Prosodic boundary information helps unsupervised word segmentation

Bogdan Ludusan, Gabriel Synnaeve and Emmanuel Dupoux
Laboratoire de Sciences Cognitives et Psycholinguistique
EHESS / ENS / CNRS
29 rue d'Ulm, 75005 Paris, France
bogdan.ludusan@ens.fr, gabriel.synnaeve@gmail.com,
emmanuel.dupoux@gmail.com

Abstract

It is well known that prosodic information is used by infants in early language acquisition. In particular, prosodic boundaries have been shown to help infants with sentence and word-level segmentation. In this study, we extend an unsupervised method for word segmentation to include information about prosodic boundaries. The boundary information used was either derived from oracle data (hand-annotated), or extracted automatically with a system that employs only acoustic cues for boundary detection. The approach was tested on two different languages, English and Japanese, and the results show that boundary information helps word segmentation in both cases. The performance gain obtained for two typologically distinct languages shows the robustness of prosodic information for word segmentation. Furthermore, the improvements are not limited to the use of oracle information, similar performances being obtained also with automatically extracted boundaries.

1 Introduction

Prosodic information is thought to play a fundamental role in early language acquisition, and provide infants with rich structural information about their language (Christophe et al., 1997). In particular, prosody has been claimed to help infants find word boundaries (Christophe and Dupoux, 1996). Newborns are able discriminate between disyllables that contains vs. does not contain a phonological phrase boundary (Christophe et al., 1994; Christophe et al., 2001), showing that they are able to encode the corresponding prosodic cues. Nine-month olds show evidence of parsing utterances into prosodic units, and show 'surprise' when a pause is inappropriately inserted inside as opposed to between these units (Jusczyk et al., 1992; Gerken et al., 1994). Ten to 13 month olds show evidence of using prosodic units to parse utterances into words, as they fail to recognize a familiar word if it appears to straddle a prosodic boundary (Gout et al., 2004).

Curiously enough, however, prosody is not used very much in unsupervised models of language acquisition, and in particular, in models of word segmentation. Most such models use text as input, and apply some form of lexical optimization. For instance, Brent and Cartwright (1996) used a Minimal Description Length Principle to optimize the size of the description of a corpus. State of the art systems use hierarchical Bayesian models (Goldwater et al., 2009) which parse a corpus into words or other linguistic units with a bias to reuse previously parsed elements. Adaptor Grammars is a generic framework which enables to formulate such Bayesian models within an overarching architecture based on probabilistic context free grammars (Johnson et al., 2007). Such models have been used to study the role of linguistic information such as syllabic structure (Johnson and Goldwater, 2009), morphology (Johnson, 2008), function words (Johnson et al., 2014), as well as the role of non-linguistic context (Synnaeve et al., 2014). To our knowledge, only one paper studied the role of prosodic information (Börschinger and Johnson, 2014). In this study, the authors used the role of word stress in constraining word segmentation (as in stress languages, there is only one main stress per word).

Human Language Technologies: The 2015 Annual Conference of the North American Chapter of the ACL, pages 953–963,
Denver, Colorado, May 31 – June 5, 2015.

Here, we test whether prosodic boundaries could directly help symbolic word segmentation by providing some word boundaries 'for free', as this was already shown to be true in the case of signal-based term discovery systems (Ludusan et al., 2014). Being a feasibility study, we will use gold prosodic boundaries in order to quantify what is the maximum gain we can expect using this type of information. In addition to that, we test whether prosodic boundaries automatically derived from the speech signal (Ludusan and Dupoux, 2014) could also provide a performance gain. As this study relies on the existence of prosodic information (either gold, or derived from speech), we did not use the standard corpora used in these studies (the Bernstein-Ratner corpus), but introduced three new corpora, two in English and one in Japanese.

The paper is structured as follows: In the next sections we introduce the systems employed in this study - the prosodic boundary detection system in section 2 and the word segmentation procedure in section 3. Next, we present the datasets used in the experiments, with the results obtained being illustrated in section 5. The paper will conclude with a general discussion and some final remarks.

2 Prosodic annotation

There are numerous studies in the speech processing literature focusing on the detection of prosodic boundaries (e.g. Wightman and Ostendorf (1991), Ananthakrishnan and Narayanan (2008), Huang et al. (2008), Jeon and Liu (2009), just to name a few). While the approaches taken vary between these studies, they tend to use either supervised learning, thus needing large, prosodically annotated corpora, or higher level information (syntactic, lexical, etc) which would also require further annotations. Since unsupervised word segmentation is a process that requires low resources (only symbolic transcription), we have decided to use for the automatic detection of prosodic boundaries a previously proposed method which employs only acoustic cues that can be extracted from the speech signal (Ludusan and Dupoux, 2014).

The algorithm takes into consideration four acoustic cues which had been shown, in the language acquisition literature, to be used by young infants for the recognition of prosodic boundaries. The cues correspond to the following phenomena that occur next to prosodic breaks: silent pauses, final lengthening, initial strengthening and F0 reset. The acoustic cues were extracted at the syllable level and they include: the duration of the pause following the syllable (pause cue), the syllable nucleus duration (nucleus cue), the distance between the nucleus onset of the current syllable and that of the following one (onset cue) and the difference between the F0 end value of the current syllable and the F0 beginning value of the following syllable (F0 reset cue). The nucleus and onset cues are computed for all the syllables, the later being a combination of the nucleus cue, pause cue and the onset of the following syllable, which is the domain of the initial strengthening phenomenon. The pause cue is set to 0 for syllables not followed by a silence pause, while F0 reset is only computed for syllables which are at a local minimum for F0, otherwise it is set to 0. Then, for each individual cue function except pause, we considered only the values which were local maxima, the other values being set to 0.

Once a numerical value for each of the cues is obtained, they are standardized between 0 and 1 and combined in a detector function, by summing them up. The local maxima of the detector function are then obtained and the syllables corresponding to the maxima will be considered as prosodic boundary candidates. Next, a thresholding of these values is applied and all the right-hand boundaries of the syllables greater or equal to this threshold are marked as prosodic boundaries. This operation is followed by a second step in which prosodic boundaries are marked based on a different rule, rule that we would call conjunction of cues. This rule was inspired by the results of several studies in the infant literature (Seidl, 2007; Wellmann et al., 2012) showing that most prosodic boundaries tend to be marked by more than one acoustic cue. Taking these findings into account, we could also mark as prosodic boundaries all syllables which are signalled by at least two different cues, regardless of the value of these cues. Thus, by employing the conjunction of cues we can give a higher weight to a group of cues which, by appearing together, mark more reliably the presence of a boundary, in the hope that it would increase recall without decreasing too much the precision.

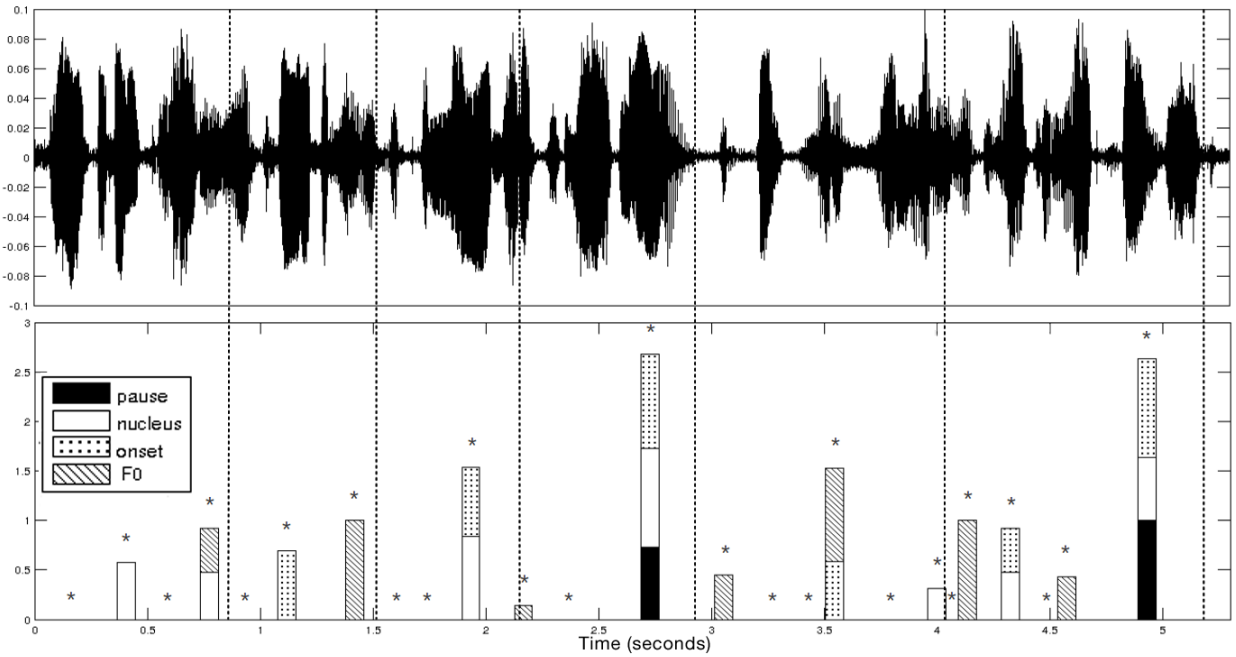

Figure 1: Speech waveform and corresponding detector function employed for prosodic boundary detection of the phrase: "My tape machine records well, but the knobs are too small, the buttons are flimsy and the counter misplaced" (for details, see (Ludusan and Dupoux, 2014)).

The parameters of the algorithm: the combination of cues, the cut-off threshold and the combination of conjunction of cues are obtained on a hold-out set, by aiming to maximize the performance of the system on that particular set.

The prosodic boundary detection procedure is illustrated in Figure 1 for the following utterance: "My tape machine records well, but the knobs are too small, the buttons are flimsy and the counter misplaced". The waveform of the speech signal is shown in the upper panel, with prosodic boundaries marked with dashed lines. In the lower panel are the values of the computed detector function, for each syllable, and the contribution of each of the cues towards the value of the function (the asterisk denotes the position of the syllable nucleus). The syllables corresponding to local maxima of the detector function (syllables 2, 4, 7, 10, 13, 17, 19, 21 and 25) would be considered as possible candidates for the position of a prosodic boundary. Provided that their value is higher than the decision threshold, they will be marked as actual boundaries. For example, if the threshold is set to the first percentile of the function, all the candidates will be kept, for the 50th percentile only syllables 10, 13, 17 and 25 will be considered, while a threshold equal to the value of the 100th percentile will leave only syllable 13 to be marked as a boundary. If we also use what we called conjunction of cues, and we set the cues to be the nucleus and the onset, syllables 10, 13, 22 and 25 will be marked as boundary placeholders, regardless of the fact they are or not a local maximum or they pass or not over the decision threshold.

3 Word segmentation models

3.1 Adaptor grammars

Adaptor Grammars (AGs) are an extension of probabilistic context-free grammars (PCFGs) that learn probability of entire subtrees as well as probabilities of rules (Johnson et al., 2007). A PCFG (N, W, R, S, θ) consists of a start symbol S, N and W disjoints sets of nonterminals and terminal symbols respectively. R is a set of rules producing elements of N or W. Finally, θ is a set of distributions over the rules $R_X, \forall X \in N$ (R_X are the rules that expand X). An AG $(N, W, R, S, \theta, A, C)$ extends the above PCFG with a subset ($A \subseteq N$) of adapted nonterminals, each of them ($X \in A$) having an associated adaptor ($C_X \in C$). An AG defines a dis-

tribution over trees $G_X, \forall X \in N \cup W$. If $X \notin A$, then G_X is defined exactly as for a PCFG:

$$G_X = \sum_{\substack{X \to Y_1 \ldots Y_n \\ \in R_X}} \theta_{X \to Y_1 \ldots Y_n} \mathrm{TD}_X(G_{Y_1} \ldots G_{Y_n})$$

With $\mathrm{TD}_X(G_1 \ldots G_n)$ the distribution over trees with root node X and each subtree $t_i \sim G_i$ i.i.d. If $X \in A$, then there is an additional indirection (composition) with the distribution H_X:

$$G_X = \sum_{\substack{X \to Y_1 \ldots Y_n \\ \in R_X}} \theta_{X \to Y_1 \ldots Y_n} \mathrm{TD}_X(H_{Y_1} \ldots H_{Y_n})$$

$$H_X \sim C_X(G_X)$$

We used C_X adaptors following the Pitman-Yor process (PYP) (Perman et al., 1992; Teh, 2006) with parameters a and b. The PYP generates (Zipfian) type frequencies that are similar to those that occur in natural language (Goldwater et al., 2011). Metaphorically, if there are n customers and m tables, the $n+1$th customer is assigned to table z_{n+1} according to (δ_k is the Kronecker delta function):

$$z_{n+1}|z_1 \ldots z_n \sim \frac{ma+b}{n+b}\delta_{m+1} + \sum_{k=1}^{m} \frac{n_k - a}{n+b}\delta_k$$

For an AG, this means that adapted non-terminals ($X \in A$) either expand to a previously generated subtree ($T(X)_k$) with probability proportional to how often it was visited (n_k), or to a new subtree ($T(X)_{m+1}$) generated through the PCFG with probability proportional to $ma + b$.

3.2 Grammars including prosodic information

The *baseline* that we are using is commonly called the "Colloc3-Syll" model (Johnson and Goldwater, 2009) and is reported at 87% token F-score on the standard Brent version of the Bernstein-Ratner corpus corpus. It posits that sentences are composed of 3 hierarchical levels of collocations, the lower level being collocations of words, and words are composed of syllables. Goldwater et al. (2009) showed how an assumption of independence between words (a unigram model) led to under-segmentation. So, above the $Word$ level, we take the collocations (co-occurring sequences) of words into account.

$$\begin{aligned}
Sentence &\to Colloc3^+ \\
\underline{Colloc3} &\to Colloc2^+ \\
\underline{Colloc2} &\to Colloc1^+ \\
\underline{Colloc1} &\to Word^+ \\
\underline{Word} &\to StructSyll
\end{aligned}$$

where the rule $\underline{Colloc2} \to Colloc1^+$ is implemented by:

$$\begin{aligned}
\underline{Colloc2} &\to Collocs1 \\
Collocs1 &\to Colloc1 \\
Collocs1 &\to Colloc1\ Collocs1
\end{aligned}$$

$Word$ splits into general syllables and initial- or final- specific syllables in $StructSyll$. In English, syllables consist of onsets or codas (producing consonants), and nuclei (vowels). Onsets, nuclei and codas are adapted, thus allowing this model to memorize sequences or consonants or sequences of vowels, dependent on their position in the word. Consonants and vowels are the pre-terminals, their derivation is specified in the grammar into phonemes of the language. In Japanese, syllables are adapted and are composed either of (Consonant-)Vowel(-Nasal) or Nasal. Phonemes are annotated either as consonant, vowel, or nasal (the moraic nasal /N/).

To allow for these grammars to use the prosodic information, we modify them so that prosodic boundaries are considered as breaks at a given level of collocations (or words). For instance we describe below how we change a Colloc3-Syll grammar to make use of the prosodic boundaries information at the lower level of collocations (Colloc1), by using the terminal symbols "|" (the rest is unchanged):

$$\begin{aligned}
\underline{Colloc2} &\to Collocs1 \\
Collocs1 &\to Colloc1 \\
Collocs1 &\to Colloc1 \mid Collocs1 \\
Collocs1 &\to Colloc1\ Collocs1 \\
\underline{Colloc1} &\to Word^+
\end{aligned}$$

We produced and tested grammars which incorporated these prosodic boundary annotations at different levels, from $Collocs3$ down to $Word$ level.

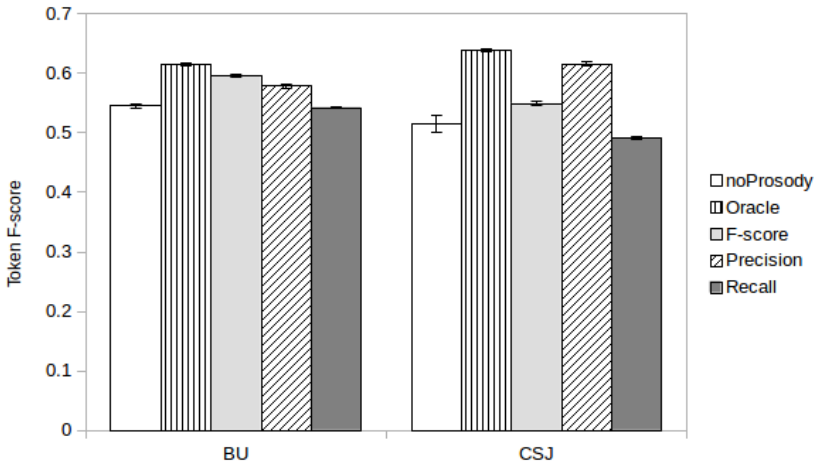

Figure 2: Colloc3-Syll based grammars scores on the BU and CSJ datasets. We show the best results without prosodic annotation, with hand-annotated prosody information (oracle), and with automatically derived annotations that maximize either F-score, precision, or recall of prosodic boundaries.

4 Materials

The experiments were performed on two distinct languages: English and Japanese. For English, we have chosen the Boston University radio news (BU) corpus (Ostendorf et al., 1995) and the LUCID corpus (Baker and Hazan, 2010). The first one, the BU corpus, consists of broadcast news recorded by professional speakers and is widely used in speech prosody research. Here, we only used the prosody annotated portion of the corpus, containing about 3 hours of recordings, labelled for accent tones and prosodic breaks following the ToBI standard for American English (Silverman et al., 1992). Level 3 and level 4 break indices, corresponding to intermediate and intonational phrase boundaries, were considered in this work. The recordings belonging to 6 speakers were used for the experiments, while those belonging to one speaker were employed as a development set, for setting the parameters of the automatic boundary detection algorithm. The evaluation set was divided into utterances, at pauses longer or equal to 200 ms, giving in total 2,273 utterances having 27,980 tokens.

While the BU corpus has the advantage of being annotated for prosodic boundaries, and thus being able to provide us with an upper bound of the performance increase that the prosodic information could bring, it is not large enough to give state-of-the-art results using AG. For this, we have taken a large corpus of spontaneous interactions, the LUCID corpus, and used it in connection to automatically detected prosodic boundaries. Due to the more spontaneous nature of these materials, we have defined utterances as being stretches of speech bounded by pauses at least 500 ms long. Since durational information is needed for the detection of the prosodic boundaries, the corpus was force aligned using the UPenn aligner (Yuan and Liberman, 2008). From the utterances obtained we have excluded all utterances containing hesitations or words not present in the dictionary of the aligner. Thus, a total of 21,649 utterances were eventually used in the experiments, corresponding to 118,640 tokens.

For Japanese, a subpart of the core of the Corpus of Spontaneous Japanese (CSJ) was used (Maekawa, 2003). It contains more than 18 hours of academic recordings from 70 speakers and it was annotated for prosodic boundaries using the X-JToBI standard (Maekawa et al., 2002). Oracle level 2 and level 3 prosodic breaks (accentual and intonational phrases) were used in this study as well as automatically obtained boundaries. The data set aside for the setting of parameters belongs to 5 speakers, with the recordings of the rest of the speakers used for the evaluation. We used the utterance markings provided with the corpus, the evaluation set containing 21,974 utterances and 195,744 tokens.

While previous studies on word segmentation have focused on infant-directed speech (IDS), we employ here corpora of adult-directed speech. The reason behind this choice is the fact that IDS corpora

Model	F-score	Precision	Recall
maxFscore	**.608**	.705	.535
maxPrecision	.391	**.986**	.244
maxRecall	.496	.377	**.724**

Table 1: Automatic prosodic boundary annotation performance on the BU corpus.

are not, generally, annotated for prosody. We would expect that experiments on ADS would improve less over the baseline, when compared to those run on IDS, due to its less exaggerated prosody and its reduced number of prosodic boundaries. Thus, any improvement found on ADS, would be found also on IDS.

The corpora used have all been transcribed phonetically, but, for the purpose of this paper, we have transformed this phonetic annotation into a phonemic one. For the English databases the mappings proposed by Lee and Hon (1989) were employed, with two notable exceptions: vowels /er/ and /axr/ were mapped to the phonemes /ah/ and /r/, while the syllabic consonants /el/, /em/ and /en/ were mapped to the label /ah/ and their corresponding consonant (/l/, /m/ or /n/). For Japanese, we employed the same mappings used by Boruta (2011).

5 Results

The prosodic boundary procedure on the BU and the CSJ used oracle segmental (phonetic) information, while phonemes were force-aligned from word-level annotation for the LUCID. The prosodic boundaries were evaluated with the classic measurements: precision, recall and F-score. The word segmentation token F-scores were obtained every 10 epochs (for less correlation due to the sampler) during the 100 epochs (BU corpus), or the 200 epochs (LUCID and CSJ corpora) centered around the point of convergence, and their mean and standard deviation computed. The convergence point was determined by smoothing the prior probability of the grammar with a sliding window and choosing the epoch where the negative log probability was the lowest.

5.1 English

The best parameters of the prosodic boundary detection system were searched for on the development set left aside for this purpose. The F-score of the system was maximized and the best combination of cues and conjunction of cues were $pause{+}onset$ and $pause + nucleus$, respectively. For these settings, we then determined the threshold values which gave the best F-score, precision and recall for boundary detection, which were further used to run the algorithm on the evaluation set. The results obtained on the evaluation set for the systems trying to maximize F-score ($maxFscore$), precision ($maxPrecision$) or recall ($maxRecall$) are presented in Table 1.

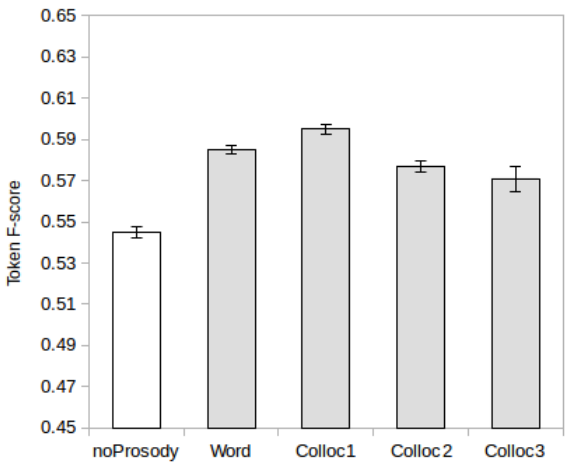

Figure 3: Colloc3-Syll based grammars scores on the BU dataset, comparing results without prosodic annotation, with those obtained by automatic prosodic boundaries that maximize F-score, added at different levels in the grammar.

The word segmentation method was then run with the grammars defined in section 3.2, with and without prosodic boundary information. For the prosody enhanced cases, both oracle and automatic boundaries were employed. The best results obtained on the BU corpus, for each of the five settings, are illustrated on the left side of Figure 2. It appears that all cases that employ prosodic information improve over the baseline, with oracle boundaries giving a 7% absolute performance gain.

Next, we looked in more detail at the behaviour of the best system that uses automatic boundaries ($maxFscore$). We present the token F-score obtained by this system for the different levels of the grammar where the prosodic information is added.

Although we obtained improvements on the BU corpus, for all cases when prosodic information was used, the overall results are far from state-of-the art

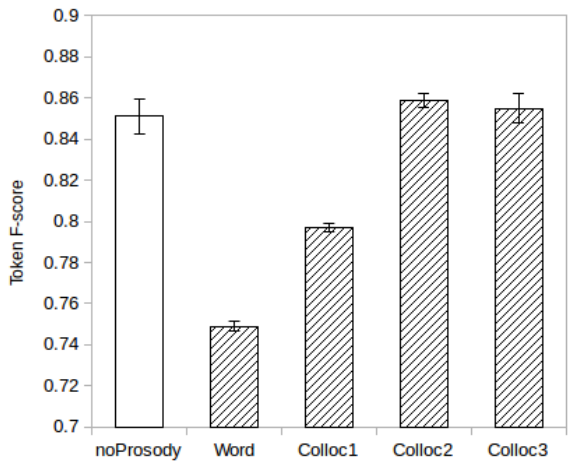

Figure 4: Colloc3-Syll based grammars scores on the LUCID dataset, comparing results without prosodic annotation, with those obtained by automatic prosodic boundaries that maximize precision, added at different levels in the grammar.

performance, due to the relatively small size of the corpus. For this reason, we chose to test on a bigger English corpus, LUCID. While this corpus is indeed larger, it has the disadvantage of not being prosodically annotated. Thus, we investigated only the cases when automatically determined prosodic boundaries are employed. The detection of prosodic boundaries used the same parameters obtained on the BU corpus but, since no prosodic annotation exists, we were not able to perform the same evaluation of the boundaries, as we did for BU.

The token F-scores for the best prosodic boundary setting ($maxPrecision$) are displayed in Figure 4. These results are closer to the state-of-the-art for English, which stand at 87% token F-score. Contrary to the results on the BU corpus, the prosody enhanced system improves over the baseline only when the boundary information is added at $Colloc2$ or $Colloc3$ level (best gain: 0.8% absolute value). While the improvements brought here tend to be quite small, compared to those obtained for BU, we are closer to ceiling value on LUCID and also the quality of the automatic boundaries might be lower, due to the different type of speech on which the parameters of the model were found.

With the Adaptor Grammar tending to slightly over-segment the results, the inclusion of prosody at $Word$ or $Colloc1$ has increased the precision slightly, at the expense of a significantly lower recall, and thus a lower overall F-score. This over-segmentation trait was instead much more pronounced for the BU corpus, where the increase in precision was accompanied only by a slight decrease in recall, brought the two measures closer together, and thus has maximized the F-score.

Model	F-score	Precision	Recall
maxFscore	**.469**	.533	.418
maxPrecision	.398	**.781**	.267
maxRecall	.431	.353	**.552**

Table 2: Automatic prosodic boundary annotation performance on the CSJ corpus.

5.2 Japanese

The same procedure for parameter detection as for the BU corpus was applied and the best cues obtained were $pause + onset$, while the best combination of conjunction of cues was $pause+f0Reset$. Table 2 illustrates the prosodic boundary results obtained on the CSJ evaluation set, for the systems maximizing F-score, precision and recall, respectively.

Since oracle prosodic information was available for this corpus, we were able to compare the performance of the baseline to that of the oracle and automatic boundaries enhanced system. This comparison is displayed in Figure 2, right hand side. Having a sizable corpus, the results are more similar to the state-of-the-art for Japanese, reported in (Fourtassi et al., 2013) (55%). Increases in performance can be observed when hand-labelled prosody is introduced (12.3% absolute value), and also when automatic boundaries ($maxPrecision$) are employed (10% absolute value).

Similarly to the previous experiments, we display in Figure 5 the comparison between the baseline and the best system employing automatic boundaries ($maxPrecision$), for the different levels where the information is added. It shows that prosody helps, regardless of the level where prosody is used, although it appears to favour the lower collocation levels.

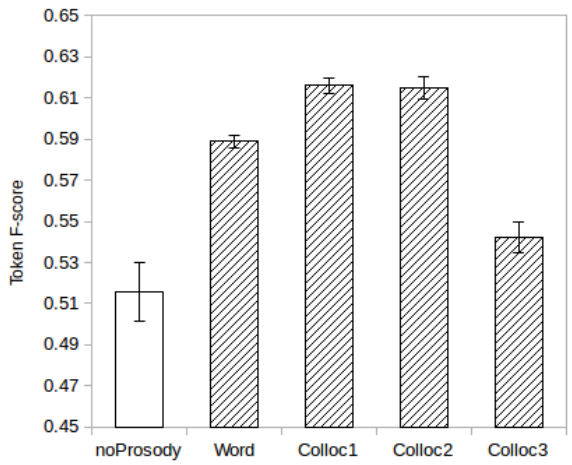

Figure 5: Colloc3-Syll based grammars scores on the CSJ dataset, comparing results without prosodic annotation, with those obtained by automatic prosodic boundaries that maximize precision, added at different levels in the grammar.

6 Discussion and Conclusions

We have investigated the use of prosodic boundary information for unsupervised word discovery in a multilingual setting. We showed that prosodic boundaries can improve word segmentation across both languages even when automatically determined boundaries are used. We also illustrated that the way in which to integrate prosody into word segmentation is not homogeneous across corpora, both in terms of the level of collocation where these boundaries are introduced, and in terms of the balance between precision and recall, when it comes to using automatic boundaries.

For the first issue, the results on BU suggest that $Word$ or $Colloc1$ would be the best level, those on LUCID show that either $Colloc2$ or $Colloc3$ would give the best performance, while the scores on CSJ favors $Colloc1$ or $Colloc2$. But, if we were to discard the results on BU, due to its heavy over-segmentation and its small size, and use the collocation level giving the most balanced scores on the other two datasets, it appears that $Colloc2$ would be the common denominator. Besides giving the most balanced token scores it also gives the most balanced boundary scores, striking a good compromise between the under-segmentation produced by adding the prosody at lower levels and the over-segmentation tendency for boundaries introduced at higher levels.

To investigate the second issue, a closer look to the tables presenting the evaluation of the automatic boundaries (Table 1 and Table 2) is needed. The best word segmentation scores on BU were obtained for the $maxFscore$ system, but we can observe that the condition also has a high precision (.705). At the same time, the best score on CSJ was obtained for the $maxPrecision$ system, the $maxFscore$ system (with a precision of .533) giving no improvement over the baseline (see Figure 2). Furthermore, $maxRecall$, which has very low precisions, seems to behave similar to, or below the baseline, for both datasets. Thus, it appears that a relatively high precision for the prosodic boundaries is needed to obtain improvements in word segmentation and, once this condition is fulfilled, any increase in recall would increase the gain over the baseline.

Further evidence supporting this can be found when performing a word-based evaluation of the automatic prosodic boundaries obtained. For the BU and CSJ corpora, we computed the percentage of word boundaries found, out of the total word boundaries in the corpora, and the proportion of incorrect word boundaries from the total number of boundaries found (see Table 3). It shows that the systems that bring improvements over the baseline ($maxFscore$ and $maxPrecision$ for BU, and $maxPrecision$ for CSJ) have a relatively low rate of false alarms (lower than 6%). At the same time, the increase in performance can be obtained even without a high coverage of the corpus, the $maxPrecision$ models achieving this with a coverage lower than 10%.

Since all the resuls reported in this paper were obtained using the state-of-the-art Adaptor Grammar model, $Colloc3-Syll$, we also verified that our results are generalizable across different models. We created several AG models, by varying the following settings in the grammar: using either one or three collocation levels, and having knowledge or not of the syllabic structure. This gave us, besides the already tested $Colloc3-Syll$ model, three new models: $Colloc3-noSyll$, $Colloc-Syll$ and $Colloc-noSyll$, which were all tested on the CSJ. When evaluating the token F-score obtained using these models, we can see improvements for all the models, regardless of the nature of the prosodic

Corpus	Model	% found	% incorr
BU	oracle	100	0
	maxPrecision	7.0	0.1
	maxFscore	20.3	5.7
	maxRecall	40.4	34.2
CSJ	oracle	100	0
	maxPrec	9.9	0.04
	maxFscore	21.0	23.5
	maxRecall	32.8	51.3

Table 3: Word boundary-based evaluation of the three systems used for prosodic boundary detection. We report the percentage of correct word boundaries found and the number of incorrect boundaries found, as a percentage of all boundaries found.

boundaries used.

Before closing, we note that prosody seem to helps differentially the segmentation of the two languages we tested. In Japanese we found improvements reaching 10 percentage points in F-score, whereas the improvements in English were more modest (5 points for the BU, 1 point for the LUCID), when automatic boundaries are used. This could be due to differences in the segmentation problem across these two languages. Indeed, words in Japanese are in their majority composed of several syllables, and many words contain embedded words, making the segmentation problem intrinsically more difficult than in English, for which the large majority of words are monosyllabic (Fourtassi et al., 2013). It is possible that prosody particularly helps those languages with a polysyllabic lexicon, by helping prevent over-segmentation.

While the current work examined the use of prosodic boundaries for word segmentation in two languages, we would like to extend the study to more languages. We would expect a similar behaviour also for other languages, but it would be interesting to investigate the interaction between boundary information and collocation level for other typologically distinct languages. Also, we have employed here oracle segmental information for the automatic detection of prosodic boundaries. In the future we plan to completely automatize the process, by employing segmental durations obtained with signal-based methods for speech segmentation. Finally, prosody was introduced here by way of a discrete symbol, forcing us to make a binary decision. A more integrated model would enable to associate prosodic break with a probability distribution, over acoustic features, thereby achieving the joint learning of segmentation and prosody.

Acknowledgments

The authors would like to thank the three anonymous reviewers for their insightful comments. The research leading to these results was funded by the European Research Council (ERC-2011-AdG-295810 BOOTPHON). It was also supported by the Agence Nationale pour la Recherche (ANR-10-LABX-0087 IEC, ANR-10-IDEX-0001-02 PSL*), the Fondation de France, the École des Neurosciences de Paris, and the Région Île-de-France (DIM cerveau et pensée).

References

Sankaranarayanan Ananthakrishnan and Shrikanth Narayanan. 2008. Automatic prosodic event detection using acoustic, lexical, and syntactic evidence. *Audio, Speech, and Language Processing, IEEE Transactions on*, 16(1):216–228.

Rachel Baker and Valerie Hazan. 2010. LUCID: a corpus of spontaneous and read clear speech in British English. In *Proceedings of DiSS-LPSS Joint Workshop*, pages 3–6.

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.

Luc Boruta. 2011. *Indicators of allophony and phonemehood*. Ph.D. thesis, Paris-Diderot University.

Michael Brent and Timothy Cartwright. 1996. Distributional regularity and phonotactics are useful for segmentation. *Cognition*, 61:3–125.

Anne Christophe and Emmanuel Dupoux. 1996. Bootstrapping lexical acquisition: The role of prosodic structure. *The Linguistic Review*, 13(3-4):383–412.

Anne Christophe, Emmanuel Dupoux, Josiane Bertoncini, and Jacques Mehler. 1994. Do infants perceive word boundaries? An empirical study of the bootstrapping of lexical acquisition. *Journal of the Acoustical Society of America*, 95:1570–1580.

Anne Christophe, Teresa Guasti, Marina Nespor, Emmanuel Dupoux, and Brit van Ooyen. 1997. Reflections on prosodic bootstrapping: its role for lexical and syntactic acquisition. *Language and Cognitive Processes*, 12:585–612.

Anne Christophe, Jacques Mehler, and Núria Sebastián-Gallés. 2001. Perception of prosodic boundary correlates by newborn infants. *Infancy*, 2(3):385–394.

Abdellah Fourtassi, Benjamin Börschinger, Mark Johnson, and Emmanuel Dupoux. 2013. Whyisenglishsoeasytosegment? In *CMCL 2013*.

LouAnn Gerken, Peter Jusczyk, and Denise Mandel. 1994. When prosody fails to cue syntactic structure: 9-month-olds' sensitivity to phonological versus syntactic phrases. *Cognition*, 51(3):237–265.

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

Sharon Goldwater, Thomas Griffiths, and Mark Johnson. 2011. Producing power-law distributions and damping word frequencies with two-stage language models. *Journal of Machine Learning Research*, 12:2335–2382.

Ariel Gout, Anne Christophe, and James Morgan. 2004. Phonological phrase boundaries constrain lexical access II. Infant data. *Journal of Memory and Language*, 51(4):548–567.

Jui-Ting Huang, Mark Hasegawa-Johnson, and Chilin Shih. 2008. Unsupervised prosodic break detection in Mandarin speech. In *Proc. of Speech Prosody*, pages 165–168.

Je Hun Jeon and Yang Liu. 2009. Semi-supervised learning for automatic prosodic event detection using co-training algorithm. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 540–548.

Mark Johnson 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*, 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.

Mark Johnson, Anne Christophe, Emmanuel Dupoux, and Katherine Demuth. 2014. Modelling function words improves unsupervised word segmentation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 282–292.

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

Peter Jusczyk, Kathy Hirsh-Pasek, Deborah Kemler-Nelson, Lori Kennedy, Amanda Woodward, and Julie Piwoz. 1992. Perception of acoustic correlates of major phrasal units by young infants. *Cognitive Psychology*, 24(2):252–293.

Kai-Fu Lee and Hsiao-Wuen Hon. 1989. Speaker-independent phone recognition using hidden Markov models. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 37(11):1641–1648.

Bogdan Ludusan and Emmanuel Dupoux. 2014. Towards low-resource prosodic boundary detection. In *Proceedings of SLTU*, pages 231–237.

Bogdan Ludusan, Guillaume Gravier, and Emmanuel Dupoux. 2014. Incorporating prosodic boundaries in unsupervised term discovery. In *Proceedings of Speech Prosody*, pages 939–943.

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.

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

Mari Ostendorf, Patti Price, and Stefanie Shattuck-Hufnagel. 1995. The Boston University radio news corpus. *Linguistic Data Consortium*, pages 1–19.

Mihael Perman, Jim Pitman, and Marc Yor. 1992. Size-biased sampling of Poisson point processes and excursions. *Probability Theory and Related Fields*, 92(1):21–39.

Amanda Seidl. 2007. Infants use and weighting of prosodic cues in clause segmentation. *Journal of Memory and Language*, 57:24–48.

Kim Silverman, Mary Beckman, John Pitrelli, Mari Ostendorf, Colin Wightman, Patti Price, Janet Pierrehumbert, and Julia Hirschberg. 1992. TOBI: a standard for labeling English prosody. In *Proceedings of ICSLP*, pages 867–870.

Gabriel Synnaeve, Isabelle Dautriche, Benjamin Börschinger, Mark Johnson, and Emmanuel Dupoux. 2014. Unsupervised word segmentation in context. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 2326–2334.

Yee Whye Teh. 2006. A hierarchical Bayesian language model based on Pitman-Yor processes. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 985–992.

Caroline Wellmann, Julia Holzgrefe, Hubert Truckenbrodt, Isabell Wartenburger, and Barbara Höhle. 2012. How each prosodic boundary cue matters: evidence from German infants. *Frontiers in psychology*, 3.

Colin Wightman and Mari Ostendorf. 1991. Automatic recognition of prosodic phrases. In *Proceedings of Acoustics, Speech, and Signal Processing, 1991 International Conference on*, pages 321–324.

Jiahong Yuan and Mark Liberman. 2008. Speaker identification on the SCOTUS corpus. In *Proceedings of Acoustics '08*.