Modelling function words improves unsupervised word segmentation

Mark Johnson[1,2], **Anne Christophe**[3,4], **Katherine Demuth**[2,6] and **Emmanuel Dupoux**[3,5]

[1] Department of Computing, Macquarie University, Sydney, Australia
[2] Santa Fe Institute, Santa Fe, New Mexico, USA
[3] Ecole Normale Supérieure, Paris, France
[4] Centre National de la Recherche Scientifique, Paris, France
[5] Ecole des Hautes Etudes en Sciences Sociales, Paris, France
[6] Department of Linguistics, Macquarie University, Sydney, Australia

Abstract

Inspired by experimental psychological findings suggesting that function words play a special role in word learning, we make a simple modification to an Adaptor Grammar based Bayesian word segmentation model to allow it to learn sequences of monosyllabic "function words" at the beginnings and endings of collocations of (possibly multi-syllabic) words. This modification improves unsupervised word segmentation on the standard Bernstein-Ratner (1987) corpus of child-directed English by more than 4% token f-score compared to a model identical except that it does not special-case "function words", setting a new state-of-the-art of 92.4% token f-score. Our function word model assumes that function words appear at the left periphery, and while this is true of languages such as English, it is not true universally. We show that a learner can use Bayesian model selection to determine the location of function words in their language, even though the input to the model only consists of unsegmented sequences of phones. Thus our computational models support the hypothesis that function words play a special role in word learning.

1 Introduction

Over the past two decades psychologists have investigated the role that function words might play in human language acquisition. Their experiments suggest that function words play a special role in the acquisition process: children learn function words before they learn the vast bulk of the associated content words, and they use function words to help identify context words.

The goal of this paper is to determine whether computational models of human language acquisition can provide support for the hypothesis that function words are treated specially in human language acquisition. We do this by comparing two computational models of word segmentation which differ solely in the way that they model function words. Following Elman et al. (1996) and Brent (1999) our word segmentation models identify word boundaries from unsegmented sequences of phonemes corresponding to utterances, effectively performing unsupervised learning of a lexicon. For example, given input consisting of unsegmented utterances such as the following:

j u w a n t t u s i ð ə b ʊ k

a word segmentation model should segment this as *ju want tu si ðə bʊk*, which is the IPA representation of "you want to see the book".

We show that a model equipped with the ability to learn some rudimentary properties of the target language's function words is able to learn the vocabulary of that language more accurately than a model that is identical except that it is incapable of learning these generalisations about function words. This suggests that there are acquisition advantages to treating function words specially that human learners could take advantage of (at least to the extent that they are learning similar generalisations as our models), and thus supports the hypothesis that function words are treated specially in human lexical acquisition. As a reviewer points out, we present no evidence that children use function words in the way that our model does, and we want to emphasise we make no such claim. While absolute accuracy is not directly relevant to the main point of the paper, we note that the models that learn generalisations about function words perform unsupervised word segmentation at 92.5% token f-score on the standard Bernstein-Ratner (1987) corpus, which improves the previous state-of-the-art by more than 4%.

As a reviewer points out, the changes we make to our models to incorporate function words can be viewed as "building in" substantive information about possible human languages. The model that achieves the best token f-score expects function words to appear at the left edge of phrases. While this is true for languages such as English, it is not true universally. By comparing the posterior probability of two models — one in which function words appear at the left edges of phrases, and another in which function words appear at the right edges of phrases — we show that a learner could use Bayesian posterior probabilities to determine that function words appear at the left edges of phrases in English, even though they are not told the locations of word boundaries or which words are function words.

This paper is structured as follows. Section 2 describes the specific word segmentation models studied in this paper, and the way we extended them to capture certain properties of function words. The word segmentation experiments are presented in section 3, and section 4 discusses how a learner could determine whether function words occur on the left-periphery or the right-periphery in the language they are learning. Section 5 concludes and describes possible future work. The rest of this introduction provides background on function words, the Adaptor Grammar models we use to describe lexical acquisition and the Bayesian inference procedures we use to infer these models.

1.1 Psychological evidence for the role of function words in word learning

Traditional descriptive linguistics distinguishes *function words*, such as determiners and prepositions, from *content words*, such as nouns and verbs, corresponding roughly to the distinction between functional categories and lexical categories of modern generative linguistics (Fromkin, 2001).

Function words differ from content words in at least the following ways:

1. there are usually far fewer function word types than content word types in a language
2. function word types typically have much higher token frequency than content word types
3. function words are typically morphologically and phonologically simple (e.g., they are typically monosyllabic)
4. function words typically appear in peripheral positions of phrases (e.g., prepositions typically appear at the beginning of prepositional phrases)
5. each function word class is associated with specific content word classes (e.g., determiners and prepositions are associated with nouns, auxiliary verbs and complementisers are associated with main verbs)
6. semantically, content words denote sets of objects or events, while function words denote more complex relationships over the entities denoted by content words
7. historically, the rate of innovation of function words is much lower than the rate of innovation of content words (i.e., function words are typically "closed class", while content words are "open class")

Properties 1–4 suggest that function words might play a special role in language acquisition because they are especially easy to identify, while property 5 suggests that they might be useful for identifying lexical categories. The models we study here focus on properties 3 and 4, in that they are capable of learning specific sequences of monosyllabic words in peripheral (i.e., initial or final) positions of phrase-like units.

A number of psychological experiments have shown that infants are sensitive to the function words of their language within their first year of life (Shi et al., 2006; Hallé et al., 2008; Shafer et al., 1998), often before they have experienced the "word learning spurt". Crucially for our purpose, infants of this age were shown to exploit frequent function words to segment neighboring content words (Shi and Lepage, 2008; Hallé et al., 2008). In addition, 14 to 18-month-old children were shown to exploit function words to constrain lexical access to known words - for instance, they expect a noun after a determiner (Cauvet et al., 2014; Kedar et al., 2006; Zangl and Fernald, 2007). In addition, it is plausible that function words play a crucial role in children's acquisition of more complex syntactic phenomena (Christophe et al., 2008; Demuth and McCullough, 2009), so it is interesting to investigate the roles they might play in computational models of language acquisition.

1.2 Adaptor grammars

Adaptor grammars are a framework for Bayesian inference of a certain class of hierarchical non-parametric models (Johnson et al., 2007b). They define distributions over the trees specified by a context-free grammar, but unlike probabilistic context-free grammars, they "learn" distributions over the possible subtrees of a user-specified set of "adapted" nonterminals. (Adaptor grammars are non-parametric, i.e., not characterisable by a finite set of parameters, if the set of possible subtrees of the adapted nonterminals is infinite). Adaptor grammars are useful when the goal is to learn a potentially unbounded set of entities that need to satisfy hierarchical constraints. As section 2 explains in more detail, word segmentation is such a case: words are composed of syllables and belong to phrases or collocations, and modelling this structure improves word segmentation accuracy.

Adaptor Grammars are formally defined in Johnson et al. (2007b), which should be consulted for technical details. Adaptor Grammars (AGs) are an extension of Probabilistic Context-Free Grammars (PCFGs), which we describe first. A *Context-Free Grammar* (CFG) $G = (N, W, R, S)$ consists of disjoint finite sets of *nonterminal symbols* N and *terminal symbols* W, a finite set of *rules* R of the form $A \rightarrow \alpha$ where $A \in N$ and $\alpha \in (N \cup W)^\star$, and a *start symbol* $S \in N$. (We assume there are no "ϵ-rules" in R, i.e., we require that $|\alpha| \geq 1$ for each $A \rightarrow \alpha \in R$).

A *Probabilistic Context-Free Grammar* (PCFG) is a quintuple $(N, W, R, S, \boldsymbol{\theta})$ where N, W, R and S are the nonterminals, terminals, rules and start symbol of a CFG respectively, and $\boldsymbol{\theta}$ is a vector of non-negative reals indexed by R that satisfy $\sum_{\alpha \in R_A} \theta_{A \rightarrow \alpha} = 1$ for each $A \in N$, where $R_A = \{A \rightarrow \alpha : A \rightarrow \alpha \in R\}$ is the set of rules expanding A.

Informally, $\theta_{A \rightarrow \alpha}$ is the probability of a node labelled A expanding to a sequence of nodes labelled α, and the probability of a tree is the product of the probabilities of the rules used to construct each non-leaf node in it. More precisely, for each $X \in N \cup W$ a PCFG associates distributions G_X over the set of trees $\mathcal{T}_X$ generated by X as follows:

If $X \in W$ (i.e., if X is a terminal) then G_X is the distribution that puts probability 1 on the single-node tree labelled X.

If $X \in N$ (i.e., if X is a nonterminal) then:

$$G_X = \sum_{X \rightarrow B_1 \ldots B_n \in R_X} \theta_{X \rightarrow B_1 \ldots B_n} \mathrm{TD}_X(G_{B_1}, \ldots, G_{B_n}) \quad (1)$$

where R_X is the subset of rules in R expanding nonterminal $X \in N$, and:

$$\mathrm{TD}_X(G_1, \ldots, G_n) \left(\begin{array}{c} X \\ \overbrace{t_1 \ldots t_n} \end{array} \right) = \prod_{i=1}^{n} G_i(t_i).$$

That is, $\mathrm{TD}_X(G_1, \ldots, G_n)$ is a distribution over the set of trees $\mathcal{T}_X$ generated by nonterminal X, where each subtree t_i is generated independently from G_i. The PCFG generates the distribution G_S over the set of trees $\mathcal{T}_S$ generated by the start symbol S; the distribution over the strings it generates is obtained by marginalising over the trees.

In a Bayesian PCFG one puts Dirichlet priors $\mathrm{Dir}(\boldsymbol{\alpha})$ on the rule probability vector $\boldsymbol{\theta}$, such that there is one Dirichlet parameter $\alpha_{A \rightarrow \alpha}$ for each rule $A \rightarrow \alpha \in R$. There are Markov Chain Monte Carlo (MCMC) and Variational Bayes procedures for estimating the posterior distribution over rule probabilities $\boldsymbol{\theta}$ and parse trees given data consisting of terminal strings alone (Kurihara and Sato, 2006; Johnson et al., 2007a).

PCFGs can be viewed as recursive mixture models over trees. While PCFGs are expressive enough to describe a range of linguistically-interesting phenomena, PCFGs are *parametric models*, which limits their ability to describe phenomena where the set of basic units, as well as their properties, are the target of learning. Lexical acqusition is an example of a phenomenon that is naturally viewed as *non-parametric inference*, where the number of lexical entries (i.e., words) as well as their properties must be learnt from the data.

It turns out there is a straight-forward modification to the PCFG distribution (1) that makes it suitably non-parametric. As Johnson et al. (2007b) explain, by inserting a Dirichlet Process (DP) or Pitman-Yor Process (PYP) into the generative mechanism (1) the model "concentrates" mass on a subset of trees (Teh et al., 2006). Specifically, an Adaptor Grammar identifies a subset $A \subseteq N$ of *adapted nonterminals*. In an Adaptor Grammar the unadapted nonterminals $N \setminus A$ expand via (1), just as in a PCFG, but the distributions of the adapted nonterminals A are "concentrated" by passing them through a DP or PYP:

$$\begin{aligned} H_X &= \sum_{X \rightarrow B_1 \ldots B_n \in R_X} \theta_{X \rightarrow B_1 \ldots B_n} \mathrm{TD}_X(G_{B_1}, \ldots, G_{B_n}) \\ G_X &= \ \mathrm{PYP}(H_X, a_X, b_X) \end{aligned}$$

Here a_X and b_X are parameters of the PYP associated with the adapted nonterminal X. As Goldwater et al. (2011) explain, such Pitman-Yor Processes naturally generate power-law distributed data.

Informally, Adaptor Grammars can be viewed as caching entire subtrees of the adapted nonterminals. Roughly speaking, the probability of generating a particular subtree of an adapted nonterminal is proportional to the number of times that subtree has been generated before. This "rich get richer" behaviour causes the distribution of subtrees to follow a power-law (the power is specified by the a_X parameter of the PYP). The PCFG rules expanding an adapted nonterminal X define the "base distribution" of the associated DP or PYP, and the a_X and b_X parameters determine how much mass is reserved for "new" trees.

There are several different procedures for inferring the parse trees and the rule probabilities given a corpus of strings: Johnson et al. (2007b) describe a MCMC sampler and Cohen et al. (2010) describe a Variational Bayes procedure. We use the MCMC procedure here since this has been successfully applied to word segmentation problems in previous work (Johnson, 2008).

2 Word segmentation with Adaptor Grammars

Perhaps the simplest word segmentation model is the *unigram model*, where utterances are modeled as sequences of words, and where each word is a sequence of segments (Brent, 1999; Goldwater et al., 2009). A unigram model can be expressed as an Adaptor Grammar with one adapted nonterminal $\underline{\mathsf{Word}}$ (we indicate adapted nonterminals by underlining them in grammars here; regular expressions are expanded into right-branching productions).

$$\mathsf{Sentence} \rightarrow \mathsf{Word}^{+} \tag{2}$$

$$\underline{\mathsf{Word}} \rightarrow \mathsf{Phone}^{+} \tag{3}$$

The first rule (2) says that a sentence consists of one or more Words, while the second rule (3) states that a Word consists of a sequence of one or more Phones; we assume that there are rules expanding Phone into all possible phones. Because Word is an adapted nonterminal, the adaptor grammar memoises Word subtrees, which corresponds to learning the phone sequences for the words of the language.

The more sophisticated Adaptor Grammars discussed below can be understood as specialising either the first or the second of the rules in (2–3). The next two subsections review the Adaptor Grammar word segmentation models presented in Johnson (2008) and Johnson and Goldwater (2009): section 2.1 reviews how phonotactic syllable-structure constraints can be expressed with Adaptor Grammars, while section 2.2 reviews how phrase-like units called "collocations" capture inter-word dependencies. Section 2.3 presents the major novel contribution of this paper by explaining how we modify these adaptor grammars to capture some of the special properties of function words.

2.1 Syllable structure and phonotactics

The rule (3) models words as sequences of independently generated phones: this is what Goldwater et al. (2009) called the "monkey model" of word generation (it instantiates the metaphor that word types are generated by a monkey randomly banging on the keys of a typewriter). However, the words of a language are typically composed of one or more syllables, and explicitly modelling the internal structure of words typically improves word segmentation considerably.

Johnson (2008) suggested replacing (3) with the following model of word structure:

$$\underline{\mathsf{Word}} \rightarrow \mathsf{Syllable}^{1:4} \tag{4}$$

$$\mathsf{Syllable} \rightarrow (\mathsf{Onset})\ \mathsf{Rhyme} \tag{5}$$

$$\underline{\mathsf{Onset}} \rightarrow \mathsf{Consonant}^{+} \tag{6}$$

$$\mathsf{Rhyme} \rightarrow \mathsf{Nucleus}\ (\mathsf{Coda}) \tag{7}$$

$$\underline{\mathsf{Nucleus}} \rightarrow \mathsf{Vowel}^{+} \tag{8}$$

$$\underline{\mathsf{Coda}} \rightarrow \mathsf{Consonant}^{+} \tag{9}$$

Here and below superscripts indicate iteration (e.g., a Word consists of 1 to 4 Syllables), while an Onset consists of an unbounded number of Consonants), while parentheses indicate optionality (e.g., a Rhyme consists of an obligatory Nucleus followed by an optional Coda). We assume that there are rules expanding Consonant and Vowel to the set of all consonants and vowels respectively (this amounts to assuming that the learner can distinguish consonants from vowels). Because Onset, Nucleus and Coda are adapted, this model learns the possible syllable onsets, nucleii and coda of the language, even though neither syllable structure nor word boundaries are explicitly indicated in the input to the model.

The model just described assumes that word-internal syllables have the same structure as word-peripheral syllables, but in languages such as English word-peripheral onsets and codas can be more complex than the corresponding word-internal onsets and codas. For example, the word "string" begins with the onset cluster *str*, which is relatively rare word-internally. Johnson (2008) showed that word segmentation accuracy improves if the model can learn different consonant sequences for word-inital onsets and word-final codas. It is easy to express this as an Adaptor

Grammar: (4) is replaced with (10–11) and (12–17) are added to the grammar.

$$\underline{\mathsf{Word}} \rightarrow \mathsf{SyllableIF} \tag{10}$$

$$\underline{\mathsf{Word}} \rightarrow \mathsf{SyllableI}\ \mathsf{Syllable}^{0:2}\ \mathsf{SyllableF} \tag{11}$$

$$\mathsf{SyllableIF} \rightarrow (\mathsf{OnsetI})\ \mathsf{RhymeF} \tag{12}$$

$$\mathsf{SyllableI} \rightarrow (\mathsf{OnsetI})\ \mathsf{Rhyme} \tag{13}$$

$$\mathsf{SyllableF} \rightarrow (\mathsf{Onset})\ \mathsf{RhymeF} \tag{14}$$

$$\underline{\mathsf{OnsetI}} \rightarrow \mathsf{Consonant}^{+} \tag{15}$$

$$\mathsf{RhymeF} \rightarrow \mathsf{Nucleus}\ (\mathsf{CodaF}) \tag{16}$$

$$\underline{\mathsf{CodaF}} \rightarrow \mathsf{Consonant}^{+} \tag{17}$$

In this grammar the suffix "I" indicates a word-initial element, and "F" indicates a word-final element. Note that the model simply has the ability to learn that different clusters can occur word-peripherally and word-internally; it is not given any information about the relative complexity of these clusters.

2.2 Collocation models of inter-word dependencies

Goldwater et al. (2009) point out the detrimental effect that inter-word dependencies can have on word segmentation models that assume that the words of an utterance are independently generated. Informally, a model that generates words independently is likely to incorrectly segment multi-word expressions such as "the doggie" as single words because the model has no way to capture word-to-word dependencies, e.g., that "doggie" is typically preceded by "the". Goldwater et al show that word segmentation accuracy improves when the model is extended to capture bigram dependencies.

Adaptor grammar models cannot express bigram dependencies, but they can capture similiar inter-word dependencies using phrase-like units that Johnson (2008) calls collocations. Johnson and Goldwater (2009) showed that word segmentation accuracy improves further if the model learns a nested hierarchy of collocations. This can be achieved by replacing (2) with (18–21).

$$\mathsf{Sentence} \rightarrow \mathsf{Colloc3}^{+} \tag{18}$$

$$\underline{\mathsf{Colloc3}} \rightarrow \mathsf{Colloc2}^{+} \tag{19}$$

$$\underline{\mathsf{Colloc2}} \rightarrow \mathsf{Colloc1}^{+} \tag{20}$$

$$\underline{\mathsf{Colloc1}} \rightarrow \mathsf{Word}^{+} \tag{21}$$

Informally, Colloc1, Colloc2 and Colloc3 define a nested hierarchy of phrase-like units. While not designed to correspond to syntactic phrases, by examining the sample parses induced by the Adaptor Grammar we noticed that the collocations often correspond to noun phrases, prepositional phrases or verb phrases. This motivates the extension to the Adaptor Grammar discussed below.

2.3 Incorporating "function words" into collocation models

The starting point and baseline for our extension is the adaptor grammar with syllable structure phonotactic constraints and three levels of collocational structure (5-21), as prior work has found that this yields the highest word segmentation token f-score (Johnson and Goldwater, 2009).

Our extension assumes that the Colloc1 − Colloc3 constituents are in fact phrase-like, so we extend the rules (19–21) to permit an optional sequence of monosyllabic words at the left edge of each of these constituents. Our model thus captures two of the properties of function words discussed in section 1.1: they are monosyllabic (and thus phonologically simple), and they appear on the periphery of phrases. (We put "function words" in scare quotes below because our model only approximately captures the linguistic properties of function words).

Specifically, we replace rules (19–21) with the following sequence of rules:

$$\underline{\mathsf{Colloc3}} \rightarrow (\mathsf{FuncWords3})\ \mathsf{Colloc2}^{+} \tag{22}$$

$$\underline{\mathsf{Colloc2}} \rightarrow (\mathsf{FuncWords2})\ \mathsf{Colloc1}^{+} \tag{23}$$

$$\underline{\mathsf{Colloc1}} \rightarrow (\mathsf{FuncWords1})\ \mathsf{Word}^{+} \tag{24}$$

$$\underline{\mathsf{FuncWords3}} \rightarrow \mathsf{FuncWord3}^{+} \tag{25}$$

$$\underline{\mathsf{FuncWord3}} \rightarrow \mathsf{SyllableIF} \tag{26}$$

$$\underline{\mathsf{FuncWords2}} \rightarrow \mathsf{FuncWord2}^{+} \tag{27}$$

$$\underline{\mathsf{FuncWord2}} \rightarrow \mathsf{SyllableIF} \tag{28}$$

$$\underline{\mathsf{FuncWords1}} \rightarrow \mathsf{FuncWord1}^{+} \tag{29}$$

$$\underline{\mathsf{FuncWord1}} \rightarrow \mathsf{SyllableIF} \tag{30}$$

This model memoises (i.e., learns) both the individual "function words" and the sequences of "function words" that modify the Colloc1 − Colloc3 constituents. Note also that "function words" expand directly to SyllableIF, which in turn expands to a monosyllable with a word-initial onset and word-final coda. This means that "function words" are memoised independently of the "content words" that Word expands to; i.e., the model learns distinct "function word" and "content word" vocabularies. Figure 1 depicts a sample parse generated by this grammar.

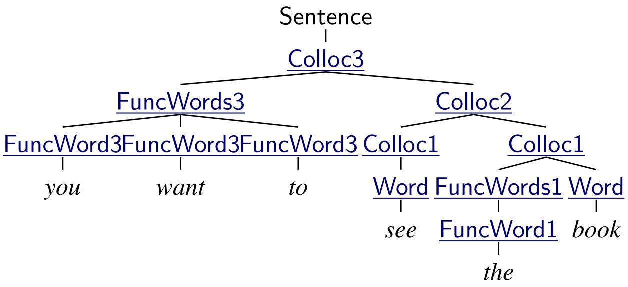

Figure 1: A sample parse generated by the "function word" Adaptor Grammar with rules (10–18) and (22–30). To simplify the parse we only show the root node and the adapted nonterminals, and replace word-internal structure by the word's orthographic form.

This grammar builds in the fact that function words appear on the left periphery of phrases. This is true of languages such as English, but is not true cross-linguistically. For comparison purposes we also include results for a mirror-image model that permits "function words" on the right periphery, a model which permits "function words" on both the left and right periphery (achieved by changing rules 22–24), as well as a model that analyses all words as monosyllabic.

Section 4 explains how a learner could use Bayesian model selection to determine that function words appear on the left periphery in English by comparing the posterior probability of the data under our "function word" Adaptor Grammar to that obtained using a grammar which is identical except that rules (22–24) are replaced with the mirror-image rules in which "function words" are attached to the right periphery.

3 Word segmentation results

This section presents results of running our Adaptor Grammar models on subsets of the Bernstein-Ratner (1987) corpus of child-directed English. We use the Adaptor Grammar software available from http://web.science.mq.edu.au/˜mjohnson/ with the same settings as described in Johnson and Goldwater (2009), i.e., we perform Bayesian inference with "vague" priors for all hyperparameters (so there are no adjustable parameters in our models), and perform 8 different MCMC runs of each condition with table-label resampling for 2,000 sweeps of the training data. At every 10th sweep of the last 1,000 sweeps we use the model to segment the entire corpus (even if it is only trained on a subset of it), so we collect 800 sample segmentations of each utterance. The most frequent segmentation in these 800 sample segmentations is the one we score in the evaluations below.

Model	**Token f-score**	**Boundary precision**	**Boundary recall**
Baseline	0.872	0.918	0.956
+ left FWs	**0.924**	0.935	**0.990**
+ left + right FWs	0.912	**0.957**	0.953

Table 1: Mean token f-scores and boundary precision and recall results averaged over 8 trials, each consisting of 8 MCMC runs of models trained and tested on the full Bernstein-Ratner (1987) corpus (the standard deviations of all values are less than 0.006; Wilcox sign tests show the means of all token f-scores differ $p < 2\text{e-}4$).

3.1 Word segmentation with "function word" models

Here we evaluate the word segmentations found by the "function word" Adaptor Grammar model described in section 2.3 and compare it to the baseline grammar with collocations and phonotactics from Johnson and Goldwater (2009). Figure 2 presents the standard token and lexicon (i.e., type) f-score evaluations for word segmentations proposed by these models (Brent, 1999), and Table 1 summarises the token and lexicon f-scores for the major models discussed in this paper. It is interesting to note that adding "function words" improves token f-score by more than 4%, corresponding to a 40% reduction in overall error rate.

When the training data is very small the Monosyllabic grammar produces the highest accuracy results, presumably because a large proportion of the words in child-directed speech are monosyllabic. However, at around 25 sentences the more complex models that are capable of finding multisyllabic words start to become more accurate.

It's interesting that after about 1,000 sentences the model that allows "function words" only on the right periphery is considerably less accurate than the baseline model. Presumably this is because it tends to misanalyse multi-syllabic words on the right periphery as sequences of monosyllabic words.

The model that allows "function words" only on the left periphery is more accurate than the model that allows them on both the left and right periphery when the input data ranges from about 100 to about 1,000 sentences, but when the training data

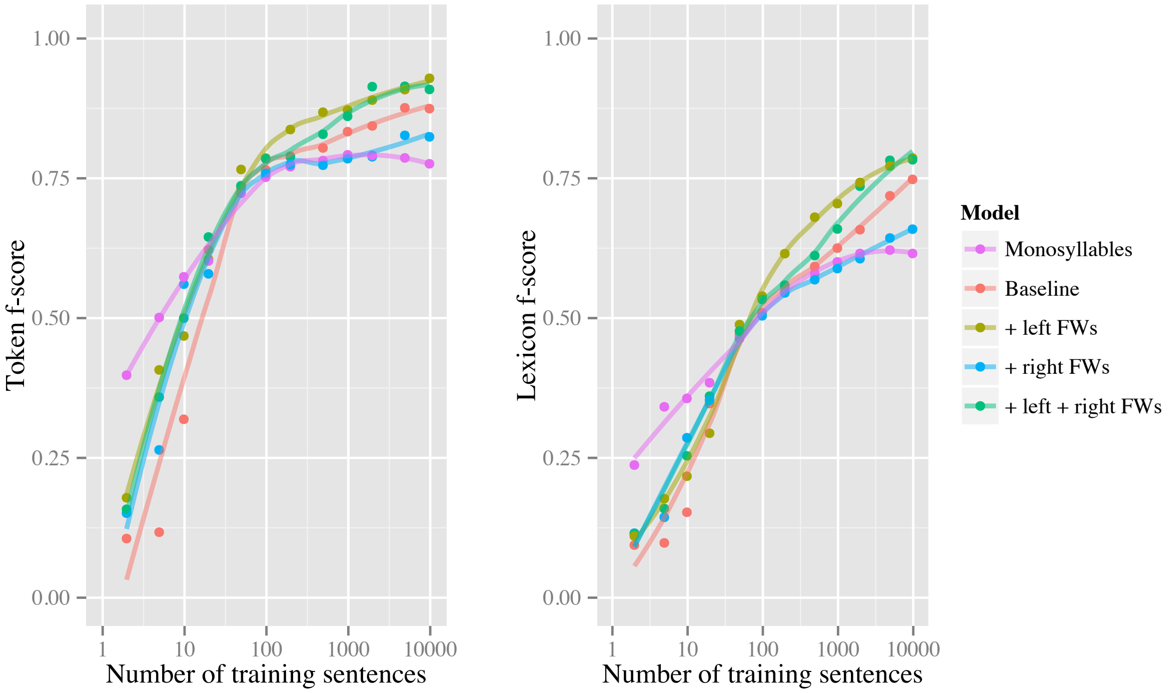

Figure 2: Token and lexicon (i.e., type) f-score on the Bernstein-Ratner (1987) corpus as a function of training data size for the baseline model, the model where "function words" can appear on the left periphery, a model where "function words" can appear on the right periphery, and a model where "function words" can appear on both the left and the right periphery. For comparison purposes we also include results for a model that assumes that all words are monosyllabic.

is larger than about 1,000 sentences both models are equally accurate.

3.2 Content and function words found by "function word" model

As noted earlier, the "function word" model generates function words via adapted nonterminals other than the Word category. In order to better understand just how the model works, we give the 5 most frequent words in each word category found during 8 MCMC runs of the left-peripheral "function word" grammar above:

Word : *book, doggy, house, want, I*
FuncWord1 : *a, the, your, little*[1], *in*
FuncWord2 : *to, in, you, what, put*
FuncWord3 : *you, a, what, no, can*

Interestingly, these categories seem fairly reasonable. The Word category includes open-class nouns and verbs, the FuncWord1 category includes noun modifiers such as determiners, while the FuncWord2 and FuncWord3 categories include prepositions, pronouns and auxiliary verbs.

Thus, the present model, initially aimed at segmenting words from continuous speech, shows three interesting characteristics that are also exhibited by human infants: it distinguishes between function words and content words (Shi and Werker, 2001), it allows learners to acquire at least some of the function words of their language (e.g. (Shi et al., 2006)); and furthermore, it may also allow them to start grouping together function words according to their category (Cauvet et al., 2014; Shi and Melançon, 2010).

4 Are "function words" on the left or right periphery?

We have shown that a model that expects function words on the left periphery performs more accurate word segmentation on English, where function words do indeed typically occur on the left periphery, leaving open the question: how could a learner determine whether function words generally appear on the left or the right periphery of phrases in the language they are learning? This question is important because knowing the side where function words preferentially occur is re-

[1] The phone 'l' is generated by both Consonant and Vowel, so "little" can be (incorrectly) analysed as one syllable.

lated to the question of the direction of syntactic headedness in the language, and an accurate method for identifying the location of function words might be useful for initialising a syntactic learner. Experimental evidence suggests that infants as young as 8 months of age already expect function words on the correct side for their language — left-periphery for Italian infants and right-periphery for Japanese infants (Gervain et al., 2008) — so it is interesting to see whether purely distributional learners such as the ones studied here can identify the correct location of function words in phrases.

We experimented with a variety of approaches that use a single adaptor grammar inference process, but none of these were successful. For example, we hoped that given an Adaptor Grammar that permits "function words" on both the left and right periphery, the inference procedure would decide that the right-periphery rules simply are not used in a language like English. Unfortunately we did not find this in our experiments; the right-periphery rules were used almost as often as the left-periphery rules (recall that a large fraction of the words in English child-directed speech are monosyllabic).

In this section, we show that learners could use Bayesian model selection to determine that function words appear on the left periphery in English by comparing the marginal probability of the data for the left-periphery and the right-periphery models.

Instead, we used Bayesian model selection techniques to determine whether left-peripheral or a right-peripheral model better fits the unsegmented utterances that constitute the training data.[2] While Bayesian model selection is in principle straight-forward, it turns out to require the ratio of two integrals (for the "evidence" or marginal likelihood) that are often intractable to compute.

Specifically, given a training corpus D of unsegmented sentences and model families G_1 and G_2 (here the "function word" adaptor grammars with left-peripheral and right-peripheral attachment respectively), the Bayes factor K is the ratio of the marginal likelihoods of the data:

$$K = \frac{\mathrm{P}(D \mid G_1)}{\mathrm{P}(D \mid G_2)}$$

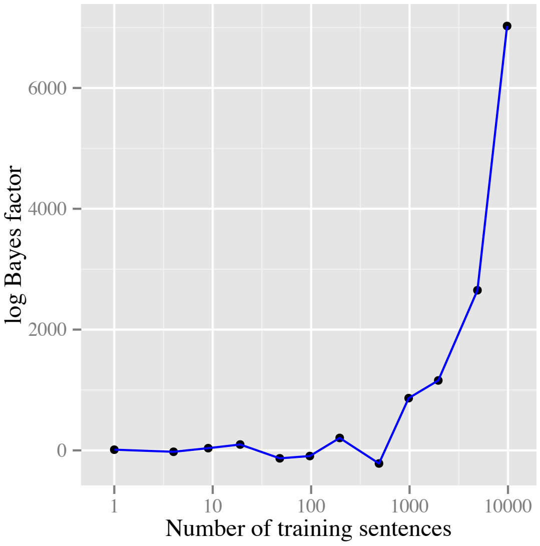

Figure 3: Bayes factor in favour of left-peripheral "function word" attachment as a function of the number of sentences in the training corpus, calculated using the Harmonic Mean estimator (see warning in text).

where the marginal likelihood or "evidence" for a model G is obtained by integrating over all of the hidden or latent structure and parameters $\boldsymbol{\theta}$:

$$\mathrm{P}(D \mid G) = \int_{\Delta} \mathrm{P}(D, \boldsymbol{\theta} \mid G)\, d\boldsymbol{\theta} \quad (31)$$

Here the variable $\boldsymbol{\theta}$ ranges over the space Δ of all possible parses for the utterances in D and all possible configurations of the Pitman-Yor processes and their parameters that constitute the "state" of the Adaptor Grammar G. While the probability of any specific Adaptor Grammar configuration $\boldsymbol{\theta}$ is not too hard to calculate (the MCMC sampler for Adaptor Grammars can print this after each sweep through D), the integral in (31) is in general intractable.

Textbooks such as Murphy (2012) describe a number of methods for calculating $\mathrm{P}(D \mid G)$, but most of them assume that the parameter space Δ is continuous and so cannot be directly applied here. The Harmonic Mean estimator (32) for (31), which we used here, is a popular estimator for (31) because it only requires the ability to calculate $\mathrm{P}(D, \boldsymbol{\theta} \mid G)$ for samples from $\mathrm{P}(\boldsymbol{\theta} \mid D, G)$:

$$\mathrm{P}(D \mid G) \approx \left(\frac{1}{n} \sum_{i=1}^{n} \frac{1}{\mathrm{P}(D, \boldsymbol{\theta}_i \mid G)} \right)^{-1}$$

where $\boldsymbol{\theta}_i, \ldots, \boldsymbol{\theta}_n$ are n samples from $\mathrm{P}(\boldsymbol{\theta} \mid$

[2] Note that neither the left-peripheral nor the right-peripheral model is correct: even strongly left-headed languages like English typically contain a few right-headed constructions. For example, "ago" is arguably the head of the phrase "ten years ago".

D, G), which can be generated by the MCMC procedure.

Figure 3 depicts how the Bayes factor in favour of left-peripheral attachment of "function words" varies as a function of the number of utterances in the training data D (calculated from the last 1000 sweeps of 8 MCMC runs of the corresponding adaptor grammars). As that figure shows, once the training data contains more than about 1,000 sentences the evidence for the left-peripheral grammar becomes very strong. On the full training data the estimated log Bayes factor is over 6,000, which would constitute overwhelming evidence in favour of left-peripheral attachment.

Unfortunately, as Murphy and others warn, the Harmonic Mean estimator is extremely unstable (Radford Neal calls it "the worst MCMC method ever" in his blog), so we think it is important to confirm these results using a more stable estimator. However, given the magnitude of the differences and the fact that the two models being compared are of similar complexity, we believe that these results suggest that Bayesian model selection can be used to determine properties of the language being learned.

5 Conclusions and future work

This paper showed that the word segmentation accuracy of a state-of-the-art Adaptor Grammar model is significantly improved by extending it so that it explicitly models some properties of function words. We also showed how Bayesian model selection can be used to identify that function words appear on the left periphery of phrases in English, even though the input to the model only consists of an unsegmented sequence of phones.

Of course this work only scratches the surface in terms of investigating the role of function words in language acquisition. It would clearly be very interesting to examine the performance of these models on other corpora of child-directed English, as well as on corpora of child-directed speech in other languages. Our evaluation focused on word-segmentation, but we could also evaluate the effect that modelling "function words" has on other aspects of the model, such as its ability to learn syllable structure.

The models of "function words" we investigated here only capture two of the 7 linguistic properties of function words identified in section 1 (i.e., that function words tend to be monosyllabic, and that they tend to appear phrase-peripherally), so it would be interesting to develop and explore models that capture other linguistic properties of function words. For example, following the suggestion by Hochmann et al. (2010) that human learners use frequency cues to identify function words, it might be interesting to develop computational models that do the same thing. In an Adaptor Grammar the frequency distribution of function words might be modelled by specifying the prior for the Pitman-Yor Process parameters associated with the function words' adapted nonterminals so that it prefers to generate a small number of high-frequency items.

It should also be possible to develop models which capture the fact that function words tend not to be topic-specific. Johnson et al. (2010) and Johnson et al. (2012) show how Adaptor Grammars can model the association between words and non-linguistic "topics"; perhaps these models could be extended to capture some of the semantic properties of function words.

It would also be interesting to further explore the extent to which Bayesian model selection is a useful approach to linguistic "parameter setting". In order to do this it is imperative to develop better methods than the problematic "Harmonic Mean" estimator used here for calculating the evidence (i.e., the marginal probability of the data) that can handle the combination of discrete and continuous hidden structure that occur in computational linguistic models.

As well as substantially improving the accuracy of unsupervised word segmentation, this work is interesting because it suggests a connection between unsupervised word segmentation and the induction of syntactic structure. It is reasonable to expect that hierarchical non-parametric Bayesian models such as Adaptor Grammars may be useful tools for exploring such a connection.

Acknowledgments

This work was supported in part by the Australian Research Council's Discovery Projects funding scheme (project numbers DP110102506 and DP110102593), the European Research Council (ERC-2011-AdG-295810 BOOTPHON), the Agence Nationale pour la Recherche (ANR-10-LABX-0087 IEC, and ANR-10-IDEX-0001-02 PSL*), and the Mairie de Paris, Ecole des Hautes Etudes en Sciences Sociales, the Ecole Normale Supérieure, and the Fondation Pierre Gilles de Gennes.

References

N. Bernstein-Ratner. 1987. The phonology of parent-child speech. In K. Nelson and A. van Kleeck, editors, *Children's Language*, volume 6, pages 159–174. Erlbaum, Hillsdale, NJ.

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

Elodie Cauvet, Rita Limissuri, Severine Millotte, Katrin Skoruppa, Dominique Cabrol, and Anne Christophe. 2014. Function words constrain on-line recognition of verbs and nouns in French 18-month-olds. *Language Learning and Development*, pages 1–18.

Anne Christophe, Séverine Millotte, Savita Bernal, and Jeffrey Lidz. 2008. Bootstrapping lexical and syntactic acquisition. *Language and Speech*, 51(1-2):61–75.

Shay B. Cohen, David M. Blei, and Noah A. Smith. 2010. Variational inference for adaptor grammars. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 564–572, Los Angeles, California, June. Association for Computational Linguistics.

Katherine Demuth and Elizabeth McCullough. 2009. The prosodic (re-)organization of childrens early English articles. *Journal of Child Language*, 36(1):173–200.

J. Elman, E. Bates, M. H. Johnson, A. Karmiloff-Smith, D. Parisi, and K. Plunkett. 1996. *Rethinking Innateness: A Connectionist Perspective on Development*. MIT Press/Bradford Books, Cambridge, MA.

Victoria Fromkin, editor. 2001. *Linguistics: An Introduction to Linguistic Theory*. Blackwell, Oxford, UK.

Judit Gervain, Marina Nespor, Reiko Mazuka, Ryota Horie, and Jacques Mehler. 2008. Bootstrapping word order in prelexical infants: A japaneseitalian cross-linguistic study. *Cognitive Psychology*, 57(1):56 – 74.

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

Sharon Goldwater, Thomas L. 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.

Pierre A. Hallé, Catherine Durand, and Bénédicte de Boysson-Bardies. 2008. Do 11-month-old French infants process articles? *Language and Speech*, 51(1-2):23–44.

Jean-Rémy Hochmann, Ansgar D. Endress, and Jacques Mehler. 2010. Word frequency as a cue for identifying function words in infancy. *Cognition*, 115(3):444 – 457.

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.

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

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, Katherine Demuth, and Michael Frank. 2012. Exploiting social information in grounded language learning via grammatical reduction. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, pages 883–891, Jeju Island, Korea, July. Association for Computational Linguistics.

Mark Johnson. 2008. 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.

Yarden Kedar, Marianella Casasola, and Barbara Lust. 2006. Getting there faster: 18- and 24-month-old infants' use of function words to determine reference. *Child Development*, 77(2):325–338.

Kenichi Kurihara and Taisuke Sato. 2006. Variational Bayesian grammar induction for natural language. In Yasubumi Sakakibara, Satoshi Kobayashi, Kengo Sato, Tetsuro Nishino, and Etsuji Tomita, editors, *Grammatical Inference: Algorithms and Applications*, pages 84–96. Springer.

Kevin P Murphy. 2012. *Machine learning: a probabilistic perspective*. The MIT Press.

Valerie L Shafer, David W Shucard, Janet L Shucard, and LouAnn Gerken. 1998. An electrophysiological study of infants' sensitivity to the sound patterns of English speech. *Journal of Speech, Language and Hearing Research*, 41(4):874.

Rushen Shi and Mélanie Lepage. 2008. The effect of functional morphemes on word segmentation in preverbal infants. *Developmental Science*, 11(3):407–413.

Rushen Shi and Andréane Melançon. 2010. Syntactic categorization in French-learning infants. *Infancy*, 15(517–533).

Rushen Shi and Janet Werker. 2001. Six-months old infants' preference for lexical words. *Psychological Science*, 12:71–76.

Rushen Shi, Anne Cutler, Janet Werker, and Marisa Cruickshank. 2006. Frequency and form as determinants of functor sensitivity in English-acquiring infants. *The Journal of the Acoustical Society of America*, 119(6):EL61–EL67.

Yee Whye Teh, Michael Jordan, Matthew Beal, and David Blei. 2006. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101:1566–1581.

Renate Zangl and Anne Fernald. 2007. Increasing flexibility in children's online processing of grammatical and nonce determiners in fluent speech. *Language Learning and Development*, 3(3):199–231.