Unsupervised Word Segmentation in Context

Gabriel Synnaeve and **Isabelle Dautriche**
LSCP, DEC
ENS Ulm, Paris, France
gabriel.synnaeve@gmail.com
isabelle.dautriche@gmail.com

Benjamin Börschinger
Institut für Computerlinguistik
Universität Heidelberg, Heidelberg, Germany.
benjamin.boerschinger@gmail.com

Mark Johnson
Department of Computer Science
Macquarie University, Sydney, Australia
mark.johnson@mq.edu.au

Emmanuel Dupoux
LSCP, DEC
EHESS, Paris, France
emmanuel.dupoux@gmail.com

Abstract

This paper extends existing word segmentation models to take non-linguistic context into account. It improves the token F-score of a top performing segmentation models by 2.5% on a 27k utterances dataset. We posit that word segmentation is easier in-context because the learner is not trying to access irrelevant lexical items. We use topics from a Latent Dirichlet Allocation model as a proxy for "activities" contexts, to label the *Providence* corpus. We present Adaptor Grammar models that use these context labels, and we study their performance with and without context annotations at test time.

1 Introduction and Previous Works

Segmentation of the speech stream into lexical units plays a central role in early language acquisition. Because words are generally not uttered in isolation, one of the first task for infants learning a language is to extract the words that make up the utterances they hear. Experimental research has shown that infants are able to segment fluent speech into word-like units within the first year of life (Jusczyk and Aslin, 1995). How does this ability emerge? There is evidence that infants use a broad array of linguistic cues to perform word segmentation (e.g., phonotactics (Jusczyk et al., 1993a), prosodic information (Jusczyk et al., 1993b), statistical regularities (Saffran et al., 1996)). Past experimental and modeling research on speech segmentation has mainly focused on linguistic cues, treating them as independent from other non-linguistic cues naturally occurring in the child learning environment. Yet, language appears in context and is constrained by the events occurring in the daily life of the child. For example, during an eating event one is most likely to speak about food, while during a zoo-visit event, people are more likely to talk about the animals they see. Activity contexts may provide a natural structure to speech that would be readily be accessible to children. A recent study using dense recordings of a single child's language development (Roy et al., 2006) showed that words appearing in specific activity contexts are learned faster (Roy et al., 2012). Relatedly, Johnson et al. (2010) showed that Adaptor Grammars (AGs) performed better on a segmentation task when the model has access to a hand-annotated set of objects present in the environment, that it can use to learn simultaneously word-object associations (see also (Frank et al., 2009)). This supports the view that integrating multiple sources of information, linguistic and non-linguistic, can improve learning.

Following this idea, we posit that information from the broader context in which a word has been uttered may simplify the learning problem faced by the child. In particular, our hypothesis postulates that speech segmentation is easier when using vocabularies that are related to a specific activity (eating,

Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers, pages 2326–2334, Dublin, Ireland, August 23-29 2014.

Table 1: Most probable words in the 7 final topics

egg	book	ball	truck	name	color	block
apple	shape	cat	car	school	bear	battery
banana	square	hat	fire	time	crayon	minute
milk	circle	tree	piece	today	hair	phone
butter	triangle	fish	train	day	head	puzzle
≈food	≈shapes	≈playing	≈toys	≈time	≈drawing	"garbage"

playing...), or place (kitchen, bedroom...). To evaluate this hypothesis, we applied topic modeling (Blei et al., 2003) to automatically derive activity contexts on a corpus of child directed speech, the *Providence* corpus (Demuth et al., 2006), and tested the influence of such topics on a word segmentation task extending the AG models used in (Börschinger et al., 2012). We found that a model augmented with the assumption that words are dependent upon the topic of the discourse (as a proxy for activity context) performs better than the same model without access to the discourse topic. This suggests that the broader context in which sentences are uttered may help in the word segmentation process, and could presumably be used at various stages of language development.

The paper is structured as follows. Section 2 presents a novel approach to augment a corpus with contextual annotations derived from topic models. Section 3 quickly explains Adaptor Grammars, the framework that we used to express all our models. Section 4 presents all the models that were used in the results. Section 5 describes the Providence corpus and the experimental setup. Section 6 shows our quantitative and qualitative results. Finally, we discuss the implications for models of language learning.

2 Topics as Proxies for Contexts

Roy et al. (2012) found high correlations between human-annotated activity contexts and topics from a latent Dirichlet allocation model (LDA) (Blei et al., 2003), thus showing that using topics as proxies for contexts is a sound approach. Topic modeling infers a topic distribution for each "document" (a bag of words) in the corpus. Since "documents" were not annotated in our corpus, we developed the following 3-step approach to automatically segment it into documents.

Firstly, for *all* the children of the Providence corpus, we used recording sessions as hard document boundaries. We considered as a "possible document" every contiguous sequence of sentences separated by at least 10 seconds of silence, according to the orthographic transcript. We also identified "possible documents" using cues such as "bye/hi", indicating a change of participants. This segmentation resulted in an over-segmented corpus (compared to context switches), yielding a total of $16,742$ documents.

Secondly, we used the *gensim* software (Řehůřek and Sojka, 2010) to train a topic model (LDA)[1], and get the topic distributions for each of these documents. We used the symmetric KL-divergence to measure the distance between two topic distributions before and after a "possible document" boundary. If the distance was above a threshold, we considered this boundary as a document boundary. Otherwise we merged both "possible documents" through this silence. The threshold was set empirically to discriminate between two topic distributions that correspond to different activity contexts. After this step, we assume that each of the resulting $8,634$ documents maps to an activity context.

Thirdly, we applied LDA again on this new segmentation to get the topic distribution, hence the activity context, of each document. The number of topics is qualitatively chosen to correspond to the number of main activity contexts (eating / playing / drawing / etc.) that occur in the Providence dataset (we used 7 topics), the resulting most topic specific words are shown in Table 1. Finally, for each document, we got a distribution on topics, and we annotated the document with the most probable topic. By doing that, we throw away graded information about the distribution on topics for each document. We could make use of the full distribution, but here we are only interested in the most probable topic as a proxy for activity context. We do not posit that the infants learn the topic models on linguistic cues while bootstrapping speech and segmentation, but rather that they get activity context from non-linguistic cues.

[1]We did LDA only on nouns (as they contain most of the semantics), weighted by TF-IDF.

3 Adaptor Grammars

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

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

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

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

$$H_X \sim C_X(G_X)$$

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

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

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

4 Word segmentation models

4.1 Unigram model

This most basic model just generates words as sequences of phonemes. As $\underline{Word}$ is underlined, it means it is adapted, and thus we learn a "word unit -like" vocabulary. $Phon$ is a nonterminal that expands to all the phonemes of the language under consideration.

$$Sentence \to Word^+$$
$$\underline{Word} \to Phon^+$$

$$\text{where}: \quad Word^+ \Leftrightarrow \begin{cases} Words \to Word \\ Words \to Word\, Words \end{cases}$$

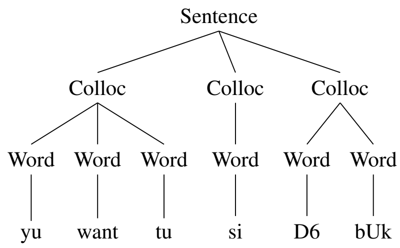
4.2 Collocations and Syllabification

The *baseline* that we are using is commonly called the "colloc-syll" model (Johnson, 2008; Börschinger et al., 2012) and is reported at 78% token F-score on the standard Brent version of the Bernstein-Ratner corpus corpus (Johnson, 2008). It posits that sentences are collocations of words, and words are composed of syllables. (Goldwater et al., 2009) showed how an assumption of independence between words (a unigram model) led to under-segmentation. So, above the $Word$ level, we take the collocations (co-occurring sequences) of words into account.

Furthermore, there is evidence that 8-month-old infants track syllable frequencies (Saffran et al., 1996), and the "colloc-syll" model can take that into account. $Word$ splits into general syllables and initial- or final- specific syllables. Syllables consist of onsets or codas (producing consonants), and nuclei (vowels). Onsets, nuclei and codas are adapted, thus allowing this model to memorize sequences or consonants or sequences of vowels, dependent on their position in the word. Consonants and vowels are the pre-terminals, their derivation is specified in the grammar into phonemes of the language.

$$Sentence \rightarrow Colloc^{+}$$
$$\underline{Colloc} \rightarrow Word^{+}$$
$$\underline{Word} \rightarrow StructSyll$$

For notations purposes, all this syllabification is appended after $Word$ by $\underline{Word} \rightarrow StructSyll$. All details about the collocations and syllabification grammars can be found in (Johnson, 2008). Here is an example of a (good) parse of "`yuwanttusiD6bUk`" with this model, skipping the $StructSyll$ derivations:



4.3 Including topics (contexts)

To allow for the model to make use of the topics (used as proxies for contexts), we modify the grammar by prefixing utterances with topic number (similarly to (Johnson et al., 2010)), $\forall K \in \#topics$:

$$Sentence \rightarrow \_tK\ Colloc_{tK}^{+}$$
$$\underline{Colloc_{tK}} \rightarrow Word_{tK}^{+}$$

For each $\underline{Word_{tK}}$, we can derive it into a common adapted $\underline{Word}$ by $\underline{Word_{tK}} \rightarrow Word$. Consider this lower level adaptor ($\underline{Word}$): it learns a shared vocabulary independently of the topic (all contexts that will derive `b U k` will increment the `Word(b U k)` pseudo-count). This $Word$-hierarchical model is called *share vocab*.

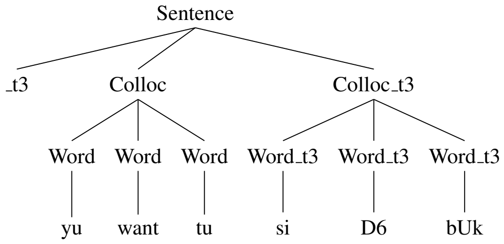
Alternatively, we can learn a separate vocabulary for each topic, by having directly: $\underline{Word_{tK}} \rightarrow StructSyll$ (note that all words then share the same syllabic structure). Words are split across different topics and need to be adapted for each topic in which they appear. This flat structure vocabulary model is called *split vocab*.

4.4 Allowing for non context-specific words

Sentences are not composed only of context-specific words, thus we need a third type of extension that allows for topic-independent and topic-specific words to mix. For this, we add topic-independent types of $Colloc$ and $Word$ that can be used across all topics, but we force each sentence to have at least one topical collocation:

$$\begin{aligned}
Sentence &\rightarrow _tK\ (Colloc^{+}|Colloc_{tK}^{+})\ Colloc_{tK}^{+} \\
&\ (Colloc^{+}|Colloc_{tK}^{+}) \\
\underline{Colloc_{tK}} &\rightarrow Word_{tK}^{+} \\
\underline{Colloc} &\rightarrow Word^{+} \\
\underline{Word_{tK}} &\rightarrow StructSyll \\
\underline{Word} &\rightarrow StructSyll
\end{aligned}$$

Parentheses denote that these terms are optionals, and "|" denotes "or". Both $\underline{Word_{tK}}$ and $\underline{Word}$ are adapted, but this time on the same level of hierarchy. This model allows the use of both topic-specific and common words in sentences, and it learns $\#topics + 1$ vocabularies. We call this model *with common*. An example of a correct parse with this model is given by:



5 Experimental setup

The *Providence* corpus (Demuth et al., 2006) consists of audio and video, weekly or bi-weekly, recordings of 6 monolingual English-speaking children home interactions. Each recording is approximatively 1 hour long. This corpus spans approximatively from their first to third year. We used the whole corpus to extract the topics to get more stable and general activity contexts. For all the following results, we used only the Naima portion between 11 months and 24 months, consisting in 26,425 utterances (sentences) and 135,389 tokens (words). The input consist in DARPABET-encoded sequences of phonemes with about 4200 word-types in the Naima subset. We followed the same preparation procedure as in (Börschinger et al., 2012), where more details about the corpus can be found.

We used the last version of Mark Johnson's Adaptor Grammars software[2]. All the additional code (preparation, topics, grammars, learning) to reproduce these experiments and results is freely available online[3], along with the datasets annotations derived from topic modeling[4]. For the adaptors, we used a $Beta(1, 1)$ (uniform) prior on the PYP a parameter, and a sparse $Gamma(100, 0.01)$ prior on the PYP b parameter. We ran 500 iterations (finishing at $\approx 0.05\%$ of log posterior variation between the lasts iterations) with several runs for each subset of the Naima dataset.

6 Results

6.1 Unsupervised words segmentation

Table 2: Mean (token and boundary) F-scores (f), precisions (p), and recalls (r) for different models depending on the size of dataset (age range).

months	baseline			share vocab			split vocab			with common		
token	f	p	r	f	p	r	f	p	r	f	p	r
11-12	**.80**	.79	.81	.77	.76	.78	.77	.75	.78	.77	.75	.78
11-15	.81	.81	.82	.76	.78	.75	.81	.79	.82	**.82**	.81	.83
11-19	.82	.82	.83	.77	.78	.76	.81	.81	.82	**.83**	.82	.84
11-22	.81	.82	.81	.77	.79	.75	.82	.81	.83	**.83**	.82	.84
boundary	f	p	r	f	p	r	f	p	r	f	p	r
11-12	**.90**	.88	.91	.88	.87	.89	.87	.85	.90	.88	.85	.90
11-15	**.91**	.91	.92	.89	.91	.86	**.91**	.89	.92	**.91**	.90	.93
11-19	**.92**	.92	.93	.90	.92	.88	**.92**	.91	.93	**.92**	.91	.94
11-22	.92	.93	.91	.90	.93	.87	.92	.91	.93	**.93**	.91	.94

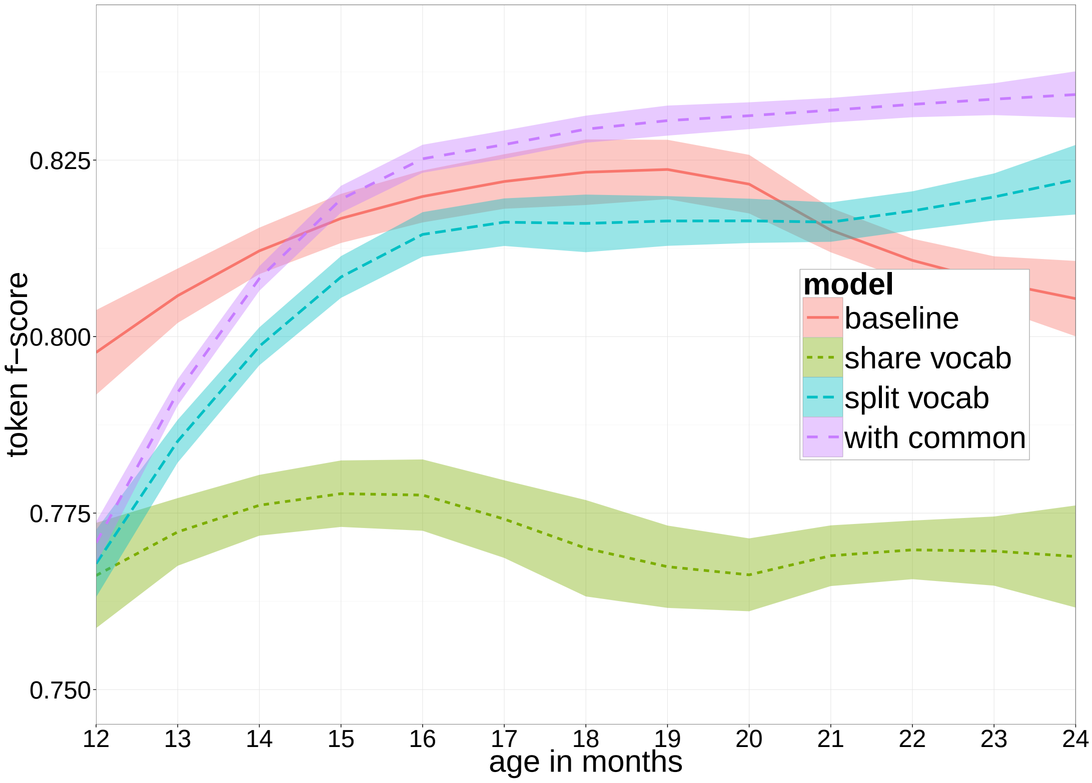
The key metric of interest is the token F-score (harmonic mean of precision and recall of words). Table 2 gives all the scores for an increasingly large dataset (as in (Börschinger et al., 2012)). Figure 1 shows the month-by-month evolution of the token F-score of the different models. We can see that

[2] `http://web.science.mq.edu.au/~mjohnson/`

[3] `https://github.com/SnippyHolloW/contextual_word_segmentation`

[4] `https://github.com/SnippyHolloW/contextual_word_segmentation/tree/master/ProvidenceFinal/Final`

Figure 1: Token F-scores (and standard deviations) evolution with an increasingly bigger and richer dataset (11 months to "X-axis value" months), computed on 8 runs of 500 iterations per data point.



context-based models need more data to get good performances (several vocabularies to learn), but they seem more resilient to over-segmentation.

Preliminary results confirm the trend of *baseline* scores getting slowly worse at **25** and **26** months while *with common* and *split vocab* stabilize (not plotted here). We also tried models for which we can have the "common vocabulary" derived only at the level of the collocations (making topic-specific collocations topic-pure as in *split vocab* for instance), or only at the level of the words (allowing for topic-specific collocations deriving in only common words if needed). Both models are worse than *split vocab* and *with common*.

Using a shared global vocabulary while being able to learn (through adaptation) different topic-specific vocabularies does not seem to be a solution: *share vocab* performs worse than the *baseline*. Token recall and boundary recall are worse off (see Table 2), suggesting that fewer words are correctly adapted. Maybe that is because this is the only model with two levels of adapted word hierarchies ($\underline{Word_{tK}}$ and $\underline{Word}$). Sharing a lower-level vocabulary ($\underline{Word}$) still does not allow for context vocabularies ($\underline{Word_{tK}}$) to mix, thus is simply harder to train. Having only one vocabulary per context (*split vocab*) is a slight improvement over the *baseline*, even though it is not significant (95% confidence interval) before 22 months. Models allowing for both topic-specific vocabularies and a common vocabulary to be learned are the best: *with common* is significantly (95% confidence interval) better than the *baseline*, starting from 20 months (Figure 1). The improvement seems to be due to better token (and boundary) recall (Table 2), suggesting that more words are learned. By looking at their lexicons at 24 months, topic-dependent models have slightly larger lexicon recalls and worse lexicon precisions than the *baseline*. This means that the additional true word-types that they learn are more frequently correctly used than the false word-types (otherwise the token F-scores would be reversed, e.g. between *split vocab* and baseline).

Figure 2: Mean token F-scores (and standard deviations) on 20% held-out test data for 6 different random splits of Naima from 11 to 22 months, 500 iterations each. Grey for *baseline* on *test*, green and blue for context-dependent models on *test* and *no prefix* conditions respectively.

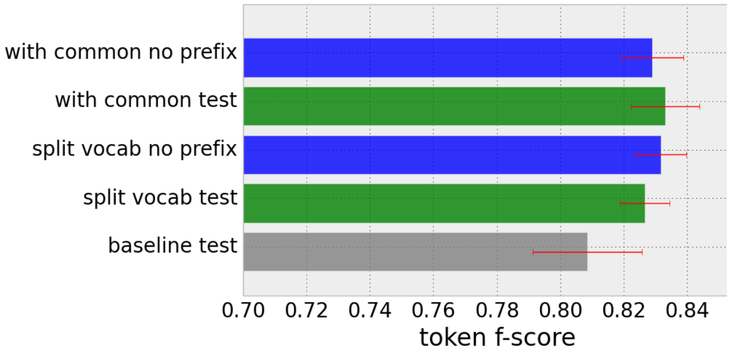


Table 3: Most probable words ($\propto P(word|topic = k)$) in the 7 recovered topics at test time without topic annotations (*no prefix* condition) for the *with common* model (we omitted phonemes clusters yielding non-words).

bread	elephant	lego	Michael	skinny	stick	bubble
delicious	owl	doctor	shorts	massage	remember	pasta
avocado	wearing	brush	towel	ostrich	track	spirals
porridge	turkey	change	shirts	nurse	forget	squirrels
raisin	haircut	squeeze	pirates	hammer	oink	thumb
biscuit	turtle	music	tangled	ruby	towed	pentagon
food	animals	play	clothes	(messy)	verbs	≈shapes

6.2 Recovery of the topics on held-out data

To check whether these models generalize to unseen utterances, and possibly unseen vocabulary, we looked at the scores of held-out data (80/20% train/test split of the Naima 11 to 22 months dataset). Token F-scores for this *test* condition are shown in green and grey in Figure 2. This separates low-frequency collocations to be used at test time and those seen at training time, both for context aware models and the basic *baseline* model. The F-scores show the same pattern as in the previous experiment, with context-aware models (*with common* and *split vocab* here) performing better than the *baseline*.

The topics are learned on the orthographic transcription of the whole *Providence* corpus (6 children), while we test only on the Naima dataset. Still, to check that these results are not simply due to additional information (leaked somehow in the form of the $\_tK$ prefix), we produced another held-out condition, without topic ($\_tK$) prefixes. Models can use topic-specific vocabularies learned during training, but they are given no context information at test time. Token F-scores for this *no prefix* condition are shown in blue (and grey for the baseline) in Figure 2. The fact that *no prefix* performance is on par with the *test* condition means that contextual cues are not only important at test time, but particularly so while learning the vocabulary. In other words, the model acquires its vocabularies making use of the additional context. In the *test* setting, it is evaluated on novel utterances for which additional context information is available. In the *no prefix* condition it is evaluated on novel utterances for which no additional context information is available. This means that topic-specific vocabulary learned during training is successfully used in a consistent way at test time. To confirm this qualitatively, we looked at the most probable words (after unsupervised segmentation from the phonemic input) in recovered topics at test time in the *no prefix* condition. They are shown in Table 3, and they exhibit some of the topics that were found on the orthographic transcript (as they are not limited to nouns, a topic for “verbs” appears).

7 Conclusion

We have shown that contextual information helps segmenting speech into word-like units. We used topic modeling as a proxy for richer contextual annotations, as (Roy et al., 2012) have shown high correlation between contexts and automatically derived topics. We modified existing Adaptor Grammar segmentation models (Johnson, 2008; Johnson and Goldwater, 2009), to be able to learn topic-specific vocabularies. We applied this approach to a large child directed speech corpus that was previously used for segmentation (Börschinger et al., 2012). Our model with the capacity to use both a topic-specific vocabulary and a common vocabulary (*with common*) produces better segmentation scores, ending up with at least 2.5% better absolute F-scores than its context-oblivious counterpart (*baseline*). More generally, both models that learn specialized vocabularies do not get worse F-scores with increasing data (Figure 1). Particularly, they seem to fix a well-known problem of previous models like "colloc-syll" (our *baseline*), that "overlearn" by over-segmenting frequent morphemes as single words (Börschinger et al., 2012). We have controlled for the additional information of giving the topic ($\_tK$), and we have found out that contextual information helps at training time.

It would be interesting to look into the link between semantics and syntax in recovered topics. Further work should integrate syntax (e.g. function words), stress cues and prosody from the audio signal (Börschinger and Johnson, 2014), use even less supervision for contexts, and be applied to other languages. We believe that language acquisition is not a simple sequential process and that segmentation, syntax, and word meaning bootstrap each others. This is only a first step towards integrating multiple sources of information and different modalities at all steps of language acquisition.

Acknowledgments

This project is funded in part by the European Research Council (ERC-2011-AdG-295810 BOOTPHON), the Agence Nationale pour la Recherche (ANR-10-LABX-0087 IEC, ANR-10-IDEX-0001-02 PSL*), the Fondation de France, the Ecole de Neurosciences de Paris, and the Region Ile de France (DIM cerveau et pense).

References

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, March.

Benjamin Börschinger and Mark Johnson. 2014. Exploring the role of stress in Bayesian word segmentation using Adaptor Grammars. *Transactions of the Association of Computational Linguistics*, 2:93–104, February.

Benjamin Börschinger, Katherine Demuth, and Mark Johnson. 2012. Studying the effect of input size for bayesian word segmentation on the providence corpus. In *COLING*, pages 325–340.

Katherine Demuth, Jennifer Culbertson, and Jennifer Alter. 2006. Word-minimality, epenthesis and coda licensing in the early acquisition of english. *Language and Speech*, 49(2):137–173.

Michael C Frank, Noah D Goodman, and Joshua B Tenenbaum. 2009. Using speakers' referential intentions to model early cross-situational word learning. *Psychological Science*, 20(5):578–585.

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(Jul):2335–2382.

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. Association for Computational Linguistics.

Mark Johnson, Thomas L Griffiths, and Sharon Goldwater. 2007. Adaptor grammars: A framework for specifying compositional nonparametric bayesian models. *Advances in neural information processing systems*, 19:641.

Mark Johnson, 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. 2008. Using adaptor grammars to identify synergies in the unsupervised acquisition of linguistic structure. In *ACL*, pages 398–406.

Peter W. Jusczyk and Richard N. Aslin. 1995. Infants detection of the sound patterns of words in fluent speech. *Cognitive psychology*, 29(1):123.

Peter W. Jusczyk, Anne Cutler, and Nancy J. Redanz. 1993a. Infants' preference for the predominant stress patterns of english words. *Child development*, 64(3):675687.

Peter W. Jusczyk, Angela D. Friederici, Jeanine MI Wessels, Vigdis Y. Svenkerud, and Ann Marie Jusczyk. 1993b. Infants sensitivity to the sound patterns of native language words. *Journal of Memory and Language*, 32(3):402420.

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

Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta, May. ELRA. `http://is.muni.cz/publication/884893/en`.

Deb Roy, Rupal Patel, Philip DeCamp, Rony Kubat, Michael Fleischman, Brandon Roy, Nikolaos Mavridis, Stefanie Tellex, Alexia Salata, Jethran Guinness, et al. 2006. The human speechome project. In *Symbol Grounding and Beyond*, pages 192–196. Springer.

Brandon C Roy, Michael C Frank, and Deb Roy. 2012. Relating activity contexts to early word learning in dense longitudinal data. In *Proceedings of the 34th Annual Cognitive Science Conference*.

Jenny R. Saffran, Richard N. Aslin, and Elissa L. Newport. 1996. Statistical learning by 8-month old infants. *Science*, 274(5294):1926–1928.

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