Unsupervised syntactic chunking with acoustic cues: computational models for prosodic bootstrapping

John K Pate (j.k.pate@sms.ed.ac.uk)
Sharon Goldwater (sgwater@inf.ed.ac.uk)
School of Informatics, University of Edinburgh
10 Crichton St., Edinburgh EH8 9AB, UK

Abstract

Learning to group words into phrases without supervision is a hard task for NLP systems, but infants routinely accomplish it. We hypothesize that infants use acoustic cues to prosody, which NLP systems typically ignore. To evaluate the utility of prosodic information for phrase discovery, we present an HMM-based unsupervised chunker that learns from only transcribed words and raw acoustic correlates to prosody. Unlike previous work on unsupervised parsing and chunking, we use neither gold standard part-of-speech tags nor punctuation in the input. Evaluated on the Switchboard corpus, our model outperforms several baselines that exploit either lexical or prosodic information alone, and, despite producing a flat structure, performs competitively with a state-of-the-art unsupervised lexicalized parser, with a substantial advantage in precision. Our results support the hypothesis that acoustic-prosodic cues provide useful evidence about syntactic phrases for language-learning infants.

1 Introduction

Young children routinely learn to group words into phrases, yet computational methods have so far struggled to accomplish this task without supervision. Previous work on unsupervised grammar induction has made progress by exploiting information such as gold-standard part of speech tags (e.g. Klein and Manning (2004)) or punctuation (e.g. Seginer (2007)). While this information may be available in some NLP contexts, our focus here is on the computational problem facing language-learning infants, who do not have access to either part of speech tags or punctuation. However, infants do have access to certain cues that have not been well explored by NLP researchers focused on grammar induction from text. In particular, we consider the cues to syntactic structure that might be available from prosody (roughly, the structure of speech conveyed through rhythm and intonation) and its acoustic realization.

The idea that prosody provides important initial cues for grammar acquisition is known as the *prosodic bootstrapping hypothesis*, and is well-established in the field of language acquisition (Gleitman and Wanner, 1982). Experimental work has provided strong support for this hypothesis, for example by showing that infants begin learning basic rhythmic properties of their language prenatally (Mehler et al., 1988) and that 9-month-olds use prosodic cues to distinguish verb phrases from non-constituents (Soderstrom et al., 2003). However, as far as we know, there has so far been no direct *computational* evaluation of the prosodic bootstrapping hypothesis. In this paper, we provide the first such evaluation by exploring the utility of acoustic cues for unsupervised syntactic chunking, i.e., grouping words into non-hierarchical syntactic phrases.

Nearly all previous work on unsupervised grammar induction has focused on learning hierarchical phrase structure (Lari and Young, 1990; Liang et al., 2007) or dependency structure (Klein and Manning, 2004); we are aware of only one previous paper on unsupervised syntactic chunking (Ponvert et al., 2010). Ponvert et al. describe a simple method for chunking that uses only bigram counts and punctuation; when the chunks are combined using a right-branching structure, the resulting trees achieve unlabeled bracketing precision and recall that is competitive with other unsupervised parsers. The sys-

Proceedings of the 2nd Workshop on Cognitive Modeling and Computational Linguistics, pages 20–29,
Portland, Oregon, June 2011.

tem's dependence on punctuation renders it inappropriate for addressing the questions we are interested in here, but its good performance reccommends syntactic chunking as a profitable approach to the problem of grammar induction, especially since chunks can be learned using much simpler models than are needed for hierarchical structure.

The models used in this paper are all variants of HMMs. Our baseline models are standard HMMs that learn from either lexical or prosodic observations only; we also consider three types of models (including a coupled HMM) that incorporate both lexical and prosodic observations, but vary the degree to which syntactic and prosodic variables are tied together in the latent structure of the models. In addition, we compare the use of hand-annotated prosodic information (ToBI annotations) to the use of direct acoustic measures (specifically, duration measures) as the prosodic observations. All of our models are unsupervised, receiving no bracketing information during training.

The results of our experiments strongly support the prosodic bootstrapping hypothesis: we find that using either ToBI annotations or acoustic measures in addition to lexical observations (i.e., word sequences) vastly improves chunking performance over any source of information alone. Interestingly, our best results are achieved using a combination of words and acoustic information as input, rather than words and ToBI annotations. Our best combined model achieves an F-score of 41% when evaluated on the lowest level of syntactic structure in the Switchboard corpus[1], as compared to 25% for a words-only model and only 3% for an acoustics-only model. Although the combined model's score is still fairly low, additional results suggest that our corpus of transcribed naturalistic speech is significantly more difficult for unsupervised parsing than the written text that is typically used for training. Specifically, we find that a state-of the-art unsupervised lexicalized parser, the Common Cover Link (CCL) parser (Seginer, 2007), achieves only 38% unlabeled bracketing F-score on our corpus, as compared to published results of 76% on WSJ10 (English) and 59% on Negra10 (German). Interestingly, we find that when evaluated against full parse trees, our best chunker achieves an F-score comparable to that of CCL despite positing only flat structure.

Before describing our models and experiments in more detail, we first present a brief review of relevant information about prosody and its relationship to syntax, including previous work combining prosody and syntax in supervised parsing systems.

2 Prosody and syntax

Prosody is a theoretical linguistic concept positing an abstract organizational structure for speech.[2] While it is often closely associated with such measurable phenomena as movement in fundamental frequency or variation in spectral tilt, these are merely observable acoustic correlates that provide evidence of varying quality about the hidden prosodic structure, which specifies such hidden variables as contrastive stress or question intonation.

Prosody has been hypothesized to be useful for learning syntax because it imposes a grouping structure on word sequences that sometimes coincides with traditional constituency analyses (Ladd, 1996; Shattuck-Hufnagel and Turk, 1996). Moreover, laboratory experiments have shown that adults use prosody both for syntactic disambiguation (Millotte et al., 2007; Price et al., 1991) and, crucially, in learning the syntax of an artificial language (Morgan et al., 1987). Accordingly, if prosodic structure is sufficiently prominent in the acoustic signal, and coincides often enough with syntactic structure, then it may provide children with useful information about how to combine words into phrases.

Although there are several theories of how to represent and annotate prosodic structure, one of the most influential is the ToBI (Tones and Break Indices) theory (Beckman et al., 2005), which we will use in some of our experiments. ToBI proposes, among other things, that the prosodic phrasing of languages can be represented in terms of sequences of break indices indicating the strength of

[1] Since our interest is in child language acquisition, we would prefer to evaluate our system on data from the CHILDES database of child-directed speech (MacWhinney, 2000). Unfortunately, there are no corpora in the database that include phrase structure annotations. We are in the process of annotating a small evaluation corpus with phrase structure trees, and hope to use this for evaluation in future work.

[2] Signed languages also exhibit prosodic phenomena, but they are not addressed here.

word boundaries. In Mainstream American English ToBI, for example, the boundary between a clitic and its base word (e.g. "do" and "n't" of "don't") is 0, representing a very weak boundary, while the boundary following a word at the end of an intonational phrase is 4, indicating a very strong boundary. Below we examine how useful these break indices are for identifying syntactic boundaries.

Finally, we note that our work is not the first computational approach to using prosody for identifying syntactic structure. However, previous work (Gregory et al., 2004; Kahn et al., 2005; Dreyer and Shafran, 2007; Nöth et al., 2000) has focused on supervised parsing rather than unsupervised chunking, and also makes different assumptions about prosody. For example, Gregory et al. (2004) assume that prosody is an acoustically-realized substitute for punctuation; our own treatment is much less constrained. Kahn et al. (2005) and Dreyer and Shafran (2007) use ToBI labels to represent prosodic information, whereas we explore both ToBI and direct acoustic measures. Finally, Nöth et al. (2000) do not use ToBI, instead developing a novel prosodic annotation system designed specifically to provide cues to syntax and for annotation efficiency. However, their system is supervised and focuses on improving parse *speed* rather than accuracy.

3 Models

Following previous work (e.g. Molina and Pla (2002) Sha and Pereira (2003)), we formulate chunking as a tagging task. We use Hidden Markov Models (HMMs) and their variants to perform the tagging, with carefully specified tags and constrained transition distributions to allow us to interpret the results as a bracketing of the input. Specifically, we use four chunk tags: **B** ("Begin") and **E** ("End") tags are interpreted as the first and last words of a chunk, respectively, with **I** ("Inside") corresponding to other words inside a chunk and **O** ("Outside") to all other words. The transition matrices are constrained to afford 0 probability to transitions that violate these definitions. Additionally, the initial probabilities are constrained to forbid the models from starting inside or at the end of a phrase.

We use this four-tag **OBIE** tagset rather than the more typical three-tag **IOB** tagset for two reasons. First, the **OBIE** set forces all chunks to be at least two words long (the shortest chunk allowed is **B E**). Imposing this requirement allows us to characterize the task in concrete terms as "learning when to group words together." Second, as we seek to incorporate acoustic correlates of prosody into chunking, we expect edge behavior to merit explicit modeling.[3]

In the following subsections, we describe the various models we use. Note that input to all models is discrete, consisting of words, ToBI annotations, and/or discretized acoustic measures (we describe these measures and their discretization in Section 3.3). See Figure 1 for examples of system input and output; different models will receive different combinations of the three kinds of input.

3.1 Baseline Models

Our baseline models are all standard HMMs, with the graphical structure shown in Figure 2(a). The first baseline uses *lexical* information only; the observation at each time step is the phonetic transcription of the current word in the sentence. To handle unseen words at test time, we use an "UNK." token to replace all words in the training and evaluation sets that appear less than twice in the training data. Our second baseline uses *prosodic* information only; the observation at each time step is the hand-annotated ToBI Break Index for the current word, which takes on one of six values: { 0, 1, 2, 3, 4, X, None }.[4] Our final baseline uses *acoustic* information only. The observations are one of six automatically determined clusters in an acoustic space, as described in Section 3.3.

We trained the HMMs using Baum-Welch, and used Viterbi for inference.[5]

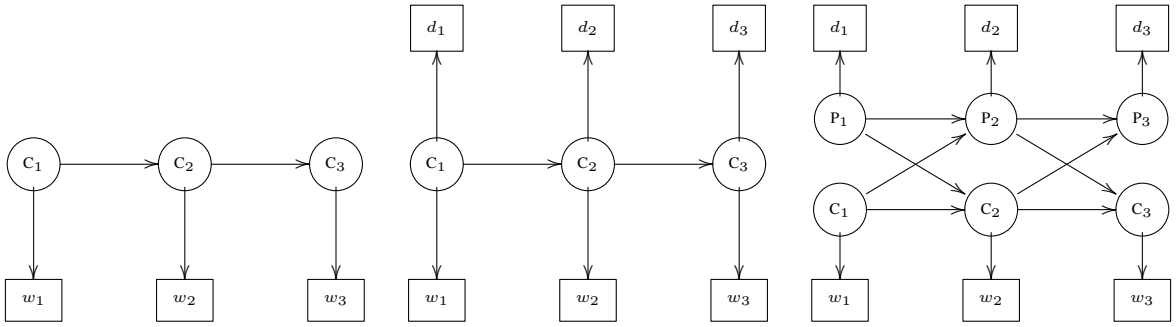
[3] Indeed, when we tried using the **IOB** tag set in preliminary experiments, dev-set performance dropped substantially, supporting this latter intuition.

[4] The numerical break indices indicate breaks of increasing strength, "X" represents a break of uncertain strength, and "None" indicates that the preceding word is outside one of the fluent prosodic phrases selected for annotation. Additional distinctions marked by "-" and "p" were ignored.

[5] We actually used the junction tree algorithm from MALLET, which, in the special case of an HMM, reduces to the Forward-Backward algorithm when using Sum-Product messages, and to the Viterbi algorithm when using Max-Product messages. Our extension of MALLET to build junction trees efficiently for Dynamic Bayes Nets is available online, and is being prepared for submission to the main MALLET project.

(a)	Words	g.aa	dh.ae.t.s	dh.ae.t	s.aw.n.d.z	p.r.ih.t.iy	b.ae.d	t.ax	m.iy
	Acoustics	4	4	6	4	5	4	5	6
	ToBI	1	2	1	1	1	1	1	3
(b)		O	O	B	I	I	E	B	E
(c)				(			)	(	)
(d)	(			(			)	(	))

Figure 1: (a) Example input sequences for the three types of input (phonetic word transcriptions, acoustic clusters, and ToBI break indices). (b) Example output tags. (c) The bracketing corresponding to (b). (d) The flat tree built from (b).



(a) Standard HMM (HMM) (b) Two Output HMM (THMM) (c) Coupled HMM (CHMM)

Figure 2: Graphical structures for our various HMMs. c_i nodes are constrained using the **OBIE** system, p_i nodes are not. w_i nodes represent lexical outputs, and d_i nodes represent acoustic or ToBI outputs. (Rectangular nodes are observed, circular nodes are hidden).

3.2 Combined Models

As discussed in Section 2, previous theoretical and experimental work suggests a combined model which models uncertainty both between prosody and acoustics, and between prosody and syntax. To measure the importance of modeling these kinds of uncertainty, we will evaluate a series of model structures that gradually divorce acoustic-prosodic cues from lexical-syntactic cues.

Our first model is the standard HMM from Figure 2(a), but generates a (word, acoustics) or (word, ToBI) pair at each time step. This model has the simplest structure, but includes a separate parameter for every unique (state, word, acoustics) triple, so may be too unconstrained to learn anything useful.

To reduce the number of parameters, we propose a second model that assumes independence between the acoustic and lexical observations, given the syntactic state. We call this a "Two-output HMM (THMM)" and present its graphical structure in Figure 2(b). It is straightforward to extend Baum-Welch to accommodate the extra outputs of the THMM.

Finally, we consider a model that explicitly represents prosodic structure distinctly from syntactic structure with a second sequence of tags. We use a Coupled HMM (CHMM) (Nefian et al., 2002), which models a set of observation sequences using a set of hidden variable sequences. Figure 2(c) presents a two-stream Coupled HMM for three time steps. The model consists of an initial state probability distribution π_s for each stream s, a transition matrix a_s for each stream s conditioning the distribution of stream s at time $t+1$ on the state of both streams at time t, and an emission matrix b_s for each stream conditioning the observation of stream s at time t on the hidden state of stream s at time t.[6]

Intuitively, the states emitting acoustic measures operationalize prosodic structure, and the states emitting words operationalize syntactic structure. Crucially, Coupled HMMs impose no *a priori* correspondence between variables of different streams, allowing our "syntactic" states to vary freely from our "prosodic" states. As two-stream CHMMs maintain two emission matrices, two transition ma-

[6] We explored a number of minor variations on this graphical structure, but preliminary experiments yielded no improvement.

trices, and two initial state distributions, they are more complex than the other combined models, but more closely embody intuitions inspired by previous work on the prosody-syntax interface.

Our Coupled HMMs were also trained using EM. Marginals for the E-step were computed using the implementation of the junction tree algorithm available in MALLET (McCallum, 2002; Sutton, 2006). During test, the Viterbi tag sequence for each model is obtained by simply replacing the sum-product messages with max-product messages.

3.3 Acoustic Cues

As explained in Section 2, prosody is an abstract hidden structure which only correlates with observable features of the acoustic signal, and we seek to select features which are both easy to measure and likely to correlate strongly with the hidden prosodic phrasal structure. While there are many possible cues, we have chosen to use duration cues. These should provide good evidence about phrases due to the phenomenon of pre-boundary lengthening (e.g. Beckman and Edwards (1990), Wightman et al. (1992)), wherein words, and their final rime, lengthen phrase-finally. This is likely especially useful for English due to the lack of confounding segmental duration contrasts (although variation in duration is unpredictably distributed (Klatt, 1976)), but should be useful in varying degrees for other languages.

We gather five duration measures:

1. Log total word duration: The annotated word end time minus the annotated word start time.
2. Log onset duration: The duration from the beginning of the word to the end of the first vowel.
3. Log offset duration: The duration from the beginning of the last vowel to the end of the word.
4. Onset proportion consonant: The duration of the non-vocalic portion of the word onset divided by the total onset duration.
5. Offset proportion consonant: The duration of the non-vocalic portion of the word offset divided by the total offset duration.

If a word contains no canonical vowels, then the first and last sonorants are counted as vocalic. If a word contains no vowels or sonorants, then the onset and offset are the entire word and the proportion consonant for both onset and offset is 1 (this occurred for 186 words in our corpus).

The potential utility of this acoustic space was verified by visual inspection of the first few PCA components, which suggested that the position of a word in this acoustic space correlated with bracket count. We discretize the raw (i.e. non-PCA) space with k-means with six initially random centers for consistency with the number of ToBI break indices.

	Train	Dev	Test
Words	68,533	7,981	8,746
Sentences	6,420	778	802

Table 1: Data set statistics

4 Experiments

4.1 Dataset

All experiments were performed on part of the Nite XML Toolkit edition of the Switchboard corpus (Calhoun et al., 2010). Specifically, we gathered all conversations which have been annotated for syntax, ToBI, and Mississippi State phonetic alignments (which lack punctuation).[7] The syntactic parses, word sequences, and ToBI break indices were hand-annotated by trained linguists, while the Mississippi State phonetic alignments were automatically produced by a forced alignment of the speech signal to a pronunciation-dictionary based phone sequence, providing an estimate of the beginning and end time of each phone. A small number of annotation errors (in which the beginning and end times of some phones had been swapped) were corrected by hand. This corpus has 74 conversations with two sides each.

We split this corpus into an 80%/10%/10% train/dev/test [8] partition by dividing the entire corpus into ten-sentence chunks, assigning the first eight to the training partition, and the ninth and tenth to the dev and test partitions, respectively. We then removed all sentences containing only one or two

[7] We threw out a small number of sentences with annotations errors, e.g. pointing to missing words.

[8] The dev set was used to explore different model structures in preliminary experiments; all reported results are on the test set.

words. Sentences this short have a trivial parse, and are usually formulaic discourse responses (Bell et al., 2009), which may influence their prosody. The final corpus statistics are presented in Table 1.

4.2 Evaluation

We use the Penn Treebank parsed version of Switchboard for evaluation. This version uses a slightly different tokenization from the Mississippi State transcriptions that were used as input to the models, so we transformed the Penn treebank tokenization to agree with the Mississippi State tokenization (primarily by concatenating clitics to their base words—i.e. "do" and "'nt" into "don't"—and splitting multi-word expressions). We also removed all gold-standard nodes spanning only `Trace` or `PUNC` (recall that the input to the models did not include punctuation) and collapsed all unary productions.[9]

In all evaluations, we convert our models' output tag sequence to a set of matched brackets by inserting a left bracket preceding each word tagged **B** tag and a right bracket following each word tagged **E**. This procedure occasionally results in a sentence with an unmatched opening bracket. If the unmatched opening bracket is one word from the end of the sentence, we delete it, otherwise we insert a closing bracket at the end of the sentence. Figure 1 shows example input sequences together with example output tags and their corresponding bracketings.

Previous work on chunking, most notably the 2000 CONLL shared task (Tjong et al., 2000), has defined gold standard chunks that are useful for finding grammatical relations but which do not correspond to any particular linguistic notion. It is not clear that such chunks should play a role in language acquisition, so instead we evaluate against traditional syntactic constituents from Penn Treebank-style parses in two different ways.

Our first evaluation method compares the output of the chunkers to what Ponvert et al. (2010) call *clumps*, which are just syntactic constituents that span only terminals. We created our clump gold-standard by taking the parse trees resulting from the preprocessing described above and deleting nodes that span a non-terminal. Figure 3 presents an example gold-standard parse tree with the clumps in boxes. This evaluation avoids penalizing chunkers for not positing hierarchical structure, but rewards chunkers only for finding very low-level structure.

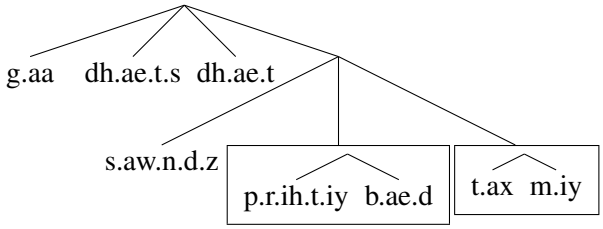


Figure 3: Example gold-standard with clumps in boxes.

In the interest of making no *a priori* assumptions about the kinds of phrases our unsupervised method recovers, we also evaluate our completely flat, non-recursive chunks directly against the fully recursive parses in the treebank. To do so, we turn our chunked utterance into a flat tree by simply putting brackets around the entire utterance as in Figure 1(d). This evaluation penalizes chunkers for never positing hierarchical structure, but makes no assumptions about which kinds of phrases ought to be found.

4.3 Models and training

In all, nine HMM models, two versions of the CCL parser, and a uniform right-branching baseline were evaluated. Three of the HMMs were standard HMMs with chunking constraints on the four hidden states (as described in Section 3.2) that received as input either words, ToBI break indices, or word duration cluster information, intended as baselines to illuminate the utility of each information source in isolation. We also ran two each of Coupled HMM and Two-output HMM models that received words in one observed chain and either ToBI break index or duration cluster in the other observed chain. In the CHMM models, chunking constraints were enforced on the chain generating the words, while variables generating the duration or ToBI information ranged over four discrete states with no constraints.[10] All non-zero parameters were initialized approximately uniformly at random,[11] and we ran EM until the log

[9] As we evaluate unlabeled bracketing precision and recall, the label of the resulting nodes is irrelevant.

[10] We also tried imposing chunking constraints on the second chain, but dev-set performance dropped slightly.

[11] In preliminary dev-set experiments, different random initializations performed within two points of each other.

Condition			Prec	Rec	F-sc
Baselines	HMM	Wds	23.5	39.9	26.3
		BI	7.2	4.8	5.8
		Ac	4.7	2.5	3.3
Combined Models	HMM	Wds+BI	24.4	22.2	23.2
		Wds+Ac	20.7	22.7	21.7
	THMM	Wds+BI	18.2	19.6	18.9
		Wds+Ac	36.1	47.8	**41.2**
	CHMM	Wds+BI	25.5	36.3	29.9
		Wds+Ac	33.6	**48.1**	39.5
CCL		Parser	15.4	41.5	22.4
		Clumper	**36.8**	37.9	37.3

Table 2: Scores for all models, evaluated on clumps. Input is words (Wds), break indices (BI), and/or acoustics.

Condition			% Covered Words	% Covered Utts	$\frac{words}{chunk}$	$\frac{chunk}{utt}$
Baselines	HMM	Wds	81.9	98.4	3.16	2.82
		BI	68.2	68.1	4.95	1.50
		Ac	46.3	71.1	4.18	1.21
Combined Models	HMM	Wds+BI	79.8	98.3	4.30	2.02
		Wds+Ac	83.3	98.5	3.71	2.45
	THMM	Wds+BI	84.6	99.0	3.84	2.40
		Wds+Ac	68.0	96.1	2.52	2.94
	CHMM	Wds+BI	83.1	99.0	2.86	3.17
		Wds+Ac	76.5	97.6	2.62	3.19
CCL Clumper			48.3	99.9	2.30	2.29

Table 3: % words in a chunk, % utterances with > 0 chunks, and mean chunk length and chunks per utterance.

corpus probability changed less than 0.001%, typically for 50-150 iterations.

The CCL parser was trained on the same word sequences provided to our models. We also evaluated the CCL parser as a clumper (CCL Clumper) by removing internal nodes spanning a non-terminal. The right-branching baseline was generated by inserting one opening bracket in front of all but the last word, and closing all brackets at the end of the sentence.

4.4 Results and Discussion

Table 2 presents results for our flat chunkers evaluated against Ponvert et al. (2010)-style clumps. Several points are apparent. First, all three HMM baselines yield very poor results, especially the prosodic baselines, whose precision and recall are both below 10%. Although the best combined models still have relatively low performance, it is markedly higher than either of the individual baselines, and also higher than the clumps identified by the CCL parser. Particularly notable is the fact that lexical and prosodic information appear to be superadditive in some cases, yielding combined performance that is higher than the sum of the individual scores. Not all combined models work equally well, however: the poor performance of the HMM combined model supports our initial hypothesis that it is over parameterized. Interestingly, our acoustic clusters work better than break indices when combined with words. Finally, we see that the THMM and CHMM obtain similar performance using words + acoustics, suggesting that modeling prosodic structure separately from syntactic structure may be unnecessary (or that the CHMM does so badly).

Condition			Prec	Rec	F-sc
Baselines	HMM	Wds	48.8(32)	26.3(15)	34.2(20)
		BI	52.4(21)	18.5(5)	27.3(8)
		Ac	52.5(15)	16.3(3)	24.9(5)
Combined Models	HMM	Wds+BI	54.4(32)	23.2(11)	32.5(16)
		Wds+Ac	51.0(32)	24.7(13)	33.3(18)
	THMM	Wds+BI	55.9(38)	26.8(15)	36.2(21)
		Wds+Ac	55.8(41)	31.0(20)	39.9(27)
	CHMM	Wds+BI	48.4(32)	28.4(17)	35.8(22)
		Wds+Ac	54.1(40)	31.9(21)	**40.1**(**28**)
CCL		Parser	38.2(28)	**37.6**(**28**)	37.9(**28**)
		Clumper	**58.8**(**42**)	27.3(16)	37.3(23)

Table 4: Model performance, evaluated on full trees. Scores in parentheses were computed after removing the full sentence bracket, which provides a free true positive.

To provide further intuition into the kinds of chunks recovered by the different models, we list some relevant statistics in Table 3. These statistics show that the models using lexical information identify at least one chunk in virtually all utterances, with the better models averaging 2-3 chunks per utterance of around 3 words each. In contrast, the unlexicalized models find longer chunks (4-5 words each) but far fewer of them, with about 30% of utterances containing none at all.

We turn now to the models' performance on full parse trees, shown in Table 4. Two different scores are given for each system: the first includes the top-level bracketing of the full sentence (which is

standard in computing bracketing accuracy, but is a free true positive), while the second does not (for a more accurate picture of the system's performance on ambiguous brackets). Comparing the second set of scores to the clumping evaluation, recall is much lower for all the chunkers; the relatively small increase in precision indicates that the chunkers are most effective at finding low-level structure. For both sets of scores, the relative F-scores of the chunkers are similar to the clumping evaluation, with the words + acoustics versions of the THMM and CHMM scoring best. Not surprisingly, the CCL parser has much higher recall than the chunkers, though the best chunkers have much higher precision. The result is that, using standard Parseval scoring (first column), the best chunkers outperform CCL on F-score; even discounting the free sentence-level bracket (second column) they do about as well.

It is worth noting that, although CCL achieves state-of-the-art performance on the English WSJ and German Negra corpora (Seginer (2007) reports 75.9% F-score on WSJ10, for example), its performance on our corpus is far lower. In fact, on this corpus the CCL parser (as well as our chunkers) underperform a uniform right-branching baseline, which obtains 42.2% precision and 64.8% recall (including the top-level bracket), leading to an overall F-score of 51.1%. This suggests that our corpus is significantly more difficult than WSJ, probably due to disfluencies and/or lack of punctuation.[12] Moreover, we stress that the use of a right-branching baseline, while useful as a measure of overall performance, is not plausible as a model of language acquisition since it is highly language-specific.

5 Conclusion

Taken together, our results indicate that a purely local model that combines lexical and acoustic-prosodic information in an appropriate way can identify syntactic phrases far more effectively than a similar model using either source of information alone. Our best combined models outperformed the baseline individual models by a wide margin when evaluated against the lowest level of syntactic structure, and their performance was comparable to CCL, a state-of-the-art unsupervised lexicalized parser, when evaluated against full parse trees. It is disappointing that all of these systems scored worse than a right-branching baseline, but this result underscores the major differences between parsing spoken utterances (even using transcriptions) and parsing written text (where CCL and other unsupervised parsers were developed and tested). Since children learning language do not (at least initially) know the head direction of their language, the right-branching baseline for English is not available to them. Thus, combining lexical and acoustic cues may provide them with initial useful information about the location of syntactic phrases, as suggested by the prosodic bootstrapping hypothesis.

Nevertheless, we caution against assuming that the usefulness of acoustic information must result from its relation to prosody (especially because we found that direct acoustic information was more useful than hand-annotated prosodic labels). The "Smooth Signal Hypothesis" (Aylett and Turk, 2004) posits that talkers modulate their communicative effort according to the predictability of their message in order to achieve efficient communication, pronouncing more predictable parts of messages more quickly or less distinctly. If talkers consider syntactic predictability in this process, then it is possible that acoustic cues help initial grammar learning not by serving as cues to prosody but by serving as cues to the talker's syntax-dependent view of predictability. In this case, it may make more sense to discuss "predictability bootstrapping" rather than "prosodic bootstrapping."

Regardless of the underlying reason, we have shown that acoustic cues can be useful for identifying syntactic structure when used in combination with lexical information. In order to further substantiate these results, we plan to replicate our experiments on a corpus of child-directed speech, which we are currently annotating for evaluation purposes. We also hope to extend our findings to a model that can identify hierarchical structure, and to analyze more carefully the reasons for CCL's poor performance on the Switchboard corpus, in hopes of developing a model that can reach levels of performance closer to those typical of unsupervised parsers for written text.

[12] Including punctuation improves CCL little, possibly because the punctuation in this corpus is nearly all sentence-final.

References

Matthew Aylett and Alice Turk. 2004. The smooth signal redundancy hypothesis: A functional explanation for relationships between redundancy, prosodic prominence, and duration in spontaneous speech. *Language and Speech*, 47(1):31 – 56.

Mary E. Beckman and Jan Edwards. 1990. Lengthenings and shortenings and the nature of prosodic constituency. In J. Kingston and Mary E. Beckman, editors, *Between the grammar and physics of speech: Papers in laboratory phonology I*, pages 152–178. Cambridge: Cambridge University Press.

M Beckman, J Hirschberg, and S Shattuck-Hufnagel. 2005. The original tobi system and the evolution of the tobi framework. In S.-A. Jun, editor, *Prosodic Typology – The Phonology of Intonation and Phrasing*. Oxford University Press.

Alan Bell, Jason M. Brenier, Michelle Gregory, Cynthia Girand, and Dan Jurafsky. 2009. Predictability effects on durations of content and function words in conversational english. *Journal of Memory and Language*, 60:92 – 111.

S Calhoun, J Carletta, J Brenier, N Mayo, D Jurafsky, M Steedman, and D Beaver. 2010. The nxt-format switchboard corpus: A rich resource for investigating the syntax, semantics, pragmatics and prosody of dialogue. *Language Resources and Evaluation*, 44(4):387 – 419.

Markus Dreyer and Izhak Shafran. 2007. Exploiting prosody for pcfgs with latent annotations. In *Proc. of Interspeech*, Antwerp, Belgium, August.

L. Gleitman and E. Wanner. 1982. Language acquisition: The state of the art. In E. Wanner and L. Gleitman, editors, *Language acquisition: The state of the art*, pages 3–48. Cambridge University Press, Cambridge, UK.

Michelle L. Gregory, Mark Johnson, and Eugene Charniak. 2004. Sentence-internal prosody does not help parsing the way punctuation does. In *Proceedings of the North American Association for Computational Linguistics (NAACL)*, pages 81–88.

J. G. Kahn, M. Lease, E. Charniak, M. Johnson, and M. Ostendorf. 2005. Effective use of prosody in parsing conversational speech. In *Proc. of HLT/EMNLP-05*.

D H Klatt. 1976. Linguistic uses of segmental durations in english: Acoustic and perceptual evidence. *JASA*, 59:1208 – 1221.

Dan Klein and Christopher D. Manning. 2004. Corpus-based induction of syntactic structure: Models of dependency and constituency. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL 2004)*, pages 479–486.

Bob Ladd. 1996. *Intonational Phonology*. Cambridge University Press.

K Lari and S J Young. 1990. The estimation of stochastic context-free grammars using the inside-outside algorithm. *Computer Speech and Language*, 5:237 – 257.

P. Liang, S. Petrov, M. I. Jordan, and D. Klein. 2007. The infinite PCFG using hierarchical Dirichlet processes. In *Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP/CoNLL)*.

Brian MacWhinney. 2000. *The CHILDES project: Tools for analyzing talk*. Lawrence Erlbaum Associates, Mahwah, NJ, third edition.

Andrew Kachites McCallum. 2002. Mallet: A machine learning for language toolkit. http://mallet.cs.umass.edu.

Jacques Mehler, Peter Juszcyk, Ghislaine Lambertz, Nilofar Halsted, Josiane Bertoncini, and Claudine Amiele-Tison. 1988. A precursor to language acquisition in young infants. *Cognition*, 29:143 – 178.

Séverine Millotte, Roger Wales, and Anne Christophe. 2007. Phrasal prosody disambiguates syntax. *Language and Cognitive Processes*, 22(6):898 – 909.

Antonio Molina and Feran Pla. 2002. Shallow parsing using specialized HMMs. *Journal of Machine Learning Research*, 2:595 – 613.

James L. Morgan, Richard P. Meier, and Elissa L. Newport. 1987. Structural packaging in the input to language learning: contributions of prosodic and morphological marking of phrases to the acquisition of language. *Cognitive Psychology*, 19:498 – 550.

Ara V. Nefian, Luhong Liang, Xiaobao Pi, Liu Xiaoxiang, Crusoe Moe, and Kevin Murphy. 2002. A coupled hmm for audiovisual speech recognition. In *International Conference on Acoustics, Speech and Signal Processing*.

Elmer Nöth, Anton Batliner, Andreas Kieling, and Ralfe Kompe. 2000. Verbmobil: The use of prosody in the linguistic components of a speech understanding system. *IEEE Transactions on Speech and Audio Processing*, 8(5).

Elias Ponvert, Jason Baldridge, and Katrin Erk. 2010. Simple unsupervised identification of low-level constituents. In *ICSC*.

P J Price, M Ostendorf, S Shattuck-Hufnagel, and C Fong. 1991. The use of prosody in syntactic disambiguation. *JASA*, pages 2956 – 2970.

Yoav Seginer. 2007. Fast unsupervised incremental parsing. In *Proceedings of the Association of Computational Linguistics*.

Fei Sha and Fernando Pereira. 2003. Shallow parsing with conditional random fields. In *Proceedings of HLT-NAACL 03*, pages 213–220.

Stefanie Shattuck-Hufnagel and Alice E Turk. 1996. A prosody tutorial for investigators of auditory sentence processing. *Journal of Psycholinguistic Research*, 25(2):193 – 247.

M. Soderstrom, A. Seidl, D. G. K. Nelson, and P. W. Jusczyk. 2003. The prosodic bootstrapping of phrases: Evidence from prelinguistic infants. *Journal of Memory and Language*, 49:249–267.

Charles Sutton. 2006. Grmm: Graphical models in mallet. http://mallet.cs.umass.edu/grmm/.

Erik F. Tjong, Kim Sang, and Sabine Buchholz. 2000. Introduction to the conll-2000 shared task: Chunking. In *Proceedings of CoNLL-2000 and LLL-2000*, Lisbon, Portugal.

C W Wightman, S Shattuck-Hufnagel, M. Ostendorf, and P J Price. 1992. Segmental durations in the vicinity of prosodic phrase boundaries. *Journal of the Acoustical Society of America*, 91(3):1707 – 1717.