segmentation algorithms, we use these as a representative sample, covering both sub-lexical and lexical strategies, as well as heuristic (TP_R, TP_A, PUDDLE) and optimal (DiBS, AG) algorithms.

1.2. The present study

This study quantitatively evaluates the effective role of prosodic boundaries on word segmentation, using a computational approach. It builds on previous computational evaluations of word segmentation in real corpora. Ludusan, Synnaeve, and Dupoux (2015) investigated one computational model in adult-directed English and Japanese corpora, finding that performance improved with prosodic boundary information. Due to the specific characteristics of infant-directed input, the effectiveness of such information has to be tested directly with IDS input. Ludusan, Mazuka, Bernard, Cristia, and Dupoux (2017) investigated the effects of prosodic information on word segmentation in Japanese IDS and two ADS corpora, using several segmentation algorithms, and found effects that varied across algorithms and corpora. However, only gold standard prosodic boundaries were employed.

Here, we fill this gap by focusing on IDS from a large naturalistic corpus and applying a phonetically-based prosodic segmentation algorithm to estimate how children may access prosodic boundaries in real life. We apply the five word segmentation algorithms defined above with and without gold prosodic breaks (posited by human annotators), as well as with automatically extracted breaks.

2. Methods

2.1. Corpus

The RIKEN corpus (Mazuka, Igarashi, & Nishikawa, 2006) contains recordings of spontaneous interactions between 22 Japanese mothers and their 18 to 24-month-old infants, while playing with toys or reading a book. This resulted in about 11 h of IDS, which were entirely annotated at the segmental and prosodic levels.

The prosody in the corpus was labelled based on X-JToBI rules (Maekawa, Kikuchi, Igarashi, & Venditti, 2002) adapted to Japanese prosodic organization (Venditti, 2005). In this standard, prosodic breaks are annotated by expert native Japanese coders, based on their perception of the degree of disjuncture, from weakest (0, corresponding to word breaks in fast speech) to strongest (3, akin to intonation phrase boundaries). The annotators had access to all sources of phonetic and linguistic information to decide the strength of a break. This study uses levels 2 (corresponding to accentual phrase boundaries) and 3.

Table 1 provides information on the number of words per utterance and number of prosodic phrases per utterance, counting as breaks only utterance breaks (without); utterance and level 3 breaks (brk3); utterance, level 2, and level 3 breaks (brk23); or utterance and automatically identified breaks (brkA). Corpus statistics are available in Section S1 of the Supplementary Materials.

2.2. Analyses

We provide here key aspects of the analyses. For more information, refer to Sections S2-S3 of the Supplementary Materials.

The automatic prosodic boundary detection algorithm (similar to the one in Ludusan & Dupoux, 2014) employs acoustic cues that have been shown to be used by infants for the recognition of prosodic boundaries (e.g., Wellmann et al., 2012). Specifically, we extracted at the syllable level the duration of following pause, the duration of current syllable

Table 1
Average number of words per utterance (wrd/utt) and prosodic phrases per utterance (phr/utt), for the four prosody conditions.

Measure	without	brk3	brk23	brkA
wrd/utt	3.52	2.48	2.06	2.92
phr/utt	1	1.419	1.714	1.243

nucleus, the nucleus-onset-to-nucleus-onset duration and the difference between the average pitch value of the next syllable nucleus and that of the current syllable. The contribution of each cue was then normalized between 0 and 1 and their sum computed, for every syllable (an equal weight was given to each cue). The resulting syllable-based function was then used for positing prosodic boundaries, as follows: The local maxima of the function were determined and prosodic boundaries were placed after the syllables corresponding to those maxima. The system does not employ any sort of learning paradigm, placing boundaries based only on the strength of the acoustic cues marking that particular syllable. Therefore, the automatically obtained boundaries do not correspond to a specific prosodic boundary level, although higher-level phrase boundaries are more likely to be discovered, given that they are stronger marked acoustically. The manual segmentation provided with the corpus was used to derive the duration cues employed by the detection algorithm. Pitch was extracted using Praat (Boersma, 2001) and any tracking errors were hand-corrected. We evaluate the goodness of boundary placing with respect to the manual annotations provided with the corpus, by computing the F-score (the harmonic mean between precision and recall). Its value varies between 0 and 1, with the latter representing a perfect system.

Word segmentation algorithms are part of the open source package WordSeg (Bernard et al., 2020). Each was run on the four prosodically-defined versions of the dataset (without, brk3, brk23, brkA). Thus, the input was a file containing on each line the sequence of phonemes (with word boundaries removed) corresponding to segmental content, with breaks as defined in each condition (e.g. for without, only utterance boundaries were breaks, whereas for brk23, breaks corresponded to utterance and levels 2–3). Prosodic boundary information should help the segmentation models by providing free information on word boundaries (right edge of word preceding boundary and left edge of word following it). This would have an impact on the values of the probabilities employed in sub-lexical models and on the dictionary items determined by lexical models.

The returned segmentation was compared to the gold orthographic word segmentation, using token F-score – the harmonic average between token precision (how many word tokens, out of the total number of segmented word tokens, were correct) and token recall (how many word tokens, out of the total number of word tokens in the reference data, were found). All algorithms were run by means of 5-fold validation, the evaluation being performed each time on the last 20% of the corpus. Although only PUDDLE is incremental, previous work suggests performance is remarkably stable as function of corpus size for the other algorithms (Bernard et al., 2020). We also computed the word correctness, defined as the percentage of correctly segmented words out of the total number of words, and the degree of under−/over-segmentation of the model. This measure represents the difference between the number of words found by the models and the actual words in the dataset, normalized by the latter. Its sign (positive/negative) gives the direction of the trend (over/under-segmentation), while its absolute value shows the degree of over/under-segmentation.

The word segmentation results were analyzed using mixed-effects models. For each group of model classes (sub-lexical/lexical), we fitted a model having the token F-score as dependent variable, the prosody condition as fixed effect and the corpus sub-part as random intercept. We used a Bayesian framework, as we also wanted to determine the posterior probability of the prosody-enabled conditions to bring performance gains compared to the baseline case. The R (R Core Team, 2020) package brms (Bürkner, 2017) was employed, with a weakly informed prior (uniform distribution) and five Markov chains, each having 3000 iterations (of which 1000 for warm-up).

3. Results

We present here the main results of the study, more details being given in the Supplementary Materials, Section S4.

3.1. Manual prosody

Fig. 1 shows that manual prosody (light grey/grey bars) helps both sub-lexical and lexical word segmentation models, with a greater gain for the former.[2] The absolute F-score gain is 3.5%–8.6% (sub-lexical) and 1.5%–2.1% (lexical). The improvement in segmentation performance was mainly due to the correct segmentation of a greater percentage of words (Fig. S3, Supplementary Materials), the word correctness increasing for all models, by 0.3%–5.8% for brk3 and by 1.3%–10.5% for brk23 condition. Prosody also helped the models reduce their over/under- segmentation, with all models except TP_A and PUD brk23 showing this effect (for more details see Supplementary Materials, Fig. S4).

3.2. Automatic prosody

Evaluating the goodness of the automatically determined prosodic boundaries, we observe a system favouring precision (0.685) over recall (0.304), with an overall F-score of 0.421. Despite this rather poor performance, the automatically obtained prosodic boundaries correspond to word boundaries in 85.8% of the cases.

The performance of the models integrating the automatic boundaries is illustrated in Fig. 1 (dark grey bars). Automatic boundaries improved performance slightly for sub-lexical models (between 0.6% and 0.8% F-score gain), but decreased it for lexical models (−3.8% and − 1.1%). The number of correctly segmented words decreased for lexical models and increased for sub-lexical models. Additionally, only TP_R and DiBS saw improvements in under/over-segmentation when integrating automatic boundaries.

4. Discussion and conclusions

May infants profit from utterance-internal prosodic breaks when segmenting words from their everyday input? Our modeling results suggest, first and foremost, that the facilitatory effect of prosodic boundary, if it exists, is modest in size, compared to the intrinsic differences between segmentation algorithms. This may relate to the fact that intra-utterance prosodic boundaries have a relatively low prevalence in IDS (see Table 1), and therefore do not have a lot of occasions to make a difference. This conclusion has to be qualified by two key issues: Which segmentation algorithm infants may be using, and how accurate their detection of such prosodic breaks are. If infants use a lexical strategy (as documented e.g. in Bortfeld et al., 2005), then they do not stand to gain much, but if they employ a sub-lexical strategy (as in Saffran et al., 1996) or treat all sentences as words (as suggested by Keren-Portnoy, Vihman, & Lindop Fisher, 2019) they would, provided they can retrieve the breaks expert adult annotators tagged in this corpus. However, as the annotators had access also to higher-level linguistic information when taking a prosodic boundary decision, these results show the highest possible improvement brought by boundaries. As infants' prosodic break detection is errorful, they would gain little if using a sub-lexical strategy, and lose performance if they were using a lexical strategy. Our findings are consistent with those of previous studies, showing that prosodic boundaries in IDS are more easily identified than in adult-directed speech (Ludusan et al., 2016) (see section S2 of the Supplementary Materials for more details) and that they offer a lower boost to segmentation compared to gold-standard boundaries (Ludusan et al., 2015).

Our computational approach suggests two key areas of further

[2] The differences between the conditions with and without prosody are rather small, though, compared to the differences between algorithms. The difference between models (in the case without prosody) is nearly three times larger (22.9%, representing the difference between the best performing model, AG, and the worst performing one, DiBS) than the largest prosody gain, 8.6%.

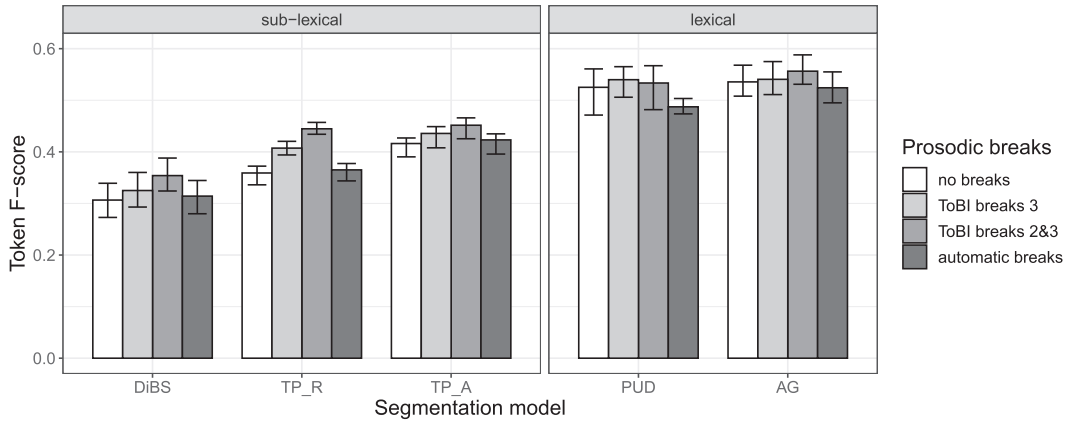

Fig. 1. Segmentation performance as a function of segmentation model and prosodic condition.

research. With respect to computational modeling, future research could explore segmentation approaches taking as input the acoustic signal, as well as the integration of language-specific cues (such as lexical stress, which could be helpful in some languages, Yang, 2004; Börschinger & Johnson, 2014). In order to better understand these cross-linguistic aspects, additional work on other languages, with diverse prosodic structures, would be desirable. A second interesting area arises from the fact that prosodic cues seem beneficial only for certain algorithms, and not others. This raises a sort of "meta-algorithmic" problem: given that infants have a whole range of segmentation strategies in their toolkit, and perceive a large number of phonetic cues, how do they know which particular combination of segmentation strategy and phonetic cue is helpful in their particular case? This may be especially difficult to solve if the strategy-cue interaction is language dependant (see Fourtassi & Dupoux, 2014 for an attempt at solving a similar meta-algorithmic problem).

Concerning infant laboratory experimentation, our study makes an interesting prediction: Utterance-internal breaks are most useful when infants employ sub-lexical strategies, and less so when using lexical ones. It is likely that infants younger than 6 months need to rely more on sub-lexical strategies, whereas older infants, who have been able to accumulate a pseudo-lexicon, should be in a better position to integrate lexical cues.[3] If infants are able to use a meta-algorithmic approach to find optimal strategies, then we should observe a greater reliance on utterance-internal breaks at earlier ages than at later ages, a prediction that could be investigated by familiarizing infants with passages in which the target word has one edge aligned with an utterance-internal break. However, besides these two strategies investigated here, there are other sources of information available to infants, also at younger ages, to help with the segmentation (e.g. words appearing in isolation; Keren-Portnoy et al., 2019) and which are worth investigating.

To conclude, this study illustrates the interest of complementing traditional language acquisition studies via laboratory experimentation with a computational modeling approach that simulates the learning process itself (Dupoux, 2018). Whereas the former addresses the existence of potential learning mechanisms or perceptual cues, the latter enables us to evaluate the effectiveness of such mechanisms and cues on realistic input data. More research is needed in this direction, notably, in linking more tightly the results of the learning simulations to actual outcome measures (see Larsen, Cristia, & Dupoux, 2017, for a potential approach).

[3] Please note that this does not map onto the speech versus statistics discussions (E. Johnson & Jusczyk, 2001; Thiessen & Saffran, 2003), because both coarticulation and statistics are sub-lexical and could be language-universal strategies. The strategy of positing word boundaries before strong syllables, on the other hand, probably needs to be learned from the native language and it may involve non-local cues, being in this sense closer to the lexical strategies studied here.

Acknowledgements

The research reported in this paper was partly funded by JSPS Grant-in-Aid for Scientific Research (16H06319, 20H05617) and MEXT Grant-in-aid on Innovative Areas #4903 (Co-creative Language Evolution), 17H06382 to RM. It was also supported by the European Research Council (ERC-2011-AdG-295810 BOOTPHON), the Agence Nationale pour la Recherche (ANR-17-CE28-0007 LangAge, ANR-16-DATA-0004 ACLEW, ANR-14-CE30-0003 MechELex, ANR-17-EURE-0017 Frontcog, ANR-10-LABX-0087 IEC, ANR-10-IDEX-0001-02 PSL*, ANR-19-P3IA-0001 PRAIRIE 3IA Institute). ED is further grateful to the CIFAR (Learning in Machines and Brain), BL to the Canon Foundation in Europe, and AC to the JS McDonnell Foundation.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.cognition.2021.104961.

References

Ananthakrishnan, S., & Narayanan, S. (2007). Automatic prosodic event detection using acoustic, lexical, and syntactic evidence. *IEEE Transactions on Audio, Speech and Language Processing, 16*(1), 216–228. https://doi.org/10.1109/TASL.2007.907570

Bernard, M., Thiolliere, R., Saksida, A., Loukatou, G., Larsen, E., Johnson, M., … Cristia, A. (2020). WordSeg: Standardizing unsupervised word form segmentation from text. *Behavior Research Methods, 52*(1), 264–278. https://doi.org/10.3758/s13428-019-01223-3

Boersma, P. (2001). Praat, a system for doing phonetics by computer. *Glot International, 5*(9), 314–345.

Börschinger, B., & Johnson, M. (2014). Exploring the role of stress in Bayesian word segmentation using adaptor grammars. *Transactions of the Association for Computational Linguistics, 2*, 93–104. https://doi.org/10.1162/tacl_a_00168

Bortfeld, H., Morgan, J., Golinkoff, R. M., & Rathbun, K. (2005). Mommy and me: Familiar names help launch babies into speech-stream segmentation. *Psychological Science, 16*(4), 298–304. https://doi.org/10.1111/j.0956-7976.2005.01531.x

Brent, M., & Cartwright, T. (1996). Distributional regularity and phonotactic constraints are useful for segmentation. *Cognition, 61*(1), 93–125. https://doi.org/10.1016/S0010-0277(96)00719-6

Bürkner, P.-C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software, 80*(1), 1–28. https://doi.org/10.18637/jss.v080.i01

Cristia, A. (2013). Input to language: The phonetics and perception of infant-directed speech. *Lang & Ling Compass, 7*(3), 157–170. https://doi.org/10.1111/lnc3.12015

Daland, R., & Pierrehumbert, J. (2011). Learning diphone-based segmentation. *Cognitive Science, 35*(1), 119–155. https://doi.org/10.1111/j.1551-6709.2010.01160.x

Dupoux, E. (2018). Cognitive science in the era of artificial intelligence: A roadmap for reverse-engineering the infant language-learner. *Cognition, 173*, 43–59. https://doi.org/10.1016/j.cognition.2017.11.008

Fourtassi, A., & Dupoux, E. (2014). A rudimentary lexicon and semantics help bootstrap phoneme acquisition. In *Proceedings of the 18th conference on computational language learning* (pp. 191–200). https://doi.org/10.3115/v1/W14-1620

Goldwater, S., Griffiths, T., & Johnson, M. (2009). A Bayesian framework for word segmentation: Exploring the effects of context. *Cognition, 112*(1), 21–54. https://doi.org/10.1016/j.cognition.2009.03.008

Hirsh-Pasek, K., Kemler Nelson, D., Jusczyk, P., Cassidy, K., Druss, B., & Kennedy, L. (1987). Clauses are perceptual units for young infants. *Cognition, 26*(3), 269–286. https://doi.org/10.1016/S0010-0277(87)80002-1

Johnson, E., & Jusczyk, P. (2001). Word segmentation by 8-month-olds: When speech cues count more than statistics. *Journal of Memory and Language, 44*(4), 548–567. https://doi.org/10.1006/jmla.2000.2755

Johnson, E., & Seidl, A. (2008). Clause segmentation by 6-month-old infants: A crosslinguistic perspective. *Infancy, 13*(5), 440–455. https://doi.org/10.1080/15250000802329321

Johnson, E., Seidl, A., & Tyler, M. (2014). The edge factor in early word segmentation: Utterance-level prosody enables word form extraction by 6-month-olds. *PLoS One, 9*(1), Article e83546. https://doi.org/10.1371/journal.pone.0083546

Johnson, M., Griffiths, T., & Goldwater, S. (2006). Adaptor grammars: A framework for specifying compositional nonparametric Bayesian models. In *Proceedings of the 19th international conference on neural information processing systems* (pp. 641–648).

Keren-Portnoy, T., Vihman, M., & Lindop Fisher, R. (2019). Do infants learn from isolated words? An ecological study. *Language Learning and Development, 15*(1), 47–63. https://doi.org/10.1080/15475441.2018.1503542

Larsen, E., Cristia, A., & Dupoux, E. (2017). Relating unsupervised word segmentation to reported vocabulary acquisition. In *Proceedings of INTERSPEECH* (pp. 2198–2202). https://doi.org/10.21437/Interspeech.2017-937

Ludusan, B., Cristia, A., Martin, A., Mazuka, R., & Dupoux, E. (2016). Learnability of prosodic boundaries: Is infant-directed speech easier? *The Journal of the Acoustical Society of America, 140*(2), 1239–1250. https://doi.org/10.1121/1.4960576

Ludusan, B., & Dupoux, E. (2014). Towards low-resource prosodic boundary detection. In *Proceedings of the workshop on spoken language technologies for under-resourced languages* (pp. 231–237).

Ludusan, B., Mazuka, R., Bernard, M., Cristia, A., & Dupoux, E. (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 (short papers)* (pp. 178–183). https://doi.org/10.18653/v1/P17-2028

Ludusan, B., Synnaeve, G., & Dupoux, E. (2015). Prosodic boundary information helps unsupervised word segmentation. In *Proceedings of human language technologies: The 2015 annual conference of the North American chapter of the ACL* (pp. 953–963). https://doi.org/10.3115/v1/N15-1096

Ludusan, B., Versteegh, M., Jansen, A., Gravier, G., Cao, X.-N., Johnson, M., & Dupoux, E. (2014). Bridging the gap between speech technology and natural language processing: An evaluation toolbox for term discovery systems. In *Proceedings of the 9th international conference on language resources and evaluation* (pp. 560–567).

Maekawa, K., Kikuchi, H., Igarashi, Y., & Venditti, J. (2002). X-JToBI: An extended J-ToBI for spontaneous speech. In *Proceedings of the 7th international conference on spoken language processing* (pp. 1545–1548).

Mattys, S., Jusczyk, P., Luce, P., & Morgan, J. (1999). Phonotactic and prosodic effects on word segmentation in infants. *Cognitive Psychology, 38*(4), 465–494. https://doi.org/10.1006/cogp.1999.0721

Mazuka, R., Igarashi, Y., & Nishikawa, K. (2006). Input for learning Japanese: RIKEN Japanese mother-infant conversation corpus. *IEICE Technical Report, 106*(165), 11–15.

Monaghan, P., & Christiansen, M. (2010). Words in puddles of sound: Modelling psycholinguistic effects in speech segmentation. *Journal of Child Language, 37*(03), 545–564. https://doi.org/10.1017/S0305000909990511

R Core Team. (2020). R: A language and environment for statistical computing [computer software manual]. Vienna, Austria. Retrieved from https://www.R-project.org/.

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science, 274*, 1926–1928. https://doi.org/10.1126/science.274.5294.1926

Saksida, A., Langus, A., & Nespor, M. (2017). Co-occurrence statistics as a language-dependent cue for speech segmentation. *Developmental Science, 20*(3), Article e12390. https://doi.org/10.1111/desc.12390

Seidl, A. (2007). Infants' use and weighting of prosodic cues in clause segmentation. *Journal of Memory and Language, 57*(1), 24–48. https://doi.org/10.1016/j.jml.2006.10.004

Seidl, A., & Johnson, E. (2008). Boundary alignment enables 11-month-olds to segment vowel initial words from speech. *Journal of Child Language, 35*(1), 1–24. https://doi.org/10.1017/S0305000907008215

Shukla, M., White, K., & Aslin, R. (2011). Prosody guides the rapid mapping of auditory word forms onto visual objects in 6-mo-old infants. *Proceedings of the National Academy of Sciences, 108*(15), 6038–6043. https://doi.org/10.1073/pnas.1017617108

Soderstrom, M., Seidl, A., Kemler Nelson, D., & Jusczyk, P. (2003). The prosodic bootstrapping of phrases: Evidence from prelinguistic infants. *Journal of Memory and Language, 49*(2), 249–267. https://doi.org/10.1016/S0749-596X(03)00024-X

Thiessen, E., & Saffran, J. (2003). When cues collide: Use of stress and statistical cues to word boundaries by 7-to 9-month-old infants. *Developmental Psychology, 39*(4), 706–716. https://doi.org/10.1037/0012-1649.39.4.706

Venditti, J. (2005). The J_ToBI model of Japanese intonation. In S.-A. Jun (Ed.), *Prosodic typology: The phonology of intonation and phrasing* (pp. 172–200). Oxford: Oxford University Press.

Wellmann, C., Holzgrefe, J., Truckenbrodt, H., Wartenburger, I., & Höhle, B. (2012). How each prosodic boundary cue matters: Evidence from German infants. *Frontiers in Psychology, 3*, 580. https://doi.org/10.3389/fpsyg.2012.00580

Yang, C. D. (2004). Universal grammar, statistics or both? *Trends in Cognitive Sciences, 8*(10), 451–456.