Studying the effect of input size for Bayesian Word Segmentation on the Providence Corpus

Benjamin Börschinger[1,3] *Katherine Demuth*[2] *Mark Johnson*[1]
(1) Department of Computer Science, Macquarie University
(2) Department of Linguistics, Macquarie University
(3) Department of Computational Linguistics, Heidelberg University
benjamin.borschinger@mq.edu.au, katherine.demuth@mq.edu.au,
mark.johnson@mq.edu.au

Abstract

Studies of computational models of language acquisition depend to a large part on the input available for experiments. In this paper, we study the effect that input size has on the performance of word segmentation models embodying different kinds of linguistic assumptions. Because currently available corpora for word segmentation are not suited for addressing this question, we perform our study on a novel corpus based on the Providence Corpus (Demuth et al., 2006). We find that input size can have dramatic effects on segmentation performance and that, somewhat surprisingly, models performing well on smaller amounts of data can show a marked decrease in performance when exposed to larger amounts of data. We also present the data-set on which we perform our experiments comprising longitudinal data for six children. This corpus makes it possible to ask more specific questions about computational models of word segmentation, in particular about intra-language variability and about how the performance of different models can change over time.[1]

KEYWORDS: word segmentation, language acquisition, resources, Bayesian modelling, unsupervised learning.

[1] The corpus and the code to run the experiments is available at `http://web.science.mq.edu.au/~bborschi/`.

Proceedings of COLING 2012: Technical Papers, pages 325–340,
COLING 2012, Mumbai, December 2012.

1 Introduction

Segmenting a stream of sounds into discrete words is one of the first tasks that children acquiring their native language have to tackle. Computational models of word segmentation enable us to study this problem in a controlled and detailed manner, allowing for example for an examination of the usefulness of different kinds of cues or different learning strategies. Just as important as the actual models, however, is the adequacy of the input used to evaluate them — if we are interested in answering questions about human language acquisition, the data we evaluate our models on needs to be comparable to what children are likely to have access to. To this end, several datasets of phonemically transcribed child directed speech (CDS) have been constructed in several languages, ranging from English to Italian, Polish, Sesotho and Chinese (Brent and Cartwright, 1996; Gervain and Erra, 2012; Boruta and Jastrzebska, 2012; Johnson, 2008a; Johnson and Demuth, 2010). In addition to cross-linguistic variation, however, adequate computational models also need to handle language-internal variation along several dimensions, a topic that has so far received little interest. In this paper, we look at a basic point of variation, namely the actual size of the input to the learner. The longer children are exposed to language, the more data they are exposed to and the better at their language they become, something one would expect from adequate models of language acquisition as well.

We run our experiments on a novel dataset that contains longitudinal data for six children from the Providence Corpus (Demuth et al., 2006). It has two advantages over the current defacto standard for word segmentation studies for English, the Bernstein-Ratner-Brent corpus (Brent, 1999, in the following, BRB Corpus). First of all, it cleanly separates the CDS that is directed at different children whereas the BRB Corpus contains data from 9 different children with no clear indication of the different portions. In addition, recording for some of the children in the BRB corpus began as late as month 21 and for others as early as month 13, making the data available for some of the children hard to compare. In contrast, the Providence Corpus provides data for all of the children starting from month 16 at the latest and starting from month 11 at the earliest and thus constitutes a much more homogeneous data set. Finally, the overall size of the BRB Corpus with a little less than 10,000 utterances altogether is relatively small, in particular if one considers the individual sub-parts which, on average, have only around 1,100 utterances; in contrast, our dataset contains more than 90,000 CDS utterances in total and spans a period of several months for all of the children. This makes it possible to both compare inter-child variability in word segmentation across comparable situations and to study developmental changes in individual children over a period of several months. As such, the resource will allow researchers to ask a wider range of questions than is currently the norm, in particular with respect to the study of incremental models that have recently received a lot of interest (Pearl et al., 2011; Börschinger and Johnson, 2012; Phillips and Pearl, 2012). While for our own experiments on the effects of input size we focus on one of the six sub-corpora of the dataset, we describe and make available the full data so as to enable other researchers to take advantage of this new resource as well.

The contribution of this paper is two-fold. We present a new CDS corpus for computational word segmentation that is derived from the Providence Corpus (Demuth et al., 2006), comprising data from six different children that have been collected in comparable situations over several months. The second and major contribution of our paper is the identification of a so far unreported "overlearning" effect of certain word segmentation models that runs counter to the plausible intuition that more data should lead to better learning outcomes. We also discuss these findings and propose an explanation for the behaviour exhibited by different models, highlighting the importance of linguistic structure for computational models of language acquisition. The outline of the paper is as follows. First, we provide background about the original Providence Corpus, our way of phonemically transcribing it and the properties of the new data-set we created. In section 3, we introduce the models of word segmentation which we examine with respect to the effect of input size in section 4. Section 5 discusses our findings and the final section concludes.

2 The Providence Corpus

The Providence Corpus was collected during 2002-2005 from participants in southern New England. It contains longitudinal audio/video recordings of 6 monolingual English-speaking mothers and their children from approximately 1-3 years during spontaneous interactions at home. The children included 3 boys (Alex, Ethan, William) and 3 girls (Lily, Naima, Violet). Each was recorded for approximately 1 hour every 2 weeks beginning at the onset of first words. Two of the girls have denser corpora, with weekly recordings from 1;3-2;10 (Naima) and 2;0-3;0 (Lily), and Naima's recordings tended to be 1.5 hours long. There is therefore more data for this mother and child. Lily's mother also talked quickly; there is therefore much data from Lily's mother as well. Recording began around one year or once the parent reported that the child was producing approximately four words.

Digital audio/video recordings took place in each child's home. Although parent and child could move freely about, the video information was useful in determining the context of what was being discussed, including possible target words. The availability of video would allow future work along the lines of Frank et al. (2009) and Jones et al. (2010) although so far, we haven't made direct use of the video recordings.

The digital audio/video recordings were downloaded onto a computer, and both adult and child speech were orthographically transcribed using CHAT conventions (cf. MacWhinney (2000)). The child data — but unfortunately not the caregivers' — were then also transcribed in phonemic transcription. All mother and child transcriptions, as well as audio/video files, can be found on the CHILDES database `http://childes.psy.cmu.edu/`. We used the XML version of the data for our transcription process.

2.1 Producing a phonemically transcribed version

To find CDS utterances, for all six children we extract the orthographic transcriptions for all utterances made by caregivers from the XML transcripts of the Providence corpus, starting from 11 months up to and including 22 months.[2] This makes our data qualitatively comparable to the BRB Corpus that includes CDS from between 13 and 21 months of the children's age. In total, we extract 101,451 utterances with a 9395 distinct (orthographic) types but the number of utterances we use is smaller than this because we do not keep all of the utterances in the process of transcription (see below).

To turn the orthographic representations into a phonemic format that is suitable for studying language acquisition, we perform a four step process of filtering, dictionary look-up, heuristic construction of pronunciations for unknown types and manual transcription of unknown types not covered by our heuristic as well as correction of mistakes made during earlier steps.

[2] Available at `http://childes.psy.cmu.edu/data-xml/Eng-USA/Providence.zip`

2.1.1 Filtering words

We manually remove types that we consider to be obvious non-words, in particular interjections such as *hmmmhmmm* or *mmmmhmmm*, obvious onomatopoeic wordplay such as *nanana* and unintelligible words which are transcribed in the Providence Corpus as *xxx* and *yyy*. This is consistent with the procedure followed by Brent (1999) and, more recently, Boruta and Jastrzebska (2012), making our corpus comparable in this respect to theirs. We do not, however, remove these items in cases where the resulting utterance would have been rendered fully unintelligible or where a word that should have been excluded according to the above criteria was used as an actual word in a large number of cases.[3]

In total, we identify 785 such non-words and we remove all occurrences of these types from the transcript, leaving the remaining words in the utterance:[4] utterances including any of the stop-words are still transcribed as long as there is at least one word left after removing all stop words. A total of 7,362 utterances are thus completely ignored, with 6,123 utterances consisting of exactly one of these filtered elements, in particular *xxx* (unintelligible, 2215), *oh* (525) and *hmmm* (521).

2.1.2 Dictionary lookup

After filtering, we perform a simple dictionary-lookup transcription using a phonemic dictionary. We use the current version of the VoxForge dictionary which uses a standard phone set for American English, corresponding to the current DARPABET coding.[5] We also provide a script that maps this representational scheme into one-character-per-phoneme representations that are required by some of the currently common word segmentation tools.[6]If there are multiple pronunciations available for a type, we always pick the first one. While this constitutes an idealization we believe that a well-understood and explicit idealization is to be preferred over an overly simplistic method of artificially introducing variability such as randomly choosing a pronunciation.

In total, the VoxForgeDictionary covers 7035 of the 8610 remaining types in the data, leaving 1575 of the types untranscribed. We transcribe these words manually, using a simple preprocessing heuristic to aid the process.

2.1.3 Heuristically constructing pronunciations for unknown words

Many of the unknown words are either forms of types that already are in the lexicon, e.g. possessives (*Elmo's*) or plurals (*Legos*), or compounds of two types that are both in the lexicon individually (*frenchtoast*, *teddybear*). We handle the former case by simple rules operating on the orthographic forms directly, looking for possible plural or possessive endings and then checking whether the orthographic form can be decomposed into a known base-form with the

[3] The former applies mostly to cases where an item is mentioned rather than used, e.g. ”Does the baby say ‘Wah wah’?”; the latter, for example, applies to ‘bonk’ which, in addition to its onomatopoeic use, also occurs as a verb in the corpus, including its preterite and participle. Our data includes the full list of filtered items as well as the scripts that perform the automatic steps of transcription from the original xml-data so that researchers can easily make their own decisions about which items to exclude.

[4] While this may seem like a lot of items to exclude, most of these are hapaxes like *bumpoopadoompadadooboom* or *doodleuhdoo*.

[5] The dictionary is available at `http://www.repository.voxforge1.org/downloads/SpeechCorpus/Trunk/Lexicon/VoxForge.tgz`

[6] E.g. dpseg (Goldwater et al. 2009).

Name	#Utt	#Tok	#Type	Avg. Utt. Len.	Avg. Tok. Len	Avg. Type Len.
Alex	8330	29423	1877	11.00	3.12	4.58
Violet	9024	39135	2343	13.43	3.10	4.68
William	10697	45689	2061	13.01	3.05	4.59
Ethan	18020	75564	2999	13.11	3.13	4.69
Lily	20641	94696	3946	14.77	3.22	4.96
Naima	27377	141990	4579	16.51	3.18	5.05

Table 1: Statistics about the different sub-corpora.

suffix added. In this case, we add the corresponding morph according to the rules of English phonology. To identify potential compounds, we try to decompose words into a prefix p and a maximal suffix s such that both p and s are known forms.

Taken together, our heuristic detects 924 cases which we then manually correct for mistakes. An alternative way of constructing pronunciations is to use letter-to-sound rules, something we plan to experiment with in future work.

2.1.4 Manual transcription

The remaining 651 word types are labelled manually, using where available the form-annotation in the XML files as guide-line.[7] Finally, we also manually inspect the forms generated by the automatic steps and correct mistakes.

2.2 Statistics

The final corpus comprises a total of 94,089 phonemically transcribed utterances and consists of six distinct sub-corpora, each corresponding to the CDS directed at one of the six children. Each sub-corpus is, in turn, subdivided into individual files corresponding to the age of the child at which recording took place, ranging from 11 up to 22 months. Both within these individual files and within the overall corpus, the order of the CDS utterances corresponds to the order in which these utterances were actually made, making them suitable for studies that look at changes over time.

Table 1 gives summary statistics over the full amount of data for each individual child. Looking at total number of utterances, we can broadly identify two groups: for Alex, Violet and William there are considerably less CDS utterances than for Ethan, Lily and Naima. This is presumably mostly due to the factors mentioned above. Yet, there also seem to be noticeable differences in terms of utterance-, token- and type-length. Although we will not do so in this paper, performing comparative evaluation of models across the children may lead to the discovery of interesting predictors of model performance and perhaps even actual language ability on behalf of the children.

For the rest of the paper, we will focus on the Naima part of the corpus and take a closer look at how the segmentation performance of different segmentation models changes as a function of the size of the input.

[7] While not always provided for caregiver utterances, some of them include phonetic markup for individual words, in particular if the words were names.

3 Bayesian Word Segmentation

The word segmentation models we study in this paper are Goldwater's Unigram and Bigram model (Goldwater, 2007; Goldwater et al., 2009) and Johnson (2008b)'s collocation-syllable Adaptor Grammar models. For the mathematical details of these models, we refer the reader to these original works.

All the models are Bayesian probabilistic models that define a generative process for the target of learning, in this case segmented utterances or sequences of words. This generative process is defined with the help of the Dirichlet Process (DP):[8] at an intuitive level, the DP can be used for word segmentation because it can bias models to identify compact ways to represent the observed unsegmented utterances, trading off the number of both tokens and types used in an analysis of the data. This trade-off is a consequence of the way that probabilities are assigned to tokens under a DP model: the probability of hypothesizing a word token[9] depends on the number of times that its type has previously been hypothesized, and the probability of a full hypothesis or segmentation of the data is the product of the probabilities for all the tokens used in the segmentation. This tends to make solutions in which a small number of words is used relatively frequently yet not over-excessively (as would be the case if every individual segment were a type) the most probable which, in most cases, also leads to linguistically reasonable results. What differentiates the different models from one-another are the specific assumptions about *the nature of possible word types* and *the relationships between word tokens*. We will quickly elaborate on these points, thus introducing all models used in our experiments.

3.1 Assumptions about possible words

A naive assumption about possible words is that they can be any arbitrary sequence of segments. Thus, both *dog* and *qfx* would be equally good candidates for possible words a priori. While obviously not true of human languages (*bnik* isn't a possible English word), such a unigram phoneme model is embodied in the original Unigram and Bigram models (Goldwater et al., 2009) and has been shown to work reasonably well on the BRB Corpus.

An alternative and linguistically more adequate assumption is that a word must conform to (arguably universal) constraints of syllable structure: words aren't just sequences of segments but consist of one or more syllables, themselves structured entities that consist of an optional initial consonantal onset, an obligatory vocalic nucleus and an optional consonantal coda (Smolensky and Legendre, 2005). In fact, Norris et al. (1997) provided evidence for a constraint on possible words along these lines in human segmentation of unsegmented speech and Johnson (2008b) showed previously how incorporating this kind of 'prior knowledge' into word-segmentation models can lead to a considerable improvement in segmentation accuracy.

From a language acquisition point of view, there are also reasons to assume that syllables and sub-syllabic units are learned by infants in addition to entire words: there is strong evidence that even very young infants track probabilities defined over entire syllables (Saffran et al., 1996) and that infants are sensitive to the phonotactics of their language from early on (Jusczyk et al., 1993), leading us to look at models that learn both entire syllables and sub-syllabic units. In addition, many languages including English are more or less restrictive about the material allowed to occur word-initially and word-finally than word-internally, having led previous work

[8] Adaptor Grammars actually use the Pitman-Yor Process, a strict generalization of the DP. Here, we gloss over this detail.

[9] Or a token of another entity, e.g. a syllable or a multi-word expression, if the model incorporates these notions.

such as Johnson (2008b) to distinguish between word-internal and word-peripheral onsets and codas.[10] The models we discuss, then, incorporate all of these ideas, that is they learn sub-syllabic units that are distinguished as to whether they occur word-peripherally or internally as well as entire syllables. The Adaptor Grammar framework makes it very easy to implement these different models, and we refer the reader to Johnson (2008b) for a discussion of how this is done.[11]

3.2 Assumptions about relations between word tokens

The simplest assumption about the relation between words within an utterance is probably that there is none. For a probabilistic model, this leads to a Unigram assumption about words, i.e. that the probability of a sequence of words is simply the product of the probability of each individual word, irrespective of its specific context. This is the assumption embodied by Goldwater's original Unigram model and a lot of previous work on word segmentation (Brent, 1999; Venkataraman, 2001).

A more plausible assumption is that words in a sequence are predictive of each other, a simple version of which can be spelt out as a Bigram assumption about words that is employed by Goldwater's Bigram model.

Yet another way of modelling the relation between words has been proposed by Johnson (2008b), employing a hierarchical instead of a sequential notion of context. The *collocation* models assume that sentences are sequences of *phrases* that, themselves, are made up of individual words. Consequently, the model not only learns words but an additional kind of entity, entire chunks of words. Importantly and in contrast to the Unigram and Bigram models, these chunks are stored in addition to and not at the expense of the words that make them up — a collocation model can infer that *thedoggie* is a high-frequent unit that is made up of the distinct words *the* and *doggie*. Johnson's Adaptor Grammar framework makes it easy to assume multiple levels of such collocations: a collocation2-grammar, for example, learns, in addition to words and collocations composed of words, collocations that are themselves composed of collocations of words. In this paper, following Johnson and Goldwater (2009) we examine models that use up to three levels of collocations.

3.3 The different models

As the different kinds of assumptions about the internal structure of words and about the relationship between words are independent, we can freely combine them, yielding a total of 9 models.[12] We refer to the different models by names that indicate which assumptions they embody. Models that do not incorporate the syllable structure assumption are simply referred to by the relationship assumed between word tokens, i.e. *unigram*, *bigram*, *colloc*, *colloc2* and *colloc3*, the last two names referring to models assuming a total of 2 and 3 levels of collocations, respectively.

Except for the bigram model, each of these models can either use a 'naive' word-assumption

[10] For example, in English *str* is only valid as an onset word-initially, as in *string* or *strong*, and the consonant cluster *dths* in *widths* is restricted to the end of words.

[11] In fact, the sole difference between Johnson's models and our own is that his do not learn entire syllables.

[12] With the exception of the bigram model for which we know of no existing implementation that makes it possible to incorporate syllable structure. In principle, however, this model is a possibility and we hope to be able to study it in the future.

allowing arbitrary sequences of segments or the syllable-structure word-assumption which, in addition to constraining the space of possible words, enables the model to explicitly learn properties of the language's phonology. To distinguish these two cases, we suffix models that embody the latter assumption with *Syll*.

4 Experiments

We are interested in how the segmentation performance of the different models varies with the amount of input available to the models during learning. As human language learners tend to get better at their native language with longer exposure, one would expect adequate computational models to exhibit something similar, initially improving as more data is observed and, at some point (probably beyond the size of samples we usually can look at in practice), asymptotically approaching some upper bound.

The longitudinal data available in the Providence Corpus suggests a natural setup for studying this question by constructing inputs that consist of all CDS utterances directed at an individual child up to a certain point of time. For our experiments, we use the Naima section of the Providence Corpus and collect CDS utterances from when Naima was 11 months old through to when she was 21 months old to construct 11 differently sized inputs, each input consisting of all CDS utterances in the corpus up to and including a given month. We will refer to the different input sets by the last month from which it includes data and will use "language exposure" and "input size" exchangeably, a simplifying yet justified choice as is evident from figure 1 that shows how the size of the input grows over time and thus with language exposure.

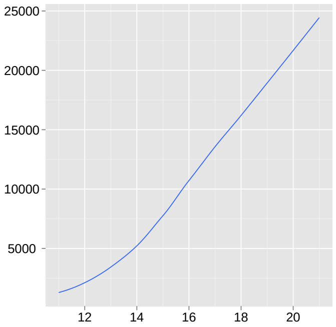

Figure 1: The months on the x-axis correspond to the different inputs, the y-axis gives the number of utterances in an input. The smallest input is month 11 with 973, the largest input month 21 with 24,327 utterances, approx 2.5 times the size of the BRB corpus.

We choose to evaluate on held-out data and construct the test-set by sampling 200 CDS utterances from the 22nd month of each of the six children's sub-corpus in the Providence Corpus. We propose this as a standard test corpus that can be used as a standard for any or all of the 6 children's data in the Providence Corpus. Our evaluation metric is token f-score, the harmonic mean between token precision (number of correct tokens identified by the model over the number of total tokens predicted by the model) and token recall (number of correct tokens identified by the model over the true number of tokens in the input) as a measure of segmentation accuracy. The segmentation on the held-out data is calculated after probabilistic inference has been performed on the input, thus implicitly defining a (probabilistic) lexicon according to which we sample a segmentation for each utterance in the test-set. Note that during this process, no novel words are added to the model's lexicon; in this sense, we evaluate the knowledge the learner has acquired after having had access to the input.

Our experimental procedure follows closely the one outlined in Johnson and Goldwater (2009). We use the current version of Mark Johnson's Adaptor Grammar implementation[13] to run

[13] Available at `http://web.science.mq.edu.au/~mjohnson/code/py-cfg-2012-08-16.tgz`.

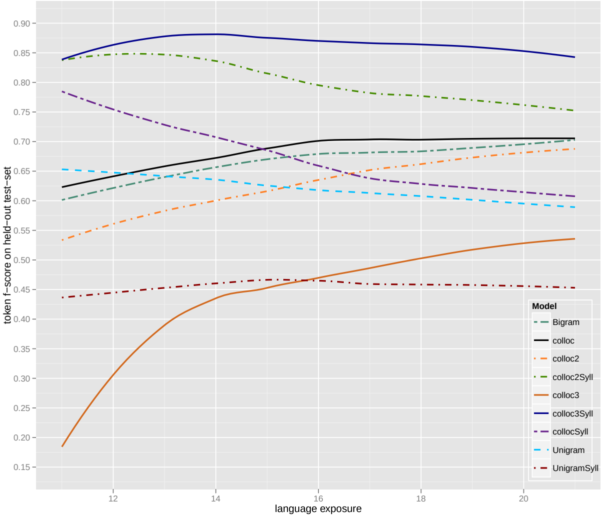

Figure 2: Word segmentation performance on the held-out test-set for the difference models, as a function of language exposure and consequently, size of the training set (see figure 1). The models incorporating syllable structure all exhibit some degree of performance decrease for larger inputs although the colloc3Syll consistently remains well above 80%. In contrast, the models without syllable structure exhibit an increase in performance over time although their segmentation accuracy is considerably worse than that of the best performing models with syllable-structures.

two Markov Chain Monte Carlo chains for each of the models for 1000 iterations, collecting sample analyses for the held-out test-set with a lag of 5 after a burn-in of 800 iterations and performing Minimum Bayes Risk decoding to get a single score for each condition at the end. For the Bigram model, the only model not expressible as an Adaptor Grammar, we use Sharon Goldwater's implementation and the experimental paradigm outlined in Goldwater (2007) and Goldwater et al. (2009), using simulated annealing for 20,000 iterations and evaluating on a single sample taken at the end. Figure 2 plots segmentation accuracy as measured by token f-score for different amounts of input size, with the size of the input growing from left to right.

Overall, we can see two broad patterns of behaviour across the different models. One group of models exhibits a degradation in performance for larger amounts of inputs, in particular the Unigram, the collocSyll and the colloc2Syll model. Neither the UnigramSyll nor the colloc3Syll model show a lot of variation although a small drop in performance for the last month as compared to their peak performance is noticeable for them as well. Another group consisting of the colloc, colloc2, colloc3 and Bigram models exhibits an increase in performance with amount of training data.

The colloc3Syll model clearly emerges as the most accurate with an accuracy of 87% at peak performance at around 14 months (4794 utterances) that drops to around 83% at month 21 (24,327 utterances). The second best model, the colloc2Syll model, peaks at around 85% at month 13 and drops by 10% to about 75% at month 21, showing a clear decrease in accuracy. A third place is harder to assign. The collocSyll model shows the most dramatic drop in performance, from around 75% at month 11 (973 utterances) to just above 60% for month 21. In contrast, both the colloc and the Bigram model start around 60% at month 11 but, for month 21, reach around 70% so that there is no single model that comes in third for all amounts of input size. Incidentally, note how the behaviour of the Bigram and the colloc model is very similar, supporting Johnson (2008b)'s hypothesis that "the collocation word adaptor grammar can capture inter-word dependencies similar to those that improve the performance of Goldwater's bigram segmentation model."[14] The most dramatic performance improvement is seen for the colloc3 model which jumps from around 18% for month 11 to about 53% for month 21. Despite this, it is the second-worst model, loosing only to the UnigramSyll model which stays consistently below 50% accuracy with little variation.

To sum up, we observe two types of behaviour. The Unigram model and the models incorporating syllable structure and the notion of a multi-word phrase exhibit what we call "overlearning": they reach their peak performance for relatively small amounts of input and gradually get worse as the size of the input grows larger. This is much more pronounced for the collocSyll and the colloc2Syll than for the colloc3Syll model, with the latter remaining well above 80% accuracy even for the largest amount of input. Despite overlearning, the colloc3Syll and the colloc2Syll perform word segmentation the most accurate for all sizes of input. On the other hand, the models lacking the assumption of syllable structure do not exhibit overlearning, at least not on the amount of data we were able to test them.

5 Discussion

Why do the models we examine exhibit these two kinds of behaviour? We begin with a detailed explanation of the "overlearning", starting from an original observation going back to Goldwater (2007) who noticed that the Unigram model tends to identify undersegmented solutions where the predicted (incorrect) words often consist of several of the (correct) words. Her explanation for this is that "groups of words that frequently co-occur violate the unigram assumption in the model since they exhibit strong word-to-word dependencies", and that "[t]he only way the model can capture these dependencies is by assuming that these collocations are in fact words themselves." (Goldwater, 2007, p.72) Why is it, however, that these "misleading" co-occurrences occur in the data in the first place, and can this explanation be extended to account for the input-size effect we detected?

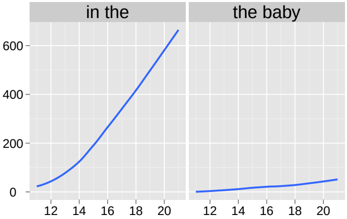

Figure 3: Frequency of a preposition+determiner pattern and determiner+noun pattern on the y-axis as a function of input size on the x-axis. Note the steep increase of frequency of *in the* and the much less dramatic increase for *the baby*.

We suspect that many of the "collocations" a model such as the Unigram model is susceptible to arise from principled regularities governing language: English syntax, for example, requires (most) prepositional phrases to begin with a preposition-determiner sequence and both preposi-

[14] Unlike Johnson (2008b) and Goldwater (2007), we did not hand-tune any of the parameters of the models which may partly explain why our scores for the Bigram model are slightly lower for the earlier months, in addition to the fact that we didn't use Minimum Bayes Risk decoding for the Bigram model.

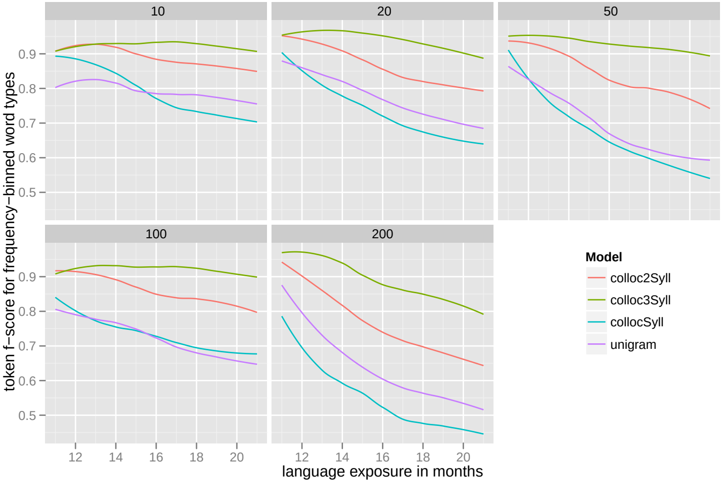

Figure 4: Token f-scores for word types of different frequencies in the test-set as a function of the size of the input. Note that the Unigram and the collocSyll model already show clear decreases in accuracy over time for types of frequency larger than 10 whereas the colloc3Syll model shows a relatively robust performance up to the highest-frequency bin of types with frequency larger than 200. The presence of the drop in performance shows that even the colloc3Syll model suffers from "overlearning" although this behaviour is much less pronounced than for the other models and only occurs dramatically for the highest-frequency types in the data.

tions and determiners are small closed classes, leading to a huge number of sequences such as *in the* in virtually any English text. Crucially, the occurrence of a sequence such as *in the* is largely independent of what is actually being talked about and the number of such occurrences can therefore be expected to grow with the amount of data considered. This is supported by figure 3 which plots the change in frequency of the preposition-determiner sequence *in the*, showing that the number of "misleading" co-occurrences of that kind does indeed increase for larger amounts of data. This leads to the prediction that the Unigram model will perform worse when trained on larger amounts of data, simply because the evidence for these kinds of spurious words that lead to undersegmentation errors grows with the input size. To our knowledge, we are the first to formulate this hypothesis and our experimental results strongly suggest that it is true. The drop in performance for the Unigram model is clear from figure 1: it reaches peak performance of around 65% when its input consists of a mere 973 utterances, and its segmentation accuracy steadily drops as it processes larger inputs down to around 58% for an input of 24,327 utterances.

More direct support for explaining the drop by the negative impact of the increasing frequency of patterns like the one in figure 3 comes from figure 4 that plots how well the model is able to identify word types of different frequencies in the test set as a function of input size. Note how for higher-frequency types, the Unigram model's performance decreases more dramatically than for low frequency types for larger amounts of input. To investigate whether this difference in

pattern	#test	month 11 (973 utterances)					month 21 (24,327 utterances)				
		#input	Unigram		colloc3Syll		#input	Unigram		colloc3Syll	
			% cor	% col	% cor	% col		% cor	% col	% cor	% col
in the	18	14	50%	50%	94%	0%	671	0%	89%	0%	100%
are you	19	14	100%	0%	100%	0%	536	0%	71%	0%	100%
on the	21	11	100%	0%	81%	0%	347	0%	86%	14%	86%
this is	9	5	100%	0%	100%	0%	196	0%	67%	56%	44%
with the	4	2	100%	0%	100%	0%	153	0%	75%	75%	0%
the baby	4	0	100%	0%	100%	0%	80	0%	100%	100%	0%

Table 2: A qualitative look at how bigram-patterns of different frequency are handled by the Unigram and the colloc3Syll model at month 11 and month 21, respectively. The %cor and %col columns give the percentage of occurrences of a pattern in the test-set that were handled correctly or were misanalysed as constituting a two-word collocation. Note that %cor and %col need not add up to 100% as the models can also make other errors than simply undersegmeting. It is clear that at month 11, both the Unigram and the colloc3Syll model handle all the cases nearly perfectly; but unlike the Unigram model, the colloc3Syll model handles lower-frequency patterns at month 21 with increasing ease, making no mistake for a unit that occurs 80 times in its input such as *the baby* but still mis-analysing a pattern such as *in the*.

performance could actually be due to high-frequency items getting "absorbed" into larger units as we suggest following Goldwater (2007), in table 2 we perform a qualitative evaluation of a number of actual examples of patterns involving high-frequency that are themselves of different frequencies. As is clear, cases that are analysed correctly at month 11 are almost consistently misanalysed as a single word at month 21, showing that the "loss" of high-frequency items is a major reason for the overlearning.

Surprisingly, perhaps, the same kind of explanation seems to apply to the collocation-syllable models. While originally proposed by Johnson (2008b) to specifically address the problem of undersegmentation, looking at figure 4 indicates that collocation models do not solve the problem of high-frequency words completely, although it seems to get less problematic with the number of collocational levels the model has at its disposal. Looking at the kind of units learned by the collocation-models helps understand why that is: among the high-frequency collocations learned by both the collocSyll and the colloc2Syll model from the largest input is, for example, the two 'word' sequence *doyou remember*; this is a better solution than the Unigram model's *doyouremember* but it still involves the undersegmented collocation *do you*. While the colloc3Syll model analyses this specific case correctly as a three-word collocation *do you remember*, it fails to acquire the collocation *do you* on its own and prefers to use the "word" *doyou* in most other cases, showing that collocations do not solve the undersegmentation problem but only push it back a level, in line with figure 4: note that while the colloc3Syll model behaves relatively stable for word-types with frequencies smaller than 200, it also shows a marked drop for the high-frequency word types from around 95% at month 11 to just below 80% for month 21. Some concrete examples of patterns which the colloc3Syll model is and isn't able to handle correctly are given in table 2, alongside the performance of the Unigram model for these cases, showing how its ability to handle word-dependencies breaks for high-frequency patterns but handles patterns the Unigram model is unable to analyse correctly. After this discussion, the fact that even for collocation-models the undersegmentation problem gets worse for larger inputs shouldn't be too surprising. As pointed out above, many of the patterns leading to undersegmentation errors are due to syntactic regularities that, for example, require prepositions to be followed (in almost all cases) by articles. Figure 3 indicates that these kinds of patterns grow continuously with the size of the input, suggesting that models that "merely" model co-occurrence statistics are bound to fail at some point. This may almost seem like a general problem for Bayesian probabilistic models of the kind discussed here that, in a sense, simply try to identify high-frequency patterns in the input. Yet this is not so. For one thing, the lack of detailed linguistic structure is not inherent to the Bayesian framework that is fully unrestricted as to what kind of structures a model is defined over. This much is clear from the ease with which syllable structure can be incorporated into the models. Secondly, even without additional linguistic structure the relative robustness of the colloc3Syll model shows that while not fully solving the problem of misleading co-occurrences, a sufficiently rich collocational structure goes a long way in alleviating the problem for input sizes that go well beyond 20,000 utterances. It suffices to handle most cases involving content words such as nouns, correctly learning for example that despite its (relatively) high frequency, *the baby* consists of two individual words. Figure 3 shows that for patterns like this, frequency grows much slower over time (although still too fast for a model lacking any ability to model larger-than-word-units such as the Unigram model), not too surprising considering that the occurrence of content words — unlike function words — is mainly dependent on what is actually being talked about and that conversation topics tend to change over time. This then suggests that a model like colloc3Syll will handle correctly these important cases for even considerably larger amounts of data, a clear desideratum for models of early language acquisition phenomena.

This leaves us to explain why the collocation models without syllable structure lead to an overall worse performance but seem to exhibit a positive relation between amount of input and segmentation accuracy. The key to this, we believe, lies again in considering the kinds of regularities the model is sensitive to. With no restriction on what an actual word may look like, high-frequency patterns of any kind — including individual segments and short n-gram like sequences of segments — can be employed by the models to explain the input they get. In particular for little input with overall few word tokens and, consequently, relatively few repetitions for each of the actual word types, the evidence for high-frequency non-words is extremely high, leading to over- rather than undersegmentation. To understand why this isn't happening for the Unigram model, recall that under the Unigram model, all tokens in a hypothesis are fully independent whereas the collocation and the Bigram model assume dependencies between tokens. Consequently, under the Unigram model a solution involving a large number of items is automatically discouraged because it can only work with marginal probabilities, whereas the other models without syllable structure can "abuse" their respective notions of context to capture dependencies not only between words but also between sub-word units. Ironically then, the oversegmentation behaviour is worst for models with a lot of additional structure such as the colloc3-model that, when combined with a syllable structure constraint, leads to the best performance. A quick glance at the kinds of units identified by it suggests why that is — at month 11, the most frequent "word" that is learned is *t* which is used in "collocations" like *t o*, *ge t* or, illustrating the problem very nicely, *j us t*. Crucially, the more input it has access to the stronger the evidence for larger (more word-like) units gets, leading to less undersegmentation at month 21 with a top-5 word list of *ing*, *you*, *z*, *the* and *to*. Yet, even for well over 20,000 utterances the colloc3 model prefers extremely short 'words' as it has another three levels of collocations that it can use to capture words; and it actually does learn collocations such as *ma mi*. The colloc2, colloc and the Bigram model are less extreme in their oversegmentation behaviour as they have fewer "levels" at their disposal, discouraging the excessive use of one-segment 'words' as in the colloc3 model more severely. In fact, as is suggested from both figure 2 and from inspecting their highest-frequency words at month 21 that include “overlearned” units like *areyou* and *doyou*, the Bigram and the colloc model are starting to overlearn just like the syllable-structure models, as we would expect from basically every model ignorant of the linguistic regularities that can explain high-frequency patterns for sufficiently large amounts of data. In conclusion, we think this shows that the intuitively attractive behaviour of these models is an artefact of their strong preference for short units that, for small amounts of data, masks the overlearning at the expense of segmentation accuracy; in particular, the fact that models lacking syllable structure do not exhibit overlearning in our experiments should not be taken as evidence that they are more adequate than their overlearning relatives.

It is possible that the overlearning behavior we describe in this paper can be addressed by manually choosing appropriate values for the models’ hyper-parameters, which control the models’ Pitman-Yor Processes (PYPs). Goldwater (2007) observed that the segmentations proposed by the unigram and bigram models depend on the choice of hyper-parameters, and Johnson (2008b) observed a similar sensitivity to hyper-parameters in Adaptor Grammar models. However, the number of hyper-parameters is twice the number of PYPs in a model, and searching for hyper-parameter values that result in the most accurate segmentation is computationally very challenging. For this reason Johnson and Goldwater (2009) placed Beta and Gamma priors on the PYP a and b hyper-parameters respectively in their Adaptor Grammar models, and reported that sampling the hyper-parameters actually improved segmentation f-score as compared to the manually-specified settings they investigated. It is reasonable to expect that more realistic models of human language will be more complex, and therefore will have even more hyper-parameters, than the models investigated here. It is certainly conceivable that the hyper-parameters are innately fixed by universal grammar to values that result in good segmentations across different languages. But all else being equal, models which do not require hyper-parameters to be fixed to specific values should be preferred on general simplicity grounds over models that do require such prespecification. For this reason we chose to use Johnson and Goldwater’s hyper-parameter sampler in the Adaptor Grammar models we studied here.

Conclusion and perspectives

We have presented a novel corpus of English CDS derived from the Providence Corpus for studying models of word segmentation. This corpus makes it possible to address a wider range of questions than is currently common, for example with respect to the study of developmental effects in incremental algorithms. We identified an interesting “overlearning” effect for state-of-the-art word segmentation models on large amounts of data that has so far gone unnoticed and proposed an explanation of this behaviour that highlights the importance of linguistic structure for Bayesian models; we have argued that the apparent lack of overlearning for linguistically less-sophisticated models is due to an undesirable preference for oversegmentation and shouldn’t be mistaken for an advantage.

In future work, we want to further explore the impact of linguistic knowledge for Bayesian models of the kind discussed here; we suspect that giving the model the ability to model more of the regularities languages exhibit more appropriately, the overlearning behaviour we observed will become less severe. Also, we want to test the cross-linguistic usefulness of different kinds of constraints such as assuming (a certain kind of) syllable-structure. This is complicated by the fact that most current data-sets do not separate CDS directed at different children and children of different ages. We hope that our presentation of a data-set that fulfills these desiderata will lead to the creation of corpora like this for other languages as well.

References

Börschinger, B. and Johnson, M. (2012). Using rejuvenation to improve particle filtering for bayesian word segmentation. In *The Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, Jeju, South Korea. Association for Computational Linguistics.

Boruta, L. and Jastrzebska, J. (2012). A phonemic corpus of polish child-directed speech. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, pages 1017–1020, Istanbul, Turkey. European Language Resources Association (ELRA).

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

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

Demuth, K., Culbertson, J., and Alter, J. (2006). Word-minimality, epenthesis, and coda licensing in the acquisition of english. *Language and Speech*, 49:137–174.

Frank, M. C., Goodman, N., and Tenenbaum, J. (2009). Using speakers' referential intentions to model early cross-situational word learning. *Psychological Science*, 20:579–585.

Gervain, J. and Erra, R. G. (2012). The statistical signature of morphosyntax: A study of hungarian and italian infant-directed speech. *Cognition*, (0):–.

Goldwater, S. (2007). *Nonparametric Bayesian Models of Lexical Acquisition*. PhD thesis, Brown University.

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

Johnson, M. (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. Association for Computational Linguistics.

Johnson, M. (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*, pages 398–406, Columbus, Ohio. Association for Computational Linguistics.

Johnson, M. and Demuth, K. (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. Coling 2010 Organizing Committee.

Johnson, M. and Goldwater, S. (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. Association for Computational Linguistics.

Jones, B. K., Johnson, M., and Frank, M. C. (2010). Learning words and their meanings from unsegmented child-directed speech. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 501–509, Los Angeles, California. Association for Computational Linguistics.

Jusczyk, P., Friederici, A., Wessels, J., Svenkerud, V., and marie Jusczyk, A. (1993). Infants' sensitivity to the sound patterns of native language words. *Jounral of Memory and Language*, 32:402–420.

MacWhinney, B. (2000). *The CHILDES Project: Tools for analyzing talk, 3rd Edition. Vol 2: The Database*. Lawrence Erlbaum Associates.

Norris, D., McQueen, J. M., and Cutler, A. (1997). The possible-word constraint in the segmentation of continuous speech. *Cognitive Psychology*, 34:191–243.

Pearl, L., Goldwater, S., and Steyvers, M. (2011). Online learning mechanisms for Bayesian models of word segmentation. *Research on Language and Computation*, 8(2):107–132.

Phillips, L. and Pearl, L. (2012). "Less is more" in Bayesian word segmentation: when cognitively plausible learners outperform the ideal. In Miyake, N., Peebles, D., and Cooper, R. P., editors, *Proceedings of the 34th Annual Conference of the Cognitive Science Society*, Austin, Texas. Cognitive Science Society.

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

Smolensky, P. and Legendre, G. (2005). *The Harmonic Mind: From Neural Computation To Optimality-Theoretic Grammar*. The MIT Press.

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