Unsupervised phonemic Chinese word segmentation using Adaptor Grammars

Mark Johnson
Department of Computing
Macquarie University
Mark.Johnson@mq.edu.au

Katherine Demuth
Department of Linguistics
Macquarie University
Katherine.Demuth@mq.edu.au

Abstract

Adaptor grammars are a framework for expressing and performing inference over a variety of non-parametric linguistic models. These models currently provide state-of-the-art performance on unsupervised word segmentation from phonemic representations of child-directed unsegmented English utterances. This paper investigates the applicability of these models to unsupervised word segmentation of Mandarin. We investigate a wide variety of different segmentation models, and show that the best segmentation accuracy is obtained from models that capture inter-word "collocational" dependencies. Surprisingly, enhancing the models to exploit syllable structure regularities and to capture tone information does improve overall word segmentation accuracy, perhaps because the information these elements convey is redundant when compared to the inter-word dependencies.

1 Introduction and previous work

The word-segmentation task is an abstraction of part of the problem facing a child learning its native language. Fluent speech, even the speech directed at children, doesn't come with silence or pauses delineating acoustic words the way that spaces separate orthographic words in writing systems like that of English. Instead, as most people listening to a language they don't understand can attest, words in fluent speech "run together", and a language user needs to learn how to segment utterances of the language they are learning into words. This kind of word segmentation is presumably an important first step in acquiring a language. It is scientifically interesting to know what information might be useful for word segmentation, and just how this information might be used. These scientific questions have motivated a body of research on computational models of word segmentation. Since as far as we can tell any child can learn any human language, our goal is to develop a single model that can learn to perform accurate word segmentation given input from any human language, rather than a model that specialised to perform well on a single language. This paper extends the previous work on word segmentation by investigating whether one class of models that work very well with English input also work with Chinese input. These models will permit us to study the role that syllable structure constraints and tone in Chinese might play in word segmentation.

While learners and fluent speakers undoubtedly use a wide variety of cues to perform word segmentation, computational models since Elman (1990) have tended to focus on the use of phonotactic constraints (e.g., syllable-structure constrains) and distributional information. Brent and Cartwright (1996) introduced the standard form of the word segmentation task still studied today. They extracted the orthographic representations of child-directed speech from the Bernstein-Ratner corpus (Bernstein-Ratner, 1987) and "phonologised" them by looking up each word in a pronouncing dictionary. For example, the orthographic utterance *you want to see the book* is mapped to the sequence of pronunciations *yu want tu si D6 bUk*, (the pronunciations are in an

Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010), pages 528–536,
Beijing, August 2010

ASCII encoding of the International Phonetic Alphabet representation of English phonemes). The input to the learner is obtained by concatenating together the phonemic representations of each utterance's words. The learner's task is to identify the locations of the word boundaries in this sequence, and hence identify the words (up to homophony). Brent and Cartwright (1996) pointed out the importance of both distributional information and phonotactic (e.g., syllable-structure) constraints for word segmentation (see also Swingley (2005) and Fleck (2008)).

Recently there has been considerable interest in applying Bayesian inference techniques for nonparametric models to this problem. Here the term "non-parametric" does not mean that the models have no parameters, rather, it is used to distinguish these models from the usual "parametric models" that have a fixed finite vector of parameters.

Goldwater et al. (2006) introduced two nonparametric Bayesian models of word segmentation, which are discussed in more detail in (Goldwater et al., 2009). The *unigram model*, which assumes that each word is generated independently to form a sentence, turned out to be equivalent to a model originally proposed by Brent (1999). The *bigram model* improves word segmentation accuracy by modelling bigram inter-word contextual dependencies, "explaining away" inter-word dependencies that would otherwise cause the unigram model to under-segment. Mochihashi et al. (2009) showed that segmentation accuracy could be improved by using a more sophisticated "base distribution" and a dynamic programming sampling algorithm very similar to the one used with the adaptor grammars below. They also applied their algorithm to Japanese and Chinese word segmentation, albeit from orthographic rather than phonemic forms, so unfortunately their results are not comparable with ours.

Johnson et al. (2007) introduced *adaptor grammars* as a grammar-based framework for expressing a variety of non-parametric models, and provided a dynamic programming Markov Chain Monte Carlo (MCMC) sampling algorithm for performing Bayesian inference on these models. For example, the unigram model can be expressed as a simple adaptor grammar as shown below, and the generic adaptor grammar inference procedure provides a dynamic programming sampling algorithm for this model. Johnson (2008b) showed how a variety of different word segmentation models can be expressed as adaptor grammars, and Johnson and Goldwater (2009) described a number of extensions and specialisations to the adaptor grammar framework that improve inference speed and accuracy (we use these techniques in our work below).

Previous work on unsupervised word segmentation from phonemic input has tended to concentrate on English. However, presumably children the world over segment their first language input in the same (innately-specified) way, so a correct procedure should work for all possible human languages. However, as far as we are aware there has been relatively little work on word segmentation from phonemic input except on English. Johnson (2008a) investigated whether the adaptor grammars models that do very well on English also apply to Sesotho (a Bantu language spoken in southern Africa with rich agglutinating morphology). He found that the models in general do very poorly (presumably because the adaptor grammars used cannot model the complex morphology found in Sesotho) and that the best segmentation accuracy was considerably worse than that obtained for English, even when that model incorporated some Bantu-specific information about morphology. Of course it may also be that the Sesotho and English corpora are not really comparable: the Bernstein-Ratner corpus that Brent and other researchers have used for English was spoken to pre-linguistic 1-year olds, while most non-English corpora are of child-directed speech to older children who are capable of talking back, and hence these corpora are presumably more complex. We discuss this issue in more detail in section 4 below.

2 A Chinese word segmentation corpus

Our goal here is to prepare a Chinese corpus of child-directed speech that parallels the English one used by Brent and other researchers. That corpus was in broad phonemic form, obtained by looking each word up in a pronouncing dictionary. Here instead we make use of a corpus in Pinyin format, which we translate into a broad phonemic IPA format using the freely-available Pinyin-to-IPA translation program "Pinyin to IPA Conversion Tools" version 2.1 available on http://sourceforge.net/projects/py2ipa.

We used the "Beijing" corpus (Tardif, 1993) available from the publicly-distributed Childes collection of corpora (MacWhinney and Snow, 1985). We are interested in child-directed speech (rather than children's speech), so we removed all utterances from participants with an Id containing "Child". (Tardif (1993) points out that Chinese-speaking children typically have a much richer social environment involving multiple adult caregivers than middle-class English-speaking children do, so we cannot simply collect only the mother's utterances, as was done for the English corpus). We also ignored all utterances with codes $INTERJ, $UNINT, $VOC and $PRMPT, as these are not always linguistic utterances. In addition, we deleted all words that could not be analysed as a sequence of syllables, such as "xxx" and "hmm", and also deleted "cluck". The first few utterances of the corpus in Pinyin format are:

zen3me gei3 ta1 bei1 shang4 lai2 (1.) ?
ta1: (.) a1yi2 gei3 de (.) ta1 gei3 de .
hen3 jian3dan1 .

We then fed these into the Pinyin-to-IPA translation program, producing output of the following format:

tsən^{214}mɤ kei^{214} t^{h}a^{55} pei^{55} ʂɑŋ^{51} lai^{35}
t^{h}a^{55} a^{55}i^{35} kei^{214} tɤ t^{h}a^{55} kei^{214} tɤ
xən^{214} tɕiɛn^{214}tan^{55}

In the IPA format, the superscript indices indicate the tone patterns associated with syllables; these appear at the end of each syllable, as is standard. While we believe there are good linguistic reasons to analyse tones as associated with syllables, we moved all the tones so they immediately followed the final vowel in each syllable. We did this because we thought that locating tones after the syllable-final consonant might give our models a strong cue as to the location of syllable boundaries, and since words often end at syllable boundaries, this would make the word segmentation problem artificially easier. (Our models take a sequence of symbols as input, so the tones must be located somewhere in the sequence. However, the linguistically "correct" solution would probably be to extend the models so they could process input in an auto-segmental format (Goldsmith, 1990) where tones would be on a separate tier and unordered with respect to the segments within a syllable.)

In order to evaluate the importance of tone for our word-segmentation models we also constructed a version of our corpus in which all tones were removed. We present results for all of our models on two versions of the corpus, one that contains tones following the vowels, and another that contains no tones at all. These two corpora constitute the "gold standard" against which our word segmentation models will be evaluated. These corpora contain 50,118 utterances, consisting of 187,533 word tokens.

The training data provided to the word segmentation models is obtained by segmenting the gold data at all possible boundary locations. Consonant clusters, diphthongs and tones (if present) are treated as single units, so the training data appears as follows:

ts ə ^{214} n m ɤ k e i ^{214} t^{h} a ^{55} p e i ^{55} ʂ ɑ ^{51} ŋ l ai ^{35}
t^{h} a ^{55} a ^{55} i ^{35} k e i ^{214} t ɤ t^{h} a ^{55} k e i ^{214} t ɤ
x ə ^{214} n tɕ iɛ ^{214} n t a ^{55} n

The task of a word-segmentation model is to identify which of these possible boundary locations correspond to actual word boundaries. The training corpus without tones contains 531,384 segments, while the training corpus with tones contains 712,318 segments.

3 Adaptor grammars for word segmentation

Adaptor grammars were first introduced by Johnson et al. (2007) as a grammar-based framework for specifying hierarchical non-parametric Bayesian models, and Johnson and Goldwater (2009) describes a number of implementation details that significantly improve performance; the interested reader should consult those papers for a full technical introduction. Johnson (2008b) proposed a number of adaptor grammars for English word segmentation, which we review and minimally modify here so they can perform Chinese word segmentation as well. In section 4 we evaluate these adaptor grammars on the Chinese corpus just described.

The grammars vary along two orthogonal dimensions, which correspond to the kinds of generalisations that the model can learn. The simplest grammar is the unigram adaptor grammar, which generates an utterance as an i.i.d. sequences of words, where each word is a sequence of phonemes. The collocation adaptor grammars capture dependencies above the word level by generating collocations, or groups of words, as memoized units. The syllable adaptor grammars capture dependencies below the word level by generating words as sequences of syllables rather than phonemes.

3.1 Unigram adaptor grammars

In order to motivate adaptor grammars as an extension to Probabilistic Context-Free Grammars (PCFGs), consider an attempt to perform unsupervised word segmentation with a PCFG containing the following rules (ignore the underlining of the Word non-terminal for now).

$$\begin{array}{l}
\text{Words} \rightarrow \text{Words } \underline{\text{Word}} \\
\text{Words} \rightarrow \underline{\text{Word}} \\
\underline{\text{Word}} \rightarrow \text{Phons} \\
\text{Phons} \rightarrow \text{Phon} \\
\text{Phons} \rightarrow \text{Phons Phon} \\
\text{Phons} \rightarrow \text{Phons Tone} \\
\text{Phon} \rightarrow \text{ai} \mid \text{o} \mid \ldots \mid \text{ʂ} \mid \text{tʂ}^{\text{h}} \mid \ldots \\
\text{Tone} \rightarrow 35 \mid 55 \mid 214 \mid \ldots
\end{array} \qquad (1)$$

In this grammar, Phon expands to all the phonemes appearing in the phonemic training data, and Tone expands to all of the tone patterns. (In this and all of the other grammars in this paper, the start symbol is the non-terminal symbol of the first rule in the grammar. This grammar, like all others in this paper, is crafted so that a Word subtree can never begin with a Tone, so the presence of tones does not make the segmentation problem harder).

The trees generated by this grammar are sufficiently expressive to *represent* any possible segmentation of any sequence of phonemes into words (including the true segmentation); a typical segmentation is shown in Figure 1. However, it should also be clear that no matter how we vary the probabilities on the rules of this grammar, *the grammar itself cannot encode the subset of trees that correspond to words of the language*. In order to do this, a model would need to memorise the probabilities of entire Word subtrees, since these are the units that correspond to individual words, but this PCFG simply is not expressive enough to do this.

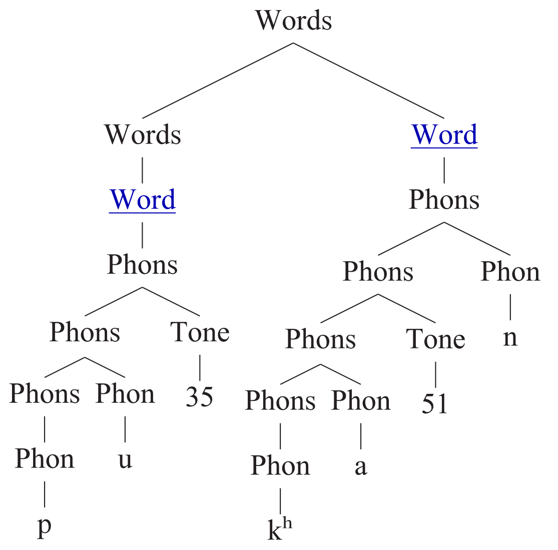

Figure 1: A parse tree generated by the unigram grammar, where adapted and non-adapted non-terminals are shown. It depicts a possible segmentation of p u ^{35} k$^{\text{h}}$ a ^{51} n.

Adaptor grammars learn the probabilities of subtrees in just this way. An adaptor grammar is specified via a set of rules or productions, just like a CFG, and the set of trees that an adaptor grammar generates is exactly the same as the CFG with those rules.

However, an adaptor grammar defines probability distributions over trees in a completely different fashion to a PCFG: for simplicity we focus here on the sampling or predictive distribution, which defines the probability of generating an entire corpus of trees. In a PCFG, the probability of each non-terminal expanding using a given rule is determined by the probability of that rule, and is independent of the expansions of the other non-terminals in the tree. In an adaptor grammar a subset of the non-terminals are designated as *adapted*. We indicate adapted non-terminals by underlining them, so $\underline{\text{Word}}$ is the only adapted non-terminal in (1). Unadapted non-terminals expand just as in a PCFG: a production is chosen according to the production probabilities. An adapted non-terminal can expand in two different ways. With probability proportional to $n(t) - a_A$ an adapted non-terminal A expands to a tree t rooted in A that has been previously generated, while with probability proportional to $m(A)a_A + b_A$ the adapted non-terminal A expands using some grammar rule, just as in a PCFG. Here $n(t)$ is the number of times tree t has been previously generated, $m(A)$ is the number of trees rooted in A that have been previously generated using grammar rules, and $0 \leqslant a_A \leqslant 1$ and $b_A > 0$ are adjustable parameters associated with the adapted non-terminal A.

Technically, this is known as a *Pitman-Yor Process* (PYP) with *concentration parameters* a_A and b_A, where the PCFG rules define the *base distribution* of the process. (The PYP is a generalisation of the Chinese Restaurant Process (CRP); a CRP is a PYP with parameter $a = 0$). Rather than setting the concentration parameters by hand (there are two for each adapted non-terminal in the grammar) we follow Johnson and Goldwater (2009) and put uniform Beta and vague Gamma priors on each of these parameters, and use sampling to explore their posterior values.

Because the probability of selecting a tree t is proportional to $n(t)$, an adaptor grammar is a kind of "rich-get-richer" process that generates power-law distributions. Depending on the values of a_A and b_A, most of the probability mass can wind up concentrated on just a few trees. An adaptor grammar is a kind of "cache" model, in which previously generated subtrees are stored and more likely to be reused in later sentences. That is, while an adapted non-terminal A can expand to any tree rooted in A that can be constructed with the grammar rules, in practice it is increasingly likely to reuse the same trees over and over again. It can be viewed as a kind of tree substitution grammar (Joshi, 2003), but where the tree fragments (as well as their probabilities) are learnt from the data.

The unigram grammar is the simplest of the word segmentation models we investigate in this paper (it is equivalent to the unigram model investigated in Goldwater et al. (2009)). Because the grammars we present below rapidly become long and complicated to read if each grammar rule is explicitly stated, we adopt the following conventions. We use regular expressions to abbreviate our grammars, with the understanding that the regular expressions are always expanded produce a left-recursive structure. For example, the unigram grammar in (1) is abbreviated as:

$$\begin{aligned}&\text{Words} \rightarrow \underline{\text{Word}}^{+}\\ &\underline{\text{Word}} \rightarrow \text{Phon (Phon} \mid \text{Tone})^{\star}\\ &\text{Phon} \rightarrow \text{ai} \mid \text{o} \mid \ldots \mid \text{ʂ} \mid \text{tʂ}^{\text{h}} \mid \ldots\\ &\text{Tone} \rightarrow 35 \mid 55 \mid 214 \mid \ldots\end{aligned} \tag{2}$$

3.2 Collocation adaptor grammars

Goldwater et al. (2006) and Goldwater et al. (2009) demonstrated the importance of contextual dependencies for word segmentation, and proposed a bigram model in order to capture some of these. It turns out that while the bigram model cannot be expressed as an adaptor grammar, a *collocation model*, which captures similar kinds of contextual dependencies, can be expressed as an adaptor grammar (Johnson et al., 2007). In a collocation grammar there are two different adapted non-terminals; $\underline{\text{Word}}$ and $\underline{\text{Colloc}}$; $\underline{\text{Word}}$ expands exactly as in the unigram grammar (2), so it is not repeated here.

$$\begin{aligned}&\text{Collocs} \rightarrow \underline{\text{Colloc}}^{+}\\ &\underline{\text{Colloc}} \rightarrow \text{Words}\\ &\text{Words} \rightarrow \underline{\text{Word}}^{+}\end{aligned} \tag{3}$$

A collocation adaptor grammar caches both words and collocations (which are sequences of words). An utterance is generated by generating one or more collocations. The PYP associated with collocations either regenerates a previously generated collocation or else generates a "fresh" collocation by generating a sequence of words according to the PYP model explained above.

The idea of aggregating words into collocations can be reapplied at a more abstract level by aggregating collocations into "super-collocations", which are sequences of collocations. This involves adding the following additional rules to the grammar in (3):

$$\begin{array}{l} \text{Colloc2s} \rightarrow \underline{\text{Colloc2}}^{+} \\ \underline{\text{Colloc2}} \rightarrow \text{Collocs}^{+} \end{array} \tag{4}$$

There are three PYPs in a grammar with 2 levels of collocations, arranged in a strict Bayesian hierarchy. It should be clear that this process can be repeated indefinitely; we investigate grammars with up to three levels of collocations below. (It should be possible to use Bayesian techniques to learn the appropriate number of levels in the hierarchy, but we leave this for future work).

3.3 Syllable structure adaptor grammars

Brent and Cartwright (1996) and others emphasise the role that syllable-structure and other phonotactic constraints might play in word segmentation. Johnson (2008b) pointed out that adaptor grammars can learn at least some of these kinds of generalisations. It's not unreasonable to assume that language learners can learn to group phonemes into syllables, and that they can exploit this syllabic structure to perform word segmentation. The syllable-structure grammars we describe below assume that word boundaries are always aligned with syllable boundaries; this is not universally true, but it is reliable enough to dramatically improve unsupervised word segmentation in English.

There is considerable cross-linguistic variation in the syllable-structure and phonotactic constraints operative in the languages of the world, so we'd like to avoid "building in" language-specific constraints into our model. We therefore make the relatively conservative assumption that the child can distinguish vowels from consonants, and that the child knows that syllables consist of Onsets, Nuclei and Codas, that Onsets and Codas consist of arbitrary sequences of consonants while Nuclei are arbitrary sequences of vowels and tones, and that Onsets and Codas are optional. Notice that syllable structure in both English and Chinese is considerably more constrained than this; we use this simple model here because it has proved successful for English word segmentation.

The syllable-structure adaptor grammars replace the rules expanding $\underline{\text{Word}}$s with the following rules:

$$\begin{array}{l} \underline{\text{Word}} \rightarrow \text{Syll} \\ \underline{\text{Word}} \rightarrow \text{Syll Syll} \\ \underline{\text{Word}} \rightarrow \text{Syll Syll Syll} \\ \underline{\text{Word}} \rightarrow \text{Syll Syll Syll Syll} \\ \text{Syll} \rightarrow (\underline{\text{Onset}})^{?}\ \underline{\text{Rhy}} \\ \underline{\text{Onset}} \rightarrow \text{C}^{+} \\ \underline{\text{Rhy}} \rightarrow \underline{\text{Nucleus}}\ (\underline{\text{Coda}})^{?} \\ \underline{\text{Nucleus}} \rightarrow \text{V}\ (\text{V} \mid \text{Tone})^{\star} \\ \underline{\text{Coda}} \rightarrow \text{C}^{+} \\ \text{C} \rightarrow \text{ʂ} \mid \text{tʂ}^{\text{h}} \mid \ldots \\ \text{V} \rightarrow \text{ai} \mid \text{o} \mid \ldots \end{array} \tag{5}$$

In these rules the superscript "?" indicates optionality. We used the relatively cumbersome mechanism of enumerating each possible number of syllables per word (we permit words to consist of from 1 to 4 syllables, although ideally this number would not be hard-wired into the grammar) because a relatively trivial modification of this grammar can distinguish word-initial and word-final consonant clusters from word-internal clusters. Johnson (2008b) demonstrated that this significantly improves English word segmentation accuracy. We do not expect this to improve Chinese word segmentation because Chinese clusters do not vary depending on their location within the word, but it will be interesting to see if the additional cluster flexibility that is useful for English segmentation hurts Chinese segmentation.

In this version of the syllable-structure grammar, we replace the $\underline{\text{Word}}$ rules in the syllable adaptor grammar with the following:

$$\begin{array}{l} \underline{\text{Word}} \rightarrow \text{SyllIF} \\ \underline{\text{Word}} \rightarrow \text{SyllI SyllF} \\ \underline{\text{Word}} \rightarrow \text{SyllI Syll SyllF} \\ \underline{\text{Word}} \rightarrow \text{SyllI Syll Syll SyllF} \end{array} \tag{6}$$

and add the following rules expanding the new kinds of syllables to the rules in (5).

$$\begin{array}{l} \text{SyllIF} \rightarrow (\underline{\text{OnsetI}})^{?}\ \underline{\text{RhyF}} \\ \text{SyllI} \rightarrow (\underline{\text{OnsetI}})^{?}\ \underline{\text{Rhy}} \\ \text{SyllF} \rightarrow (\underline{\text{OnsetI}})^{?}\ \underline{\text{RhyF}} \\ \text{Syll} \rightarrow (\underline{\text{Onset}})^{?}\ \underline{\text{Rhy}} \\ \underline{\text{OnsetI}} \rightarrow \text{C}^{+} \\ \underline{\text{RhyF}} \rightarrow \underline{\text{Nucleus}}\ (\underline{\text{CodaF}})^{?} \\ \underline{\text{CodaF}} \rightarrow \text{C}^{+} \end{array} \tag{7}$$

	Syllables		
	None	General	Specialised
Unigram	0.57	0.50	0.50
Colloc	0.69	0.67	0.67
Colloc^2	0.72	0.75	0.75
Colloc^3	*0.64*	**0.77**	**0.77**

Table 1: F-score accuracies of word segmentations produced by the adaptor grammar models on the Chinese corpus *with tones*.

	Syllables		
	None	General	Specialised
Unigram	0.56	0.46	0.46
Colloc	0.70	0.65	0.65
Colloc^2	0.74	0.74	0.73
Colloc^3	0.75	0.76	**0.77**

Table 2: F-score accuracies of word segmentations produced by the adaptor grammar models on the Chinese corpus *without tones*.

These rules distinguish syllable onsets in word-initial position and syllable codas in word-final position; the standard adaptor grammar machinery will then learn distributions over onsets and codas in these positions that possibly differ from those in word-internal positions.

4 Results on Chinese word segmentation

The previous section described two dimensions along which adaptor grammars for word segmentation can independently vary. Above the Word level, there can be from zero to three levels of collocations, yielding four different values for this dimension. Below the Word level, phonemes can either be treated as independent entities, or else they can be grouped into onset, nuclei and coda clusters, and these can vary depending on where they appear within a word. Thus there are three different values for the syllable dimension, so there are twelve different adaptor grammars overall. In addition, we ran all of these grammars on two versions of the corpus, one with tones and one without tones, so we report results for 24 different runs here.

The adaptor grammar inference procedure we used is the one described in Johnson and Goldwater (2009). We ran 1,000 iterations of 8 MCMC chains for each run, and we discarded all but last 200 iterations in order to "burn-in" the sampler. The segmentation we predict is the one that occurs the most frequently in the samples that were not discarded. As is standard, we evaluate the models in terms of token f-score; the results are presented in Tables 1 and 2.

In these tables, "None" indicates that the grammar does not model syllable structure, "General" indicates that the grammar does not distinguish word-peripheral from word-internal clusters, while "Specialised" indicates that it does. "Unigram" indicates that the grammar does not model collocational structure, otherwise the superscript indicates the number of collocational levels that the grammar captures.

Broadly speaking, the results are consistent with the English word segmentation results using adaptor grammars presented by Johnson (2008b). The unigram grammar segmentation accuracy is similar to that obtained for English, but the results for the other models are lower than the results for the corresponding adaptor grammars on English.

We see a general improvement in segmentation accuracy as the number of collocation levels increases, just as for English. However, we do not see any general improvements associated with modelling syllables; indeed, it seems modelling syllables causes accuracy to decrease unless collocational structure is also modelled. This is somewhat surprising, as Chinese has a very regular syllabic structure. It is not surprising that distinguishing word-peripheral and word-medial clusters does not improve segmentation accuracy, as Chinese does not distinguish these kinds of clusters. There is also no sign of the "synergies" when modelling collocations and syllables together that Johnson (2008b) reported.

It is also surprising that tones seem to make little difference to the segmentation accuracy, since they are crucial for disambiguating lexical items. The segmentation accuracy of the models that capture little or no inter-word dependencies (e.g., Unigram, Colloc) improved slightly when the input contains tones, but the best-performing models that capture a more complex set of inter-word dependencies do equally well on the corpus without tones as they do on the corpus with tones. Because these models capture rich inter-word context (they model three levels of collocational structure), it is possible that this context provides sufficient information to segment words even in the absence of tone information, i.e., the tonal information is redundant given the richer inter-word dependencies that these models capture. It is also possible that word segmentation may simply require less information than lexical disambiguation.

One surprising result is the relatively poor performance of the Colloc^3 model without syllables but with tones; we have no explanation for this. However, all 8 of the MCMC chains in this run produced lower f-scores, so it unlikely to be simply a random fluctuation produced by a single outlier.

Note that one should be cautious when comparing the absolute f-scores from these experiments with those of the English study, as the English and Chinese corpora differ in many ways. As Tardif (1993) (the creator of the Chinese corpus) emphasises, this corpus was collected in a much more diverse linguistic environment with child-directed speech from multiple caregivers. The children involved in the Chinese corpus were also older than the children in the English corpus, which may also have affected the nature of the corpus.

5 Conclusion

This paper applied adaptor grammar models of phonemic word segmentation originally developed for English to Chinese data. While the Chinese data was prepared in a very different way to the English data, the adaptor grammars used to perform Chinese word segmentation were very similar to those used for the English word segmentation. They also achieved quite respectable f-score accuracies, which suggests that the same models can do well on both languages.

One puzzling result is that incorporating syllable structure phonotactic constraints, which enhances English word segmentation accuracy considerably, doesn't seem to improve Chinese word segmentation to a similar extent. This may reflect the fact that the word segmentation adaptor grammars were originally designed and tuned for English, and perhaps differently formulated syllable-structure constraints would work well for Chinese. But even if one can "tune" the adaptor grammars to improve performance on Chinese, the challenge is doing this in a way that improves performance on all languages, rather than just one.

Acknowledgments

The authors would like to thank the US NSF for support for this research, which was begun while they were on the faculty at Brown University in the USA. The NSF supported this work through NSF awards 0544127 and 0631667 to Mark Johnson and Katherine Demuth.

The adaptor grammar software is freely available for download from http://web.science.mq.edu.au/˜mjohnson, and the Chinese data was obtained from the Childes archive.

References

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

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

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

Elman, Jeffrey. 1990. Finding structure in time. *Cognitive Science*, 14:197–211.

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

Goldsmith, John A. 1990. *Autosegmental and Metrical Phonology*. Basil Blackwell, Oxford, England.

Goldwater, Sharon, Thomas L. Griffiths, and Mark Johnson. 2006. Contextual dependencies in unsupervised word segmentation. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 673–680, Sydney, Australia. Association for Computational Linguistics.

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

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

Johnson, Mark, Thomas L. Griffiths, and Sharon Goldwater. 2007. Adaptor Grammars: A framework for specifying compositional nonparametric Bayesian models. In Schölkopf, B., J. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems 19*, pages 641–648. MIT Press, Cambridge, MA.

Johnson, Mark. 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.

Johnson, Mark. 2008b. Using adaptor grammars to identifying synergies in the unsupervised acquisition of linguistic structure. In *Proceedings of the 46th Annual Meeting of the Association of Computational Linguistics*, Columbus, Ohio. Association for Computational Linguistics.

Joshi, Aravind. 2003. Tree adjoining grammars. In Mikkov, Ruslan, editor, *The Oxford Handbook of Computational Linguistics*, pages 483–501. Oxford University Press, Oxford, England.

MacWhinney, Brian and Catherine Snow. 1985. The child language data exchange system. *Journal of Child Language*, 12:271–296.

Mochihashi, Daichi, Takeshi Yamada, and Naonori Ueda. 2009. Bayesian unsupervised word segmentation with nested Pitman-Yor language modeling. 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 100–108, Suntec, Singapore, August. Association for Computational Linguistics.

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

Tardif, Twila. 1993. *Adult-to-child speech and language acquisition in Mandarin Chinese*. Ph.D. thesis, Yale University.