Whyisenglishsoeasytosegment?

Abdellah Fourtassi[1], Benjamin Börschinger[2,3]
Mark Johnson[3] and Emmanuel Dupoux[1]
(1) Laboratoire de Sciences Cognitives et Psycholinguistique, ENS/EHESS/CNRS, Paris
(2) Department of Computing, Macquarie University
(3) Department of Computational Linguistics, Heidelberg University
{abdellah.fourtassi, emmanuel.dupoux}@gmail.com , {benjamin.borschinger, mark.johnson}@mq.edu.au

Abstract

Cross-linguistic studies on unsupervised word segmentation have consistently shown that English is easier to segment than other languages. In this paper, we propose an explanation of this finding based on the notion of segmentation ambiguity. We show that English has a very low segmentation ambiguity compared to Japanese and that this difference correlates with the segmentation performance in a unigram model. We suggest that segmentation ambiguity is linked to a trade-off between syllable structure complexity and word length distribution.

1 Introduction

During the course of language acquisition, infants must learn to segment words from continuous speech. Experimental studies show that they start doing so from around 7.5 months of age (Jusczyk and Aslin, 1995). Further studies indicate that infants are sensitive to a number of word boundary cues, like prosody (Jusczyk et al., 1999; Mattys et al., 1999), transition probabilities (Saffran et al., 1996; Pelucchi et al., 2009), phonotactics (Mattys et al., 2001), coarticulation (Johnson and Jusczyk, 2001) and combine these cues with different weights (Weiss et al., 2010).

Computational models of word segmentation have played a major role in assessing the relevance and reliability of different statistical cues present in the speech input. Some of these models focus mainly on *boundary detection*, and assess different strategies to identify them (Christiansen et al., 1998; Xanthos, 2004; Swingley, 2005; Daland and Pierrehumbert, 2011). Other models, sometimes called *lexicon-building algorithms*, learn the lexicon and the segmentation at the same time and use knowledge about the extracted lexicon to segment novel utterances. State-of-the-art lexicon-building segmentation algorithms are typically reported to yield better performance than word boundary detection algorithms (Brent, 1999; Venkataraman, 2001; Batchelder, 2002; Goldwater, 2007; Johnson, 2008b; Fleck, 2008; Blanchard et al., 2010).

As seen in Table 1, however, the performance varies considerably across languages with English winning by a high margin. This raises a generalizability issue for NLP applications, but also for the modeling of language acquisition since, obviously, it is not the case that in some languages, infants fail to acquire an adult lexicon. Are these performance differences only due to the fact that the algorithms might be optimized for English? Or do they also reflect some intrinsic linguistic differences between languages?

Lang.	F-score	Model	Reference
English	0.89	AG	Johnson (2009)
Chinese	0.77	AG	Johnson (2010)
Spanish	0.58	DP Bigram	Fleck (2008)
Arabic	0.56	WordEnds	Fleck (2008)
Sesotho	0.55	AG	Johnson (2008)
Japanese	0.55	BootLex	Batchelder (2002)
French	0.54	NGS-u	Boruta (2011)

Table 1: State-of-the-art unsupervised segmentation scores for eight languages.

The aim of the present work is to understand why English usually scores better than other languages, as far as unsupervised segmentation is concerned. As a comparison point, we chose Japanese because it is among the languages that have given the poorest word segmentation scores. In fact, Boruta et al. (2011) found an F-score around 0.41 using both Brent (1999)'s MBDP-1 and Venkataraman (2001)'s NGS-u models, and Batchelder (2002) found an F-score that goes from 0.40 to 0.55 depending on the corpus used. Japanese also differs typologically from English along several phonological dimensions such as

Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics, pages 1–10,
Sofia, Bulgaria, August 8, 2013. ©2013 Association for Computational Linguistics

number of syllabic types, phonotactic constraints and rhythmic structure. Although most lexicon-building segmentation algorithms do not attempt to model these dimensions, they still might be relevant to speech segmentation and help explain the performance difference.

The structure of the paper is as follows. First, we present the class of lexical-building segmentation algorithm that we use in this paper (Adaptor Grammar), and our English and Japanese corpora. We then present data replicating the basic finding that segmentation performance is better for English than for Japanese. We then explore the hypothesis that this finding is due to an intrinsic difference in segmentation ambiguity in the two languages, and suggest that the source of this difference rests in the structure of the phonological lexicon in the two languages. Finally, we use these insights to try and reduce the gap between Japanese and English segmentation through a modification of the Unigram model where multiple linguistic levels are learned jointly.

2 Computational Framework and Corpora

2.1 Adaptor Grammar

In this study, we use the Adaptor Grammar framework (Johnson et al., 2007) to test different models of word segmentation on English and Japanese Corpora. This framework makes it possible to express a class of hierarchical non-parametric Bayesian models using an extension of probabilistic context-free grammars called Adaptor Grammar (AG). It allows one to easily define models that incorporate different assumptions about linguistic structure and is therefore a useful practical tool for exploring different hypotheses about word segmentation (Johnson, 2008b; Johnson, 2008a; Johnson et al., 2010; Börschinger et al., 2012).

For mathematical details and a description of the inference procedure for AGs, we refer the reader to Johnson et al. (2007). Briefly, AG uses the non-parametric Pitman-Yor-Process (Pitman and Yor, 1997) which, as in Minimum Description lengths models, finds a compact representation of the input by re-using frequent structures (here, words).

2.2 Corpora

In the present study, we used both Child Directed Speech (CDS) and Adult Directed Speech (ADS) corpora. English CDS was derived from the Bernstein-Ratner corpus (Bernstein-Ratner, 1987), which consists in transcribed verbal interaction of parents with nine children between 1 and 2 years of age. We used the 9,790 utterances that were phonemically transcribed by Brent and Cartwright (1996). Japanese CDS consists in the first 10, 000 utterances of the Hamasaki corpus (Hamasaki, 2002). It provides a phonemic transcript of spontaneous speech to a single child collected from when the child was 2 up to when it was 3.5 years old. Both CDS corpora are available from the CHILDES database (MacWhinney, 2000).

As for English ADS, we used the first 10,000 utterances of the Buckeye Speech Corpus (Pitt et al., 2007) which consists in spontaneous conversations with 40 speakers in American English. To make it comparable to the other corpora in this paper, we only used the idealized phonemic transcription. Finally, for Japanese ADS, we used the first 10,000 utterances of a phonemic transcription of the Corpus of Spontaneous Japanese (Maekawa et al., 2000). It consists of recorded spontaneous conversations, or public speeches in different fields ranging from engineering to humanities. For each corpus, we present elementary statistics in Table 2.

3 Unsupervised segmentation with the Unigram Model

3.1 Setup

In this experiment we used the Adaptor Grammar framework to implement a Unigram model of word segmentation (Johnson et al., 2007). This model has been shown to be equivalent to the original MBDP-1 segmentation model (see Goldwater (2007)). The model is defined as:

$$
\begin{aligned}
&Utterance \rightarrow \underline{Word}^{+} \\
&\underline{Word} \rightarrow Phoneme^{+}
\end{aligned}
$$

In the AG framework, an underlined non-terminal indicates that this non-terminal is adapted, i.e. that the AG will cache (and learn probabilities for) entire sub-trees rooted in this non-terminal. Here, *Word* is the only unit that the model effectively learns, and there are no dependencies between the words to be learned. This grammar states that an utterance must be analyzed in terms of one or more Words, where a Word is a

Corpus	Child Directed Speech		Adult Directed Speech	
	English	Japanese	English	Japanese
Tokens				
Utterances	$9,790$	$10,000$	$10,000$	$10,000$
Words	$33,399$	$27,362$	$57,185$	$87,156$
Phonemes	$95,809$	$108,427$	$183,196$	$289,264$
Types				
Words	$1,321$	$2,389$	$3,708$	$4,206$
Phonemes	50	30	44	25
Average Lengths				
Words per utterance	3.41	2.74	5.72	8.72
Phonemes per utterance	9.79	10.84	18.32	28.93
Phonemes per word	2.87	3.96	3.20	3.32

Table 2 : Characteristics of phonemically transcribed corpora

sequence of Phonemes.

We ran the model twice on each corpus for 2,000 iterations with hyper-parameter sampling and we collected samples throughout the process, following the methodology of Johnson and Goldwater (2009)[1]. For evaluation, we performed their Minimum Bayes Risk decoding using the collected samples to get a single score.

3.2 Evaluation

For the evaluation, we used the same measures as Brent (1999), Venkataraman (2001) and Goldwater (2007), namely token Precision (P), Recall (R) and F-score (F). Precision is defined as the number of correct word tokens found out of all tokens posited. Recall is the number of correct word tokens found out of all tokens in the gold standard. The F-score is defined as the harmonic mean of Precision and Recall , $F = \frac{2*P*R}{P+R}$.

We will refer to these scores as the *segmentation* scores. In addition, we define similar measures for word *boundaries* and word types in the *lexicon*.

3.3 Results and discussion

The results are shown in Table 3. As expected, the model yields substantially better scores in English than Japanese, for both CDS and ADS. In addition, we found that in both languages, ADS yields slightly worse results than CDS. This is to be expected because ADS uses between 60% and 300% longer utterances than CDS, and as a result presents the learner with a more difficult segmentation problem. Moreover, ADS includes between 70% and 280% more word types than CDS, making it a more difficult lexical learning problem. Note, however, that despite these large differences in corpus statistics, the difference in segmentation performance between ADS and CDS are small compared to the differences between Japanese and English.

An error analysis on English data shows that most errors come from the Unigram model mistaking high frequency collocations for single words (see also Goldwater (2007)). This leads to an under-segmentation of chunks like “a boy” or “is it” [2]. Yet, the model also tends to break off frequent morphological affixes, especially “-ing” and “-s” , leading to an over-segmentation of words like “talk ing” or “black s”.

Similarly, Japanese data shows both over- and under-segmentation errors. However, over-segmentation is more severe than for English, as it does not only affect affixes, but surfaces as breaking apart multi-syllabic words. In addition, Japanese segmentation faces another kind of error which acts across word boundaries. For example, “ni kashite” is segmented as “nika shite” and “nurete inakatta” as “nure tei na katta”. This leads to an output lexicon that, on the one hand, allows for a more compact analysis of the corpus than the true lexicon: the number of word types drops from 2,389 to 1,463 in CDS and from 4,206 to 2,372 in ADS although the average token length – and consequently, overall number of tokens – does not change as dramatically, dropping from 3.96 to

[1] We used incremental initialization

[2] For ease of presentation, we use orthography to present examples although all experiments are run on phonemic transcripts.

	Child Directed Speech						Adult Directed Speech					
	English			Japanese			English			Japanese		
	F	P	R	F	P	R	F	P	R	F	P	R
Segmentation	**0.77**	0.76	0.77	**0.55**	0.51	0.61	**0.69**	0.66	0.73	**0.50**	0.48	0.52
Boundaries	0.87	0.87	0.88	0.72	0.63	0.83	0.86	0.81	0.91	0.76	0.74	0.79
Lexicon	0.62	0.65	0.59	0.33	0.43	0.26	0.41	0.48	0.36	0.30	0.42	0.23

Table 3 : Word segmentation scores of the Unigram model

3.31 for CDS and from 3.32 to 3.12 in ADS. On the other hand, however, most of the output lexicon items are not valid Japanese words and this leads to the bad lexicon F-scores. This, in turn, leads to the bad overall segmentation performance.

In brief, we have shown that, across two different corpora, English yields consistently better segmentation results than Japanese for the Unigram model. This confirms and extends the results of Boruta et al. (2011) and Batchelder (2002). It strongly suggests that the difference is neither due to a specific choice of model nor to particularities of the corpora, but reflects a fundamental property of these two languages.

In the following section, we introduce the notion of *segmentation ambiguity*, it to English and Japanese data, and show that it correlates with segmentation performance.

4 Intrinsic Segmentation Ambiguity

Lexicon-based segmentation algorithms like MBDP-1, NGS-u and the AG Unigram model learn the lexicon and the segmentation at the same time. This makes it difficult, in case of poor performance, to see whether the problem comes from the intrinsic segmentability of the language or from the quality of the extracted lexicon. Our claim is that Japanese is intrinsically more difficult to segment than English, even when a good lexicon is already assumed. We explore this hypothesis by studying segmentation alone, assuming a perfect (Gold) lexicon.

4.1 Segmentation ambiguity

Without any information, a string of N phonemes could be segmented in 2^{N-1} ways. When a lexicon is provided, the set of possible segmentations is reduced to a smaller number. To illustrate this, suppose we have to segment the input utterance: /ay s k r iy m/ [3], and that the lexicon contains the following words : /ay/ (I), /s k r iy m/ (scream), /ay s/ (ice), /k r iy m/ (cream). Only two segmentations are possible : /ay skriym/ (I scream) and /ays kriym/ (ice cream).

We are interested in the ambiguity generated by the different possible parses that result from such a supervised segmentation. In order to quantify this idea in general, we define a *Normalized Segmentation Entropy*. To do this, we need to assign a probability to every possible segmentation. To this end, we use a unigram model where the probability of a lexical item is its normalized frequency in the corpus and the probability of a parse is the product of the probabilities of its terms. In order to obtain a measure that does not depend on the utterance length, we normalize by the number of possible boundaries in the utterance. So for an utterance of length N, the Normalized Segmentation Entropy (NSE) is computed using Shannon formula (Shannon, 1948) as follows:

$$NSE = -\sum_i P_i log_2(P_i)/(N-1)$$

where P_i is the probability of the parse i .

For CDS data we found Normalized Segmentation Entropies of 0.0021 bits for English and 0.0156 bits for Japanese. In ADS data we found similar results with 0.0032 bits for English and 0.0275 bits for Japanese. This means that Japanese needs between 7 and 8 times more bits than English to encode segmentation information. This is a very large difference, which is of the same magnitude in CDS and ADS. These differences clearly show that intrinsically, Japanese is more ambiguous than English with regards to segmentation.

One can refine this analysis by distinguishing two sources of ambiguity: ambiguity *across word boundaries*, as in ”ice cream / [ay s] [k r iy m]”

[3] We use ARPABET notation to represent phonemic input.

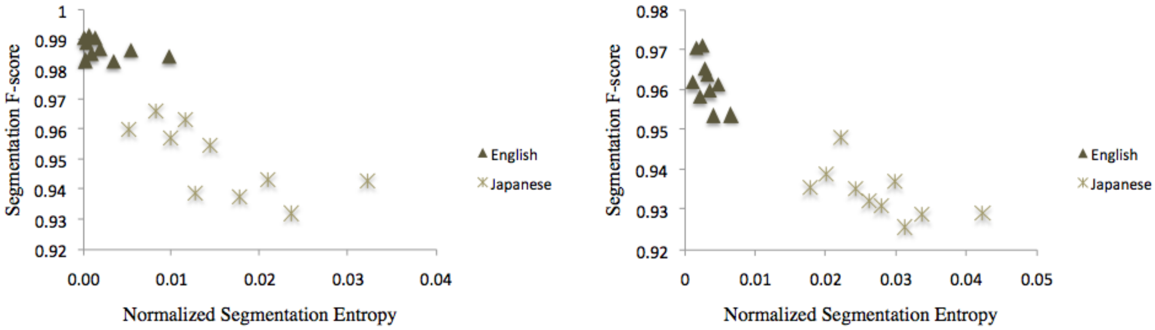

Figure 1 : Correlation between Normalized Segmentation Entropy (in bits) and the segmentation F-score for CDS (left) and ADS (Right)

vs ”I scream / [ay] [s k r iy m]”. And ambiguity *within the lexicon*, that occurs when a lexical item is composed of two or more sub-words (like in “Butterfly”).

Since we are mainly investigating lexicon-building models, it is important to measure the ambiguity within the lexicon itself, in the ideal case where this lexicon is perfect. To this end, we computed the average number of segmentations for a lexicon item. For example, the word “butterfly” has two possible segmentations : the original word “butterfly” and a segmentation comprising the two sub-words : “butter” and “fly”. For English tokens, we found an average of 1.039 in CDS and 1.057 in ADS. For Japanese tokens, we found an average of 1.811 in CDS and 1.978 in ADS. English’s averages are close to 1, indicating that it doesn’t exhibit lexicon ambiguity. Japanese, however, has averages close to 2 which means that lexical ambiguity is quite systematic in both CDS and ADS.

4.2 Segmentation ambiguity and supervised segmentation

The intrinsic ambiguity in Japanese only shows that a given sentence has multiple possible segmentations. What remains to be demonstrated is that these multiple segmentations result in systematic segmentation errors. To do this we propose a supervised segmentation algorithm that enumerates all possible segmentations of an utterance based on the gold lexicon, and selects the segmentation with the highest probability. In CDS data, this algorithm yields a segmentation F-score equal to 0.99 for English and 0.95 for Japanese. In ADS we find an F-score of 0.96 for English and 0.93 for Japanese. These results show that lexical information alone plus word frequency eliminates almost all segmentation errors in English, especially for CDS. As for Japanese, even if the scores remain impressively high, the lexicon alone is not sufficient to eliminate all the errors. In other words, even with a gold lexicon, English remains easier to segment than Japanese.

To quantify the link between segmentation entropy and segmentation errors, we binned the sentences of our corpus in 10 bins according to the Normalized Segmentation Entropy, and correlate this with the average segmentation F-score for each bin. As shown Figure 1, we found significant correlations: ($R = -0.86$, $p < 0.001$) for CDS and ($R = -0.93$, $p < 0.001$) for ADS, showing that segmentation ambiguity has a strong effect even on supervised segmentation scores. The correlation within language was also significant but only in the Japanese data : $R = -0.70$ for CDS and $R = -0.62$ for ADS.

Next, we explore one possible reason for this structural difference between Japanese and English, especially at the level of the lexicon.

4.3 Syllable structure and lexical composition of Japanese and English

One of the most salient differences between English and Japanese phonology concerns their syllable structure. This is illustrated in Figure 2 (above), where we plotted the frequency of the different syllabic structures of monosyllabic tokens in English and Japanese CDS. The statistics show that English has a very rich syllabic composition where a diversity of consonant clusters is allowed, whereas Japanese syllable structure is quite simple and mostly composed of the default CV type. This difference is bound to have an effect on the structure of the lexicon. Indeed, Japanese has to use

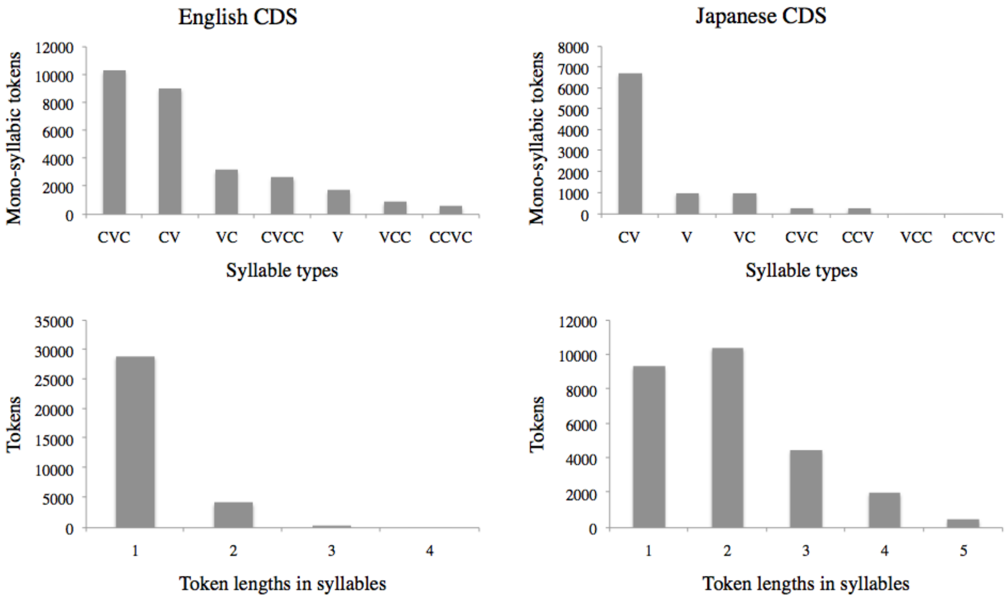

Figure 2 : Trade-off between the complexity of syllable structure (above) and the word token length in terms of syllables (below) for English and Japanese CDS.

multisyllabic words in order to achieve a large size lexicon, whereas, in principle, English could use mostly monosyllables. In Figure 2 (below) we display the distribution of word length as measured in syllables in the two languages for the CDS corpora. The English data is indeed mostly composed of mono-syllabic words whereas the Japanese one is made of words of more varied lengths. Overall, we have documented a trade-off between the diversity of syllable structure on the one hand, and the diversity of word lengths on the other (see Table 4 for a summary of this tradeoff expressed in terms of entropy).

	CDS		ADS	
	Eng.	Jap.	Eng.	Jap.
Syllable types	2.40	1.38	2.58	1.03
Token lengths	0.62	2.04	0.99	1.69

Table 4 : Entropies of syllable types and token lengths in terms of syllables (in bits)

We suggest that this trade-off is responsible for the difference in the lexicon ambiguity across the two languages. Specifically, the combination of a small number of syllable types and, as a consequence, the tendency for multi-syllabic word types in Japanese makes it likely that a long word will be composed of smaller ones. This cannot happen very often in English, since most words are monosyllabic, and words smaller than a syllable are not allowed.

5 Improving Japanese unsupervised segmentation

We showed in the previous section that ambiguity impacts segmentation even with a gold lexicon, mainly because the lexicon itself could be ambiguous. In an unsupervised segmentation setting, the problem is worse because ambiguity within and across word boundaries leads to a bad lexicon, which in turn results in more segmentation errors. In this section, we explore the possibility of mitigating some of these negative consequences.

In section 3, we saw that when the Unigram model tries to learn Japanese words, it produces an output lexicon composed of both over- and under-segmented words in addition to words that result from a segmentation across word boundaries. One way to address this is by learning multiple kinds of units jointly, rather than just words; indeed, previous work has shown that richer models with multiple levels improve segmentation for English (Johnson, 2008a; Johnson and Goldwater, 2009).

5.1 Two dependency levels

As a first step, we will allow the model to not just learn words but to also memorize sequences of words. Johnson (2008a) introduced these units as "collocations" but we choose to use the more neutral notion of *level* for reasons that become clear shortly. Concretely, the grammar is:

	CDS						ADS					
	English			Japanese			English			Japanese		
	F	P	R	F	P	R	F	P	R	F	P	R
Level 1												
Segmentation	**0.81**	0.77	0.86	0.42	0.33	0.55	**0.70**	0.63	0.78	0.42	0.35	0.50
Boundaries	0.91	0.84	0.98	0.63	0.47	0.96	0.86	0.76	0.98	0.73	0.61	0.90
Lexicon	0.64	0.79	0.54	0.18	0.55	0.10	0.36	0.56	0.26	0.15	0.68	0.08
Level 2												
Segmentation	0.33	0.45	0.26	**0.59**	0.65	0.53	0.50	0.60	0.43	**0.45**	0.54	0.38
Boundaries	0.56	0.98	0.40	0.71	0.87	0.60	0.76	0.95	0.64	0.73	0.92	0.60
Lexicon	0.36	0.25	0.59	0.47	0.44	0.49	0.46	0.38	0.56	0.43	0.37	0.50

Table 5 : Word segmentation scores of the two levels model

$$Utterance \rightarrow \underline{level2}^{+}$$
$$\underline{level2} \rightarrow \underline{level1}^{+}$$
$$\underline{level1} \rightarrow Phoneme^{+}$$

We run this model under the same conditions as the Unigram model but evaluate two different situations. The model has no inductive bias that would force it to equate $\underline{level1}$ with words, rather than $\underline{level2}$. Consequently, we evaluate the segmentation that is the result of taking there to be a boundary between every $\underline{level1}$ constituent (Level 1 in Table 5) and between every $\underline{level2}$ constituent (Level 2 in Table 5). From these results , we see that English data has better scores when the lower level represents the Word unit and when the higher level captures regularities above the word. However, Japanese data is best segmented when the higher level is the Word unit and the lower level captures sub-word regularities.

Level 1 generally tends to over-segment utterances as can be seen by comparing the Boundary Recall and Precision scores (Goldwater, 2007). In fact when the Recall is much higher than the Precision, we can say that the model has a tendency to over-segment. Conversely, we see that Level 2 tends to under-segment utterances as the Boundary Precision is higher than the Recall.

Over-segmentation at Level 1 seems to benefit English since it counteracts the tendency of the Unigram model to cluster high frequency collocations. As far as segmentation is concerned, this effect seems to outweigh the negative effect of breaking words apart (especially in CDS), as English words are mostly monosyllabic.

For Japanese, under-segmentation at Level 2 seems to be slightly less harmful than over-segmentation at Level 1, as it prevents, to some extent, multi-syllabic words to be split. However, the scores are not very different from the ones we had with the Unigram model and slightly worse for the ADS. What seems to be missing is an intermediate level where over- and under-segmentation would counteract one another.

5.2 Three dependency levels

We add a third dependency level to our model as follows :

$$Utterance \rightarrow \underline{level3}^{+}$$
$$\underline{level3} \rightarrow \underline{level2}^{+}$$
$$\underline{level2} \rightarrow \underline{level1}^{+}$$
$$\underline{level1} \rightarrow Phoneme^{+}$$

As with the previous model, we test each of the three levels as the word unit, the results are shown in Table 6.

Except for English CDS, all the corpora have their best scores with this intermediate level. Level 1 tends to over-segment Japanese utterances into syllables and English utterances into morphemes. Level 3, however, tends to highly under-segment both languages. English CDS seems to be already under-segmented at Level 2, very likely caused by the large number of word collocations like ”is-it” and ”what-is”, an observation also made by Börschinger et al. (2012) using different English CDS corpora. English ADS is quantitatively more sensitive to over-segmentation than CDS mainly because it has a richer morphological structure and relatively longer words in terms of syllables (Table 4).

	CDS						ADS					
	English			Japanese			English			Japanese		
	F	P	R	F	P	R	F	P	R	F	P	R
Level 1												
Segmentation	**0.79**	0.74	0.85	0.27	0.20	0.41	0.35	0.28	0.48	0.37	0.30	0.47
Boundaries	0.89	0.81	0.99	0.56	0.39	0.99	0.68	0.52	0.99	0.70	0.57	0.93
Lexicon	0.58	0.76	0.46	0.10	0.47	0.05	0.13	0.39	0.07	0.10	0.70	0.05
Level 2												
Segmentation	0.49	0.60	0.42	**0.70**	0.70	0.70	**0.77**	0.76	0.79	**0.60**	0.65	0.55
Boundaries	0.71	0.97	0.56	0.81	0.82	0.81	0.90	0.88	0.92	0.81	0.90	0.74
Lexicon	0.51	0.41	0.64	0.53	0.59	0.47	0.58	0.69	0.50	0.51	0.57	0.46
Level 3												
Segmentation	0.18	0.31	0.12	0.39	0.53	0.30	0.43	0.55	0.36	0.28	0.42	0.21
Boundaries	0.26	0.99	0.15	0.46	0.93	0.31	0.71	0.98	0.55	0.59	0.96	0.43
Lexicon	0.17	0.10	0.38	0.32	0.25	0.41	0.37	0.28	0.51	0.27	0.20	0.42

Table 6 : Word segmentation scores of the three levels model

6 Conclusion

In this paper we identified a property of language, *segmentation ambiguity*, which we quantified through Normalized Segmentation Entropy. We showed that this quantity predicts performance in a supervised segmentation task.

With this tool we found that English was intrinsically less ambiguous than Japanese, accounting for the systematic difference found in this paper. More generally, we suspect that Segmentation Ambiguity would, to some extent, explain much of the difference observed across languages (Table 1). Further work needs to be carried out to test the robustness of this hypothesis on a larger scale.

We showed that allowing the system to learn at multiple levels of structure generally improves performance, and compensates partially for the negative effect of segmentation ambiguity on unsupervised segmentation (where a bad lexicon amplifies the effect of segmentation ambiguity). Yet, we end up with a situation where the best level of structure may not be the same across corpora or languages, which raises the question as to how to determine which level is the correct lexical level, i.e., the level that can sustain successful grammatical and semantic learning. Further research is needed to answer this question.

Generally speaking, ambiguity is a challenge in many speech and language processing tasks: for example part-of-speech tagging and word sense disambiguation tackle lexical ambiguity, probabilistic parsing deals with syntactic ambiguity and speech act interpretation deals with pragmatic ambiguities. However, to our knowledge, ambiguity has rarely been considered as a serious problem in word segmentation tasks.

As we have shown, the lexicon-based approach does not completely solve the segmentation ambiguity problem since the lexicon itself could be more or less ambiguous depending on the language. Evidently, however, infants in all languages manage to overcome this ambiguity. It has to be the case, therefore, that they solve this problem through the use of alternative strategies, for instance by relying on sub-lexical cues (see Jarosz and Johnson (2013)) or by incorporating semantic or syntactic constraints (Johnson et al., 2010). It remains a major challenge to integrate these strategies within a common model that can learn with comparable performance across typologically distinct languages.

Acknowledgements

The research leading to these results has received funding from the European Research Council (FP/2007-2013) / ERC Grant Agreement n. ERC-2011-AdG-295810 BOOTPHON, from the Agence Nationale pour la Recherche (ANR-2010-BLAN-1901-1 BOOTLANG, ANR-11-0001-02 PSL* and ANR-10-LABX-0087) and the Fondation de France. This research was also supported under the Australian Research Council's Discovery Projects funding scheme (project numbers DP110102506 and DP110102593).

References

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

N. Bernstein-Ratner. 1987. The phonology of parent-child speech. In K. Nelson and A. van Kleeck, editors, *Children's Language*, volume 6. Erlbaum, Hillsdale, NJ.

Daniel Blanchard, Jeffrey Heinz, and Roberta Golinkoff. 2010. Modeling the contribution of phonotactic cues to the problem of word segmentation. *Journal of Child Language*, 37(3):487–511.

Benjamin Börschinger, Katherine Demuth, and Mark Johnson. 2012. Studying the effect of input size for Bayesian word segmentation on the Providence corpus. In *Proceedings of the 24th International Conference on Computational Linguistics (Coling 2012)*, pages 325–340, Mumbai, India. Coling 2012 Organizing Committee.

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

M. Brent and T. Cartwright. 1996. Distributional regularity and phonotactic constraints are useful for segmentation. *Cognition*, 61:93–125.

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

Morten H Christiansen, Joseph Allen, and Mark S Seidenberg. 1998. Learning to segment speech using multiple cues: A connectionist model. *Language and cognitive processes*, 13(2-3):221–268.

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

Margaret M. Fleck. 2008. Lexicalized phonotactic word segmentation. In *Proceedings of ACL-08: HLT*, pages 130–138, Columbus, Ohio, June. Association for Computational Linguistics.

Sharon Goldwater. 2007. *Nonparametric Bayesian Models of Lexical Acquisition*. Ph.D. thesis, Brown University.

Naomi Hamasaki. 2002. The timing shift of two-year-olds responses to caretakers yes/no questions. In *Studies in language sciences (2)Papers from the 2nd Annual Conference of the Japanese Society for Language Sciences*, pages 193–206.

Gaja Jarosz and J Alex Johnson. 2013. The richness of distributional cues to word boundaries in speech to young children. *Language Learning and Development*, (ahead-of-print):1–36.

Mark Johnson and Katherine Demuth. 2010. Unsupervised phonemic Chinese word segmentation using Adaptor Grammars. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 528–536, Beijing, China, August. Coling 2010 Organizing Committee.

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, Boulder, Colorado, June. Association for Computational Linguistics.

Elizabeth K. Johnson and Peter W. Jusczyk. 2001. Word segmentation by 8-month-olds: When speech cues count more than statistics. *Journal of Memory and Language*, 44:1–20.

Mark Johnson, Thomas Griffiths, and Sharon Goldwater. 2007. Bayesian inference for PCFGs via Markov chain Monte Carlo. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 139–146, Rochester, New York. Association for Computational Linguistics.

Mark Johnson, Katherine Demuth, Michael Frank, and Bevan Jones. 2010. Synergies in learning words and their referents. In J. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R.S. Zemel, and A. Culotta, editors, *Advances in Neural Information Processing Systems 23*, pages 1018–1026.

Mark Johnson. 2008a. 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, Columbus, Ohio, June. Association for Computational Linguistics.

Mark Johnson. 2008b. Using Adaptor Grammars to identify synergies in the unsupervised acquisition of linguistic structure. In *Proceedings of the 46th Annual Meeting of the Association of Computational Linguistics*, pages 398–406, Columbus, Ohio. Association for Computational Linguistics.

Peter W Jusczyk and Richard N Aslin. 1995. Infants detection of the sound patterns of words in fluent speech. *Cognitive psychology*, 29(1):1–23.

Peter W. Jusczyk, E. A. Hohne, and A. Bauman. 1999. Infants' sensitivity to allophonic cues for word segmentation. *Perception and Psychophysics*, 61:1465–1476.

Brian MacWhinney. 2000. *The CHILDES Project: Tools for Analyzing Talk. Transcription, format and programs*, volume 1. Lawrence Erlbaum.

Kikuo Maekawa, Hanae Koiso, Sadaoki Furui, and Hitoshi Isahara. 2000. Spontaneous speech corpus of japanese. In *proc. LREC*, volume 2, pages 947–952.

Sven L Mattys, Peter W Jusczyk, Paul A Luce, James L Morgan, et al. 1999. Phonotactic and prosodic effects on word segmentation in infants. *Cognitive psychology*, 38(4):465–494.

Sven L Mattys, Peter W Jusczyk, et al. 2001. Do infants segment words or recurring contiguous patterns? *Journal of experimental psychology, human perception and performance*, 27(3):644–655.

Bruna Pelucchi, Jessica F Hay, and Jenny R Saffran. 2009. Learning in reverse: Eight-month-old infants track backward transitional probabilities. *Cognition*, 113(2):244–247.

J. Pitman and M. Yor. 1997. The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator. *Annals of Probability*, 25:855–900.

M. A. Pitt, L. Dilley, K. Johnson, S. Kiesling, W. Raymond, E. Hume, and Fosler-Lussier. 2007. Buckeye corpus of conversational speech.

J. Saffran, R. Aslin, and E. Newport. 1996. Statistical learning by 8-month-old infants. *Science*, 274:1926–1928.

Claude Shannon. 1948. A mathematical theory of communication. *Bell System Technical Journal*, 27(3):379–423.

Daniel Swingley. 2005. Statistical clustering and the contents of the infant vocabulary. *Cognitive Psychology*, 50:86–132.

A. Venkataraman. 2001. A statistical model for word discovery in transcribed speech. *Computational Linguistics*, 27(3):351–372.

Daniel J Weiss, Chip Gerfen, and Aaron D Mitchel. 2010. Colliding cues in word segmentation: the role of cue strength and general cognitive processes. *Language and Cognitive Processes*, 25(3):402–422.

Aris Xanthos. 2004. Combining utterance-boundary and predictability approaches to speech segmentation. In *First Workshop on Psycho-computational Models of Human Language Acquisition*, page 93.