Memory-Bounded Left-Corner Unsupervised Grammar Induction on Child-Directed Input

Cory Shain
The Ohio State University
shain.3@osu.edu

William Bryce
University of Illinois
at Urbana-Champaign
bryce2@illinois.edu

Lifeng Jin
The Ohio State University
jin.544@osu.edu

Victoria Krakovna
Harvard University
vkrakovna@fas.harvard.edu

Finale Doshi-Velez
Harvard University
finale@saes.harvard.edu

Timothy Miller
Boston Children's Hospital &
Harvard Medical School
timothy.miller@childrens.harvard.edu

William Schuler
The Ohio State University
schuler@ling.osu.edu

Lane Schwartz
University of Illinois
at Urbana-Champaign
lanes@illinois.edu

Abstract

This paper presents a new memory-bounded left-corner parsing model for unsupervised raw-text syntax induction, using unsupervised hierarchical hidden Markov models (UHHMM). We deploy this algorithm to shed light on the extent to which human language learners can discover hierarchical syntax through distributional statistics alone, by modeling two widely-accepted features of human language acquisition and sentence processing that have not been simultaneously modeled by any existing grammar induction algorithm: (1) a left-corner parsing strategy and (2) limited working memory capacity. To model realistic input to human language learners, we evaluate our system on a corpus of child-directed speech rather than typical newswire corpora. Results beat or closely match those of three competing systems.

1 Introduction

The success of statistical grammar induction systems (Klein and Manning, 2002; Seginer, 2007; Ponvert et al., 2011; Christodoulopoulos et al., 2012) seems to suggest that sufficient statistical information is available in language to allow grammar acquisition on this basis alone, as has been argued for word segmentation (Saffran et al., 1999). But existing grammar induction systems make unrealistic assumptions about human learners, such as the availability of part-of-speech information and access to an index-addressable parser chart, which are not independently cognitively motivated. This paper explores the possibility that a memory-limited incremental left-corner parser, of the sort independently motivated in sentence processing theories (Gibson, 1991; Lewis and Vasishth, 2005), can still acquire grammar by exploiting statistical information in child-directed speech.

2 Related Work

This paper bridges work on human sentence processing and syntax acquisition on the one hand and unsupervised grammar induction (raw-text parsing) on the other. We discuss relevant literature from each of these areas in the remainder of this section.

Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, pages 964–975, Osaka, Japan, December 11-17 2016.

2.1 Human sentence processing and syntax acquisition

Related work in psycholinguistics and cognitive psychology has provided evidence that humans have a limited ability to store and retrieve structures from working memory (Miller, 1956; Cowan, 2001; McElree, 2001), and may therefore employ a left-corner-like strategy during incremental sentence processing (Johnson-Laird, 1983; Abney and Johnson, 1991; Gibson, 1991; Resnik, 1992; Stabler, 1994; Lewis and Vasishth, 2005). Schuler et al. (2010) show that nearly all naturally-occurring sentences can be parsed using no more than four disjoint derivation fragments in a left-corner parser, suggesting that general-purpose working memory resources are all that is needed to account for information storage and retrieval during online sentence processing. These findings motivate our left-corner parsing strategy and depth-bounded memory store.

An extensive literature indicates that memory abilities develop with age (see e.g. Gathercole, 1998 for a review). Newport (1990) proposed that limited processing abilities actually facilitate language acquisition by constraining the hypothesis space (the 'less-is-more' hypothesis). This theory has been supported by a number of subsequent computational and laboratory studies (e.g, Elman, 1993; Goldowski & Newport, 1993; Kareev et al., 1997) and parallels similar developments in the 'curriculum learning' training regimen for machine learning (Bengio et al., 2009).[1] Research on the acquisition of syntax has shown that infants are sensitive to syntactic structure (Newport et al., 1977; Seidl et al., 2003) and that memory limitations constrain the learning of syntactic dependencies (Santelman and Jusczyk, 1998). Together, these results suggest both (1) that the memory constraints in infants and young children are even more extreme than those attested for adults and (2) that these constraints impact – and may even facilitate – learning. By implementing these constraints in a domain-general computational model, we can explore the extent to which human learners might exploit distributional statistics during syntax acquisition (Lappin and Shieber, 2007).

2.2 Unsupervised grammar induction

The process of grammar induction learns the syntactic structure of a language from a sample of unlabeled text, rather than a gold-standard treebank. The constituent context model (CCM) (Klein and Manning, 2002) uses expectation-maximization (EM) to learn differences between observed and unobserved bracketings, and the dependency model with valence (DMV) (Klein and Manning, 2004) uses EM to learn distributions that generate child dependencies, conditioned on valence (left or right direction) in addition to the lexical head. Both of these algorithms induce on gold part-of-speech tag sequences.

A number of successful unsupervised raw-text syntax induction systems also exist. Seginer (2007) (CCL) uses a non-probabilistic scoring system and a dependency-like syntactic representation to bracket raw-text input. Ponvert et al. (2011) (UPPARSE) use a cascade of hidden Markov model (HMM) chunkers for unsupervised raw-text parsing. Christodoulopoulos et al. (2012) (BMMM+DMV) induce part-of-speech (PoS) tags from raw text using the Bayesian multinomial mixture model (BMMM) of Christodoulopoulos et al. (2011), induce dependencies from those tags using DMV, and iteratively re-tag and reparse using the induced dependencies as features in the tagging process. In contrast to ours, none of these systems employ a left-corner parsing strategy or model working memory limitations.

3 Methods

Experiments described in this paper use a memory-bounded probabilistic sequence model implementation of a left-corner parser (Aho and Ullman, 1972; van Schijndel et al., 2013) to determine whether natural language grammar can be acquired on the basis of statistics in transcribed speech within human-like memory constraints. The model assumes access to episodic memories of training sentences, but imposes constraints on working memory usage during sentence processing. The core innovation of this paper is the adaptation of this processing model to Bayesian unsupervised induction using constrained priors.

[1] The 'less-is-more' hypothesis has been a subject of controversy, however. See e.g. Rohde and Plaut (2003) for a critical review.

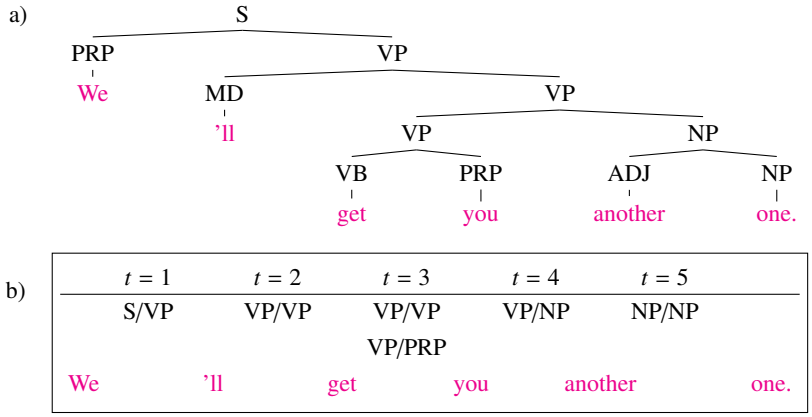


Figure 1: Trees and partial analyses for the sentence '*We'll get you another one*', taken from the training corpus. Derivation fragments are shown vertically stacked between words, using '/' to delimit top and bottom signs.

3.1 Left-corner parsing

Left-corner parsing is attractive as a sentence processing model because it maintains a very small number of disjoint derivation fragments during processing (Schuler et al., 2010), in keeping with human working memory limitations (Miller, 1956; Cowan, 2001; McElree, 2001), and correctly predicts difficulty in recognizing center-embedded, but not left- or right-embedded structures (Chomsky and Miller, 1963; Miller and Isard, 1964; Karlsson, 2007). A left-corner parser maintains a sequence of derivation fragments $a/b, a'/b', \ldots$, each consisting of an active category a lacking an awaited category b yet to come. It incrementally assembles trees by forking off and joining up these derivation fragments, using a pair of binary decisions about whether to use a word w to start a new derivation fragment (initially a complete category c):[2]

$$\frac{a/b \quad w}{a/b \quad c} \; b \xrightarrow{+} c \; \ldots \; ; \quad c \rightarrow w \tag{F=1}$$

$$\frac{a/b \quad w}{c} \; a = c; \; b \rightarrow w \tag{F=0}$$

and whether to use a grammatical inference rule to connect a complete category c to a previously disjoint derivation fragment a/b:

$$\frac{a/b \quad c}{a/b'} \; b \rightarrow c \; b' \tag{J=1}$$

$$\frac{a/b \quad c}{a/b \quad a'/b'} \; b \xrightarrow{+} a' \; \ldots \; ; \quad a' \rightarrow c \; b' \tag{J=0}$$

These two binary decisions have four possible outcomes in total: the parser can fork only (which increases the number of derivation fragments by one), join only (which decreases the number of derivation fragments by one), both fork and join (which keeps the number of derivation fragments the same), or neither fork nor join (which also preserves the number of derivation fragments).

An example derivation of the sentence '*We'll get you another one,*' in shown in Figure 1.

3.2 Probabilistic sequence model

A left-corner parser can be modeled as a probabilistic sequence model using hidden random variables at every time step for *Active* categories A, *Awaited* categories B, *Preterminal* or part-of-speech (POS) tags P, and an observed random variable W over *Words*. The model also makes use of two binary switching

[2]Here, $b \xrightarrow{+} c \; \ldots$ constrains c to be a leftmost descendant of b at some depth.

variables at each time step, F (for *Fork*) and J (for *Join*) that guide the transitions of the other states. These two binary switching variables yield four cases: 1/1, 1/0, 0/1 and 0/0 at each time step.

Let D be the depth of the memory store at position t in the sequence, and let the state $q_t^{1..D}$ be the stack of derivation fragments at t, consisting of one active category a_t^d and one awaited category b_t^d at each depth d. The joint probability of the hidden state $q_t^{1..D}$ and observed word w_t, given their previous context, are defined using Markov independence assumptions and the fork-join variable decomposition of van Schijndel et al. (2013), which preserves PCFG probabilities in incremental sentence processing:

$$\begin{aligned}
\mathsf{P}(q_t^{1..D}\ w_t \mid q_{1..t-1}^{1..D}\ w_{1..t-1}) &= \mathsf{P}(q_t^{1..D}\ w_t \mid q_{t-1}^{1..D}) && (1)\\
&\stackrel{\text{def}}{=} \mathsf{P}(p_t\ w_t\ f_t\ j_t\ a_t^{1..D}\ b_t^{1..D} \mid q_{t-1}^{1..D}) && (2)\\
&= \mathsf{P}_{\theta_P}(p_t \mid q_{t-1}^{1..D}) \cdot \\
&\quad \mathsf{P}_{\theta_W}(w_t \mid q_{t-1}^{1..D}\ p_t) \cdot \\
&\quad \mathsf{P}_{\theta_F}(f_t \mid q_{t-1}^{1..D}\ p_t\ w_t) \cdot \\
&\quad \mathsf{P}_{\theta_J}(j_t \mid q_{t-1}^{1..D}\ p_t\ w_t\ f_t) \cdot \\
&\quad \mathsf{P}_{\theta_A}(a_t^{1..D} \mid q_{t-1}^{1..D}\ p_t\ w_t\ f_t\ j_t) \cdot \\
&\quad \mathsf{P}_{\theta_B}(b_t^{1..D} \mid q_{t-1}^{1..D}\ p_t\ w_t\ f_t\ j_t\ a_t^{1..D}) && (3)
\end{aligned}$$

The part-of-speech p_t only depends on the lowest awaited (b_{t-1}^d) category at the previous time step, where d is the depth of the stack at the previous time step and $q_\perp$ is an empty derivation fragment:

$$\mathsf{P}_{\theta_P}(p_t \mid q_{t-1}^{1..D}) \stackrel{\text{def}}{=} \mathsf{P}_{\theta_P}(p_t \mid d\ b_{t-1}^d);\ \ d = \max_{d'}\{q_{t-1}^{d'} \neq q_\perp\} \tag{4}$$

The lexical item (w_t) only depends on the part of speech tag (p_t) at the same time step:

$$\mathsf{P}_{\theta_W}(w_t \mid q_{t-1}^{1..D}\ p_t) \stackrel{\text{def}}{=} \mathsf{P}_{\theta_W}(w_t \mid p_t) \tag{5}$$

The fork decision f_t is assumed to be independent of previous state $q_{t-1}^{1..D}$ variables except for the previous lowest awaited category b_{t-1}^d and part of speech tag p_t:

$$\mathsf{P}_{\theta_F}(f_t \mid q_{t-1}^{1..D}\ p_t\ w_t) \stackrel{\text{def}}{=} \mathsf{P}_{\theta_F}(f_t \mid d\ b_{t-1}^d\ p_t);\ \ d = \max_{d'}\{q_{t-1}^{d'} \neq q_\perp\} \tag{6}$$

The join decision j_t is decomposed into fork and no-fork cases depending on the outcomes of the fork decision:

$$\mathsf{P}_{\theta_J}(j_t \mid q_{t-1}^{1..D}\ f_t\ p_t\ w_t) \stackrel{\text{def}}{=} \begin{cases} \mathsf{P}_{\theta_J}(j_t \mid d\ a_{t-1}^d\ b_{t-1}^{d-1});\ \ d = \max_{d'}\{q_{t-1}^{d'} \neq q_\perp\} & \text{if } f_t = 0 \\ \mathsf{P}_{\theta_J}(j_t \mid d\ p_t\ b_{t-1}^d);\ \ d = \max_{d'}\{q_{t-1}^{d'} \neq q_\perp\} & \text{if } f_t = 1 \end{cases} \tag{7}$$

When f_t=1, that is, a fork has been created, the decision of j_t is whether to immediately integrate the newly forked derivation fragment and transition the awaited category above it (j_t=1) or keep the newly forked derivation fragment (j_t=0). When f_t=0, that is, no fork has been created, the decision of j_t is whether to reduce a stack level (j_t=1) or to transition both the active and awaited categories at the current level (j_t=0).

Decisions about the active categories $a_t^{1..D}$ are decomposed into fork- and join-specific cases depending on the previous state $q_{t-1}^{1..D}$ and the current preterminal p_t. Since the fork and join outcomes only allow a single derivation fragment to be initiated or integrated, each case of the active category model only nondeterministically modifies at most one a_t^d variable from the previous time step:[3]

$$\mathsf{P}_{\theta_A}(a_t^{1..D} \mid q_{t-1}^{1..D}\ f_t\ p_t\ w_t\ j_t) \stackrel{\text{def}}{=}$$

$$\begin{cases} [\![a_t^{1..d-2} = a_{t-1}^{1..d-2}]\!] \cdot [\![a_t^{d-1} = a_{t-1}^{d-1}]\!] \cdot [\![a_t^{d+0..D} = a_\perp]\!];\ \ d = \max_{d'}\{q_{t-1}^{d'} \neq q_\perp\} & \text{if } f_t = 0,\ j_t = 1 \\ [\![a_t^{1..d-1} = a_{t-1}^{1..d-1}]\!] \cdot \mathsf{P}_{\theta_A}(a_t^d \mid d\ b_{t-1}^{d-1}\ a_{t-1}^d) \cdot [\![a_t^{d+1..D} = a_\perp]\!];\ \ d = \max_{d'}\{q_{t-1}^{d'} \neq q_\perp\} & \text{if } f_t = 0,\ j_t = 0 \\ [\![a_t^{1..d-1} = a_{t-1}^{1..d-1}]\!] \cdot [\![a_t^{d} = a_{t-1}^{d}]\!] \cdot [\![a_t^{d+1..D} = a_\perp]\!];\ \ d = \max_{d'}\{q_{t-1}^{d'} \neq q_\perp\} & \text{if } f_t = 1,\ j_t = 1 \\ [\![a_t^{1..d-0} = a_{t-1}^{1..d-0}]\!] \cdot \mathsf{P}_{\theta_A}(a_t^{d+1} \mid d\ b_{t-1}^d\ p_t) \cdot [\![a_t^{d+2..D} = a_\perp]\!];\ \ d = \max_{d'}\{q_{t-1}^{d'} \neq q_\perp\} & \text{if } f_t = 1,\ j_t = 0 \end{cases} \tag{8}$$

[3]Here $[\![\phi]\!]$ is a (deterministic) indicator function, equal to one when ϕ is true and zero otherwise.

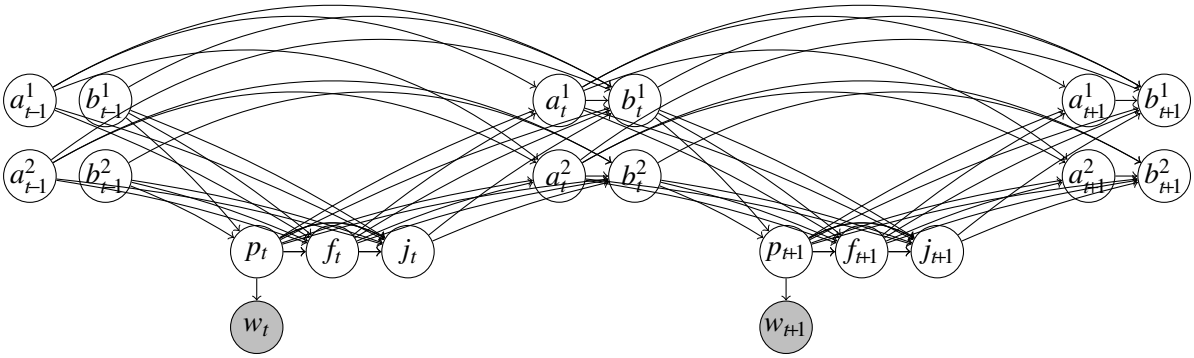


Figure 2: Graphical representation of probabilistic left-corner parsing model expressed in Equations 6–9 across two time steps, with $D = 2$.

Decisions about the awaited categories $b_t^{1..D}$ also depend on the outcome of the fork and join variables. Again, since the fork and join outcomes only allow a single derivation fragment to be initiated or integrated, each case of the awaited category model only nondeterministically modifies at most one b_t^d variable from the previous time step:

$$\mathsf{P}_{\theta_B}(b_t^{1..D} \mid q_{t-1}^{1..D}\ f_t\ p_t\ w_t\ j_t\ a_t^{1..D}) \stackrel{\text{def}}{=}$$

$$\begin{cases} [\![b_t^{1..d-2} = b_{t-1}^{1..d-2}]\!] \cdot \mathsf{P}_{\theta_B}(b_t^{d-1} \mid d\ b_{t-1}^{d-1}\ a_{t-1}^{d}) \cdot [\![b_t^{d+0..D} = b_\perp]\!]; \ d = \max_{d'}\{q_{t-1}^{d'} \neq q_\perp\} & \text{if } f_t = 0,\ j_t = 1 \\ [\![b_t^{1..d-1} = b_{t-1}^{1..d-1}]\!] \cdot \mathsf{P}_{\theta_B}(b_t^{d} \mid d\ a_t^{d}\ a_{t-1}^{d}) \cdot [\![b_t^{d+1..D} = b_\perp]\!]; \ d = \max_{d'}\{q_{t-1}^{d'} \neq q_\perp\} & \text{if } f_t = 0,\ j_t = 0 \\ [\![b_t^{1..d-1} = b_{t-1}^{1..d-1}]\!] \cdot \mathsf{P}_{\theta_B}(b_t^{d} \mid d\ b_{t-1}^{d}\ p_t) \cdot [\![b_t^{d+1..D} = b_\perp]\!]; \ d = \max_{d'}\{q_{t-1}^{d'} \neq q_\perp\} & \text{if } f_t = 1,\ j_t = 1 \\ [\![b_t^{1..d-0} = b_{t-1}^{1..d-0}]\!] \cdot \mathsf{P}_{\theta_B}(b_t^{d+1} \mid d\ a_t^{d+1}\ p_t) \cdot [\![b_t^{d+2..D} = b_\perp]\!]; \ d = \max_{d'}\{q_{t-1}^{d'} \neq q_\perp\} & \text{if } f_t = 1,\ j_t = 0 \end{cases} \tag{9}$$

Thus, the parser has a fixed number of probabilistic decisions to make as it encounters each word, regardless of the depth of the stack. A graphical representation of this model is shown in Figure 2.

3.3 Model priors

Induction in this model follows the approach of Van Gael et al. (2008) by applying nonparametric priors over the active, awaited, and part-of-speech variables. This approach allows the model to learn not only the parameters of the model—such as what parts of speech are likely to be created from what awaited categories—but also the cardinality of how many active, awaited, and part of speech categories are present, in a fully unsupervised fashion. No labels are needed for inference, which alternates between inferring these unseen categories and the associated model parameters.

The probabilistic sequence model defined above, augmented with priors, can be repeatedly sampled to obtain an estimate of the posterior distribution of its hidden variables given a set of observed word sequences. Priors over the syntactic models are based on the infinite hidden Markov model (iHMM) used for part-of-speech tagging (van Gael et al., 2009). In that model, a hierarchical Dirichlet process HMM (Teh et al., 2006) is used to allow the observed number of states—corresponding to parts of speech—in the HMM to grow as the data requires. The hierarchical structure of the iHMM ensures that transition distributions share the same set of states, which would not be possible if we used a flat infinite mixture model.

A fully infinite version of this model uses nonparametric priors on each of the active, awaited, and part-of-speech variables, allowing the cardinality of each of these variables to grow as the data requires. Each model draws a base distribution from a root Dirichlet process, which is then used as a parameter to an infinite set of Dirichlet processes, one each for each applicable combination of the conditioning

variables a_{t-1}, b_{t-1}, p_{t-1}, j_t, f_t, a_t, and b_t:

$$\beta_A \sim GEM(\gamma_A) \tag{10}$$

$$\mathsf{P}_{\theta_A}(a_t^d \mid d\ b_{t-1}^{d-1}\ a_{t-1}^d) \sim DP(\alpha_A, \beta_A) \tag{11}$$

$$\mathsf{P}_{\theta_A}(a_t^{d+1} \mid d\ b_{t-1}^d\ p_t) \sim DP(\alpha_A, \beta_A) \tag{12}$$

$$\beta_B \sim GEM(\gamma_B) \tag{13}$$

$$\mathsf{P}_{\theta_B}(b_t^{d-1} \mid d\ b_{t-1}^{d-1}\ a_{t-1}^{d-1}) \sim DP(\alpha_B, \beta_B) \tag{14}$$

$$\mathsf{P}_{\theta_B}(b_t^d \mid d\ a_t^d\ a_{t-1}^d) \sim DP(\alpha_B, \beta_B) \tag{15}$$

$$\mathsf{P}_{\theta_B}(b_t^d \mid d\ b_{t-1}^d\ p_t) \sim DP(\alpha_B, \beta_B) \tag{16}$$

$$\mathsf{P}_{\theta_B}(b_t^{d+1} \mid d\ a_t^{d+1}\ p_t) \sim DP(\alpha_B, \beta_B) \tag{17}$$

$$\beta_P \sim GEM(\gamma_P) \tag{18}$$

$$\mathsf{P}_{\theta_P}(p_t \mid d\ b_{t-1}^d) \sim DP(\alpha_P, \beta_P) \tag{19}$$

where DP is Dirichlet process and GEM is the stick-breaking construction for DPs (Sethuraman, 1994). Models at depth greater than one use the corresponding model at the previous depth as a prior.

3.4 Inference

Inference is based on the beam sampling approach employed in van Gael et al. (2009) for part-of-speech induction. This inference approach alternates between two phases in each iteration. First, given the distributions θ_F, θ_J, θ_A, θ_B, θ_P, and θ_W, the model resamples values for all the hidden states $\{q_t^d, p_t\}$. Next, given the state values $\{q_t^d, p_t\}$, it resamples each set of multinomial distributions θ_F, θ_J, θ_A, θ_B, θ_P, and θ_W. The sampler is initialized by conservatively setting the cardinalities of the number of active, awaited, and part-of-speech states we expect to see in the data set, randomly initializing the state space, and then sampling the parameters for each distribution θ_F, θ_J, θ_A, θ_B, θ_P, and θ_W given the randomly initialized states and fixed hyperparameters.

As noted by Van Gael et al. (2008), token-level Gibbs sampling in a sequence model can be slow to mix. Preliminary work found that mixing with token-level Gibbs sampling is even slower in this model due to the tight constraints imposed by the switching variables—it is technically ergodic but exploring the state space requires many low probability moves. Therefore, the experiments described in this paper use sentence-level sampling instead of token-level sampling, first computing forward probabilities for the sequence and then doing sampling in a backwards pass; resampling the parameters for the probability distributions only requires computing the counts from the sampled sequence and combining with the hyperparameters. To account for the infinite size of the state spaces, these experiments employ the beam sampler (Van Gael et al., 2008), with some modifications for computational speed.

The standard beam sampler introduces an auxiliary variable u at each time step, which acts as a threshold below which transition probabilities are ignored. This auxiliary variable u is drawn from $Uniform(0, p(q_t^{1..D}|q_{t-1}^{1..D}))$, so it will be between 0 and the probability of the previously sampled transition. The joint distribution over transitions, emissions, and auxiliary variables can be reduced so that the transition matrix is transformed into a boolean matrix with a 1 indicating an allowed transition. Depending on the cut-off value u, the size of the instantiated transition matrix will be different for every time-step.

Values of u can be sampled for active, awaited, and POS variables at every time step, rather than a single u for the transition matrix. It is possible to compile all the operations at each time step into a single large transition matrix, but computing this matrix is prohibitively slow for an operation that must be done at each time step in the data.

To address this issue, the learner may interleave several iterations holding the cardinality of the instantiated space fixed with full beam-sampling steps in which the cardinality of the state space can change.

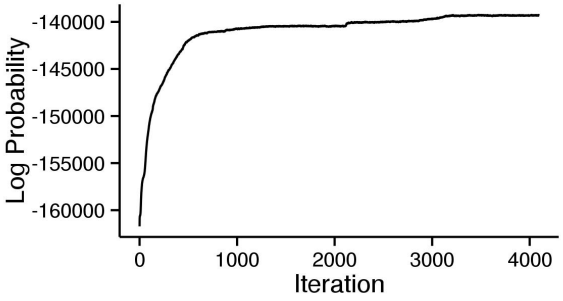


Figure 3: Log Probability (with punc)

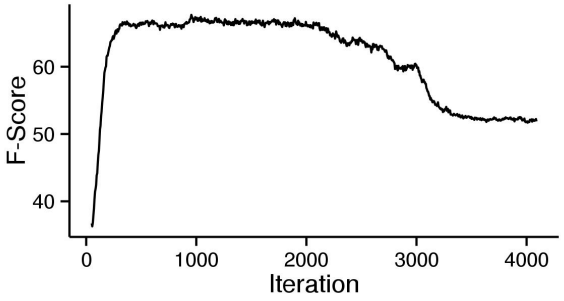


Figure 4: F-Score (with punc)

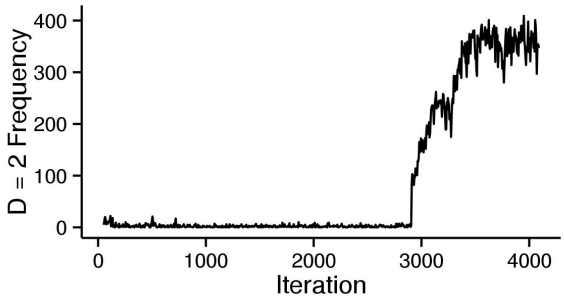


Figure 5: Depth=2 Frequency (with punc)

When the cardinality of the state space is fixed, the learner can multiply out the states into one large, structured transition matrix that is valid for all time steps. The forward pass is thus reduced to an HMM forward pass (albeit one over a much larger set of states), vastly improving the speed of inference. Alternating between sampling the parameters of this matrix and the state values themselves corresponds to updating a finite portion of the infinite possible state space; by interleaving these finite steps with occasional full beam-sampling iterations, the learner is still properly exploring the posterior over models.

3.5 Parsing

There are multiple ways to extract parses from an unsupervised grammar induction system such as this. The optimal Bayesian approach would involve averaging over the values sampled for each model across many iterations, and then use those models in a Viterbi decoding parser to find the best parse for each sentence. Alternatively, if the model parameters have ceased to change much between iterations, the learner can be assumed to have found a local optimum. It can then use a single sample from the end of the run as its model and the analyses of each sentence in that run as the parses to be evaluated. This latter method is used in the experiments described below.

4 Experimental Setup

We ran the UHHMM learner for 4,000 iterations on the approximately 14,500 child-directed utterances of the Eve section of the Brown corpus from the CHILDES database (MacWhinney, 2000).[4] To model the limited memory capacity of young language learners, we restricted the depth of the store of derivation fragments to two.[5] The input sentences were tokenized following the Penn Treebank convention and converted to lower case. Punctuation was initially left in the input as a proxy for intonational phrasal cues (Seginer, 2007; Ponvert et al., 2011), then removed in a follow-up experiment.

[4]We used 4 active states; 4 awaited states; 8 parts of speech; and parameter values 0.5 for α_a, α_b, and α_c, and 1.0 for α_f, α_j, and γ. The burnin period was 50 iterations.

[5]This limited stack depth permits discovery of interesting syntactic features – like subject-aux inversion – while modeling the severe memory limitations of infants (see §2.1). Greater depths are likely unnecessary to parse child-directed input (e.g., Newport et al., 1977).

	With punc			No punc		
	P	R	F1	P	R	F1
UPPARSE	60.50	51.96	55.90	38.17	48.38	42.67
CCL	64.70	53.47	58.55	56.87	47.69	51.88
BMMM+DMV (directed)	62.08	62.51	62.30	61.01	59.24	60.14
BMMM+DMV (undirected)	63.63	**64.02**	**63.82**	61.34	**59.33**	**60.32**
UHHMM-4000, binary	46.68	58.28	51.84	37.62	46.97	41.78
UHHMM-4000, flattened	**68.83**	57.18	62.47	**61.78**	45.52	52.42
Right-branching	68.73	**85.81**	**76.33**	**68.73**	**85.81**	**76.33**

Table 1: Parsing accuracy on Eve with and without punctuation (phrasal cues) in the input. The UHHMM systems were given 8 PoS categories while the BMMM+DMV systems were given 45. UPPARSE and CCL do not learn PoS tags. Only the UHHMM systems model limited working memory capacity or incremental left-corner parsing.

To generate accuracy benchmarks, we parsed the same data set using the three competing raw-text induction systems discussed in §2: CCL (Seginer, 2007), UPPARSE (Ponvert et al., 2011),[6] and both directed and undirected variants of BMMM+DMV (Christodoulopoulos et al., 2012).[7] The BMMM+DMV system generates dependency graphs which are not directly comparable to our phrase-structure output, so we used the algorithm of Collins et al. (1999) to convert the BMMM+DMV output to the flattest phrase structure trees permitted by the dependency graphs.

We evaluated accuracy against hand-corrected gold-standard Penn Treebank-style annotations for Eve (Pearl and Sprouse, 2013). All evaluations were of unlabeled bracketings with punctuation removed.[8] Accuracy results reported for our system are extracted from arbitrary samples taken after convergence had been reached: iteration 4000 for the with-punc model, and iteration 1500 for the no-punc model (see Figures 3 and 6, respectively).

5 Results

Figures 3, 4, and 5 show (respectively) log probability, f-score, and depth=2 frequency by iteration for the UHHMM trained on data containing punctuation. As the figures show, the model remains effectively depth 1 until around iteration 3000, at which point it discovers depth 2, rapidly overgeneralizes it, then scales back to around 350 uses over the entire corpus. Around this time, parsing accuracy drops considerably. This result is consistent with the 'less-is-more' hypothesis (Newport, 1990), since accuracy decreases near the point when the number of plausible hypotheses suddenly grows. In our system, we believe this is because the model reallocates probability mass to deeper parses. Nonetheless, as we show below, final results are state of the art.

We sampled parses from iteration 4000 of our learner for evaluation. As shown in Table 1, initial accuracy measures are worse than all four competitors. However, our system generates exclusively binary-branching output, while all competitors can produce the higher arity trees attested in the PTB-like evaluation standard (notice that our recall measure for the binary branching output beats both CCL and UPPARSE). To correct this disadvantage, we flattened the UHHMM output by first converting binary trees to dependencies using a heuristic that selects for each parent the most frequently co-occurring child category as the head, then converting these dependencies back into phrase structures using the Collins et al. (1999) algorithm. As shown in Table 1, recall remains approximately the same while precision predictably improves, resulting in higher overall F-measures that beat or closely match those of all competing systems.[9]

[6]Using the best cascaded parser settings from that work: probabilistic right-linear grammar with uniform initialization.

[7]We ran both variants of the BMMM+DMV system for 10 generations, with 500 iterations of BMMM and 20 EM iterations of DMV per generation, as was done by Christodoulopoulos et al. (2012).

[8]Note that while punctuation was removed for all evaluations, inclusion/removal of punctuation in the training data was an independent variable in our experiment.

[9]It happens to be the case that these child-directed sentences are heavily right-branching, likely due to the simplicity and

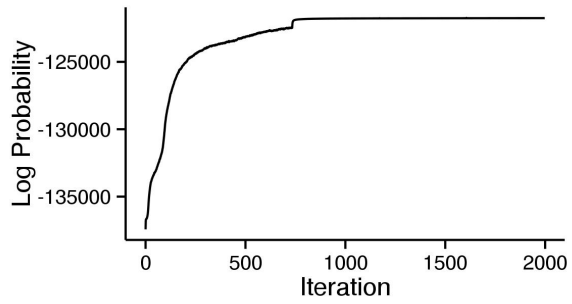


Figure 6: Log Probability (no punc)

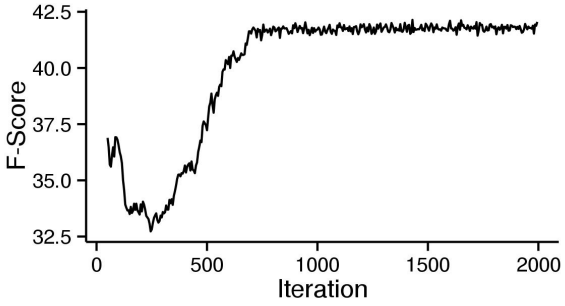


Figure 7: F-Score (no punc)

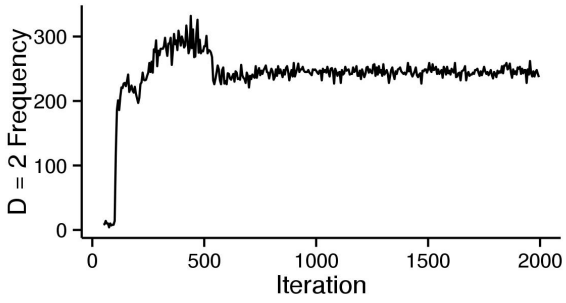


Figure 8: Depth=2 Frequency (no punc)

Figures 6, 7, and 8 show (respectively) log probability, f-score, and depth=2 frequency by iteration for the UHHMM trained on data containing no punctuation. Somewhat surprisingly, the model discovers depth 2 and converges much more quickly than it did for the with-punc corpus, requiring fewer than 1000 iterations to converge. This is possibly due to the slight reduction in corpus size. As in the case of the with-punc trained learner, once depth 2 is discovered, the system quickly overgeneralizes, then converges in a consistent range (in this case around 250 uses of depth 2).

To evaluate accuracy on the punctuation-free data, we sampled parses from iteration 1500 of our learner. Results are given in Table 1. Binary UHHMM results are on par with UPPARSE, worse than CCL, and considerably worse than BMMM+DMV, while flattened UHHMM results show higher overall F-measures than both CCL and UPPARSE. BMMM+DMV suffers less in the absence of punctuation than the other systems (and therefore generally provides the best induction results on no-punc). The large drop in UHHMM accuracy with the removal of punctuation provides weak evidence for the use of intonational phrasal cues in human syntax acquisition.

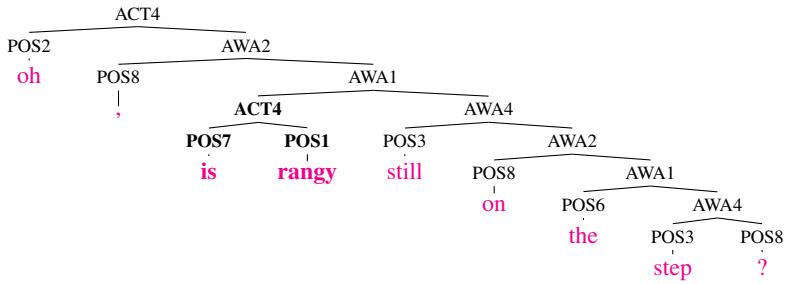
While the BMMM+DMV results are on par with ours, it is important to note that we used a severely restricted number of categories in order to improve computational efficiency. For example, our system was given 8 PoS tags to work with, while BMMM+DMV was given 45. Finer grained state spaces in a more efficient implementation of our learner will hopefully improve upon the results presented here.

Finally, it is interesting to observe that the uses of depth 2 shown in Figures 5 and 8 are in general linguistically well-motivated. They tend to occur in subject-auxiliary inversion, ditransitive, and contraction constructions, in which depth 2 is often necessary in order to bracket auxiliary+subject, verb+object, and verb+contraction together, as illustrated in Figure 9. Unfortunately, due to the flat representation of these constructions in the gold standard trees, this insight on the part of our learner is not reflected in the accuracy measures in Table 1.

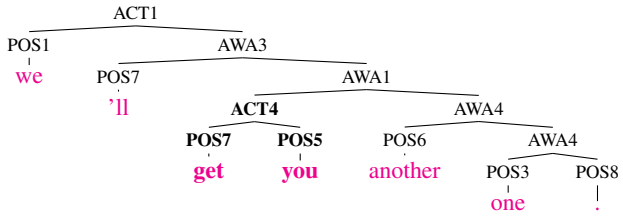
---

short length of child-directed utterances, and therefore the right-branching baseline (RB) outperforms all systems by a wide margin on this corpus. However, we argue that such utterances are a more realistic model of input to human language learners than newswire text, and therefore preferable for evaluation of systems that purport to model human language acquisition. Our system learns this directional bias from data, and does so at least as successfully as its competitors.

1. **Subject-auxiliary inversion:**



2. **Ditransitive:**



3. **Contraction:**

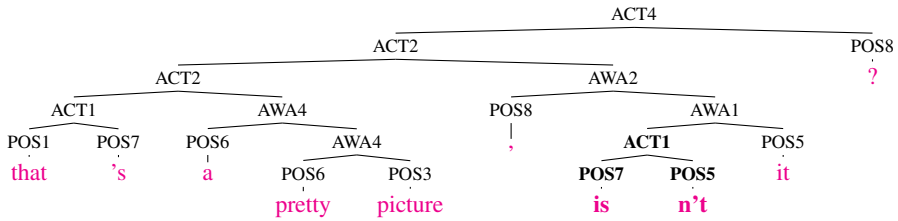


Figure 9: Actual parses from UHHMM-4000 (with punctuation), illustrating the use of depth 2 (bold) for subject-aux inversion, ditransitives, and contractions.

6 Conclusion

This paper presented a grammar induction sytem that models the working memory limitations of young language learners and employs a cognitively plausible left-corner incremental parsing strategy, in contrast to existing raw-text induction systems. The fact that our system can model these aspects of human language acquisition and sentence processing while achieving the competitive results shown here on a corpus of child-directed speech indicates that humans can in principle learn a good deal of natural language syntax from distributional statistics alone. It also shows that modeling cognition more closely can match or improve on existing approaches to the task of raw-text grammar induction.

In future research, we intend to make use of parallel processing techniques to increase the speed of inference and (1) allow the system to infer the optimal number of states in each component of the model, permitting additional granularity that might enable it to discover subtler patterns than is possible with our currently-restricted state inventories, and (2) allow the system to make use of depths 3 and 4, modeling working memory capacities of older learners.

Acknowledgements

The authors would like to thank the anonymous reviewers for their comments. This project was sponsored by the Defense Advanced Research Projects Agency award #HR0011-15-2-0022. The content of the information does not necessarily reflect the position or the policy of the Government, and no official endorsement should be inferred.

References

Steven P. Abney and Mark Johnson. 1991. Memory requirements and local ambiguities of parsing strategies. *J. Psycholinguistic Research*, 20(3):233–250.

Alfred V. Aho and Jeffery D. Ullman. 1972. *The Theory of Parsing, Translation and Compiling, Vol. 1: Parsing*. Prentice-Hall, Englewood Cliffs, New Jersey.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum learning. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 41–48, Montreal.

Noam Chomsky and George A. Miller. 1963. Introduction to the formal analysis of natural languages. In *Handbook of Mathematical Psychology*, pages 269–321. Wiley, New York, NY.

Christos Christodoulopoulos, Sharon Goldwater, and Mark Steedman. 2011. A Bayesian mixture model for part-of-speech induction using multiple features. In *Proceedings of EMNLP*, pages 638–647, Edinburgh, Scotland, 7.

Christos Christodoulopoulos, Sharon Goldwater, and Mark Steedman. 2012. Turning the pipeline into a loop: Iterated unsupervised dependency parsing and PoS induction. In *NAACL-HLT Workshop on the Induction of Linguistic Structure*, pages 96–99, Montreal, Canada, 6.

Michael Collins, Jan Hajic, Lance A. Ramshaw, and Christoph Tillman. 1999. A statistical parser for Czech. In *Proceedings of ACL*.

Nelson Cowan. 2001. The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences*, 24:87–185.

Jeffrey L. Elman. 1993. Learning and development in neural networks: The importance of starting small. *Cognition*, 48:71–99.

Susan E. Gathercole. 1998. The development of memory. *Journal of Child Psychology and Psychiatry*, 39:3–27.

Edward Gibson. 1991. *A computational theory of human linguistic processing: Memory limitations and processing breakdown*. Ph.D. thesis, Carnegie Mellon.

Boris Goldowsky and Elissa Newport. 1993. Modeling the effects of processing limitations on the acquisition of morphology: the less is more hypothesis. In Jonathan Mead, editor, *Proceedings of the 11th West Coast Conference on Formal Linguistics*, pages 234–247.

Philip N. Johnson-Laird. 1983. *Mental models: Towards a cognitive science of language, inference, and consciousness*. Harvard University Press, Cambridge, MA, USA.

Yakoov Kareev, Iris Lieberman, and Miri Lev. 1997. Through a narrow window: Sample size and the perception of correlation. *Journal of Experimental Psychology*, 126:278–287.

Fred Karlsson. 2007. Constraints on multiple center-embedding of clauses. *Journal of Linguistics*, 43:365–392.

Dan Klein and Christopher D. Manning. 2002. A generative constituent-context model for improved grammar induction. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*.

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*.

Shalom Lappin and Suart M. Shieber. 2007. Machine learning theory and practice as a source of insight into universal grammar. *Journal of Linguistics*, 43:1–34.

Richard L. Lewis and Shravan Vasishth. 2005. An activation-based model of sentence processing as skilled memory retrieval. *Cognitive Science*, 29(3):375–419.

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

Brian McElree. 2001. Working memory and focal attention. *Journal of Experimental Psychology, Learning Memory and Cognition*, 27(3):817–835.

George A. Miller and Stephen Isard. 1964. Free recall of self-embedded english sentences. *Information and Control*, 7:292–303.

George A. Miller. 1956. The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, 63:81–97.

Elissa Newport, Henry Gleitman, and Lila Gleitman. 1977. Mother, I'd rather do it myself: Some effects and non-effects of maternal speech style. In Catherine F. Snow, editor, *Talking to Children*, pages 109–149. Cambridge University Press, Cambridge.

Elissa Newport. 1990. Maturational constraints on language learning. *Cognitive Science*, 14:11–28.

Lisa Pearl and Jon Sprouse. 2013. Syntactic islands and learning biases: Combining experimental syntax and computational modeling to investigate the language acquisition problem. *Language Acquisition*, 20:23–68.

Elias Ponvert, Jason Baldridge, and Katrin Erik. 2011. Simple unsupervised grammar induction from raw text with cascaded finite state models. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, pages 1077–1086, Portland, Oregon, 6.

Philip Resnik. 1992. Left-corner parsing and psychological plausibility. In *Proceedings of COLING*, pages 191–197, Nantes, France.

Douglas L.T. Rohde and David C. Plaut. 2003. Less is less in language acquisition. In Philip Quinlan, editor, *Connectionist modelling of cognitive development*. Psychology Press, Hove, UK.

Jenny R Saffran, Elizabeth K Johnson, Richard N Aslin, and Elissa L Newport. 1999. Statistical learning of tone sequences by human infants and adults. *Cognition*, 70(1):27–52.

Lynn Santelman and Peter W. Jusczyk. 1998. Sensitivity to discontinuous dependencies in language learners: Evidence for limitations in processing space. *Cognition*, 69:105–34.

William Schuler, Samir AbdelRahman, Tim Miller, and Lane Schwartz. 2010. Broad-coverage incremental parsing using human-like memory constraints. *Computational Linguistics*, 36(1):1–30.

Yoav Seginer. 2007. Fast unsupervised incremental parsing. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 384–391.

Amanda Seidl, George Hollich, and Peter W. Jusczyk. 2003. Early understanding of subject and object wh-questions. *Infancy*, 4(3):423–436.

Jayaram Sethuraman. 1994. A constructive definition of Dirichlet priors. *Statistica Sinica*, 4:639–650.

Edward Stabler. 1994. The finite connectivity of linguistic structure. In *Perspectives on Sentence Processing*, pages 303–336. Lawrence Erlbaum.

Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. 2006. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581.

Jurgen Van Gael, Yunus Saatci, Yee Whye Teh, and Zoubin Ghahramani. 2008. Beam sampling for the infinite hidden Markov model. In *Proceedings of the 25th international conference on Machine learning*, pages 1088–1095. ACM.

Jurgen van Gael, Andreas Vlachos, and Zoubin Ghahramani. 2009. The infinite HMM for unsupervised PoS tagging. (August):678–687.

Marten van Schijndel, Andy Exley, and William Schuler. 2013. A model of language processing as hierarchic sequential prediction. *Topics in Cognitive Science*, 5(3):522–540.