Depth-Bounded Statistical PCFG Induction as a Model of Human Grammar Acquisition

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

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

Finale Doshi-Velez
Harvard University
Department of Computer Science
finale@seas.harvard.edu

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

William Schuler*
The Ohio State University
Department of Linguistics
schuler@ling.osu.edu

This article describes a simple PCFG induction model with a fixed category domain that predicts a large majority of attested constituent boundaries, and predicts labels consistent with nearly half of attested constituent labels on a standard evaluation data set of child-directed speech. The article then explores the idea that the difference between simple grammars exhibited by child learners and fully recursive grammars exhibited by adult learners may be an effect of increasing working memory capacity, where the shallow grammars are constrained images of the

Submission received: 4 March 2020; revised version received: 18 December 2020; accepted for publication: 20 December 2020.

https://doi.org/10.1162/COLI_a_00399

recursive grammars. An implementation of these memory bounds as limits on center embedding in a depth-specific transform of a recursive grammar yields a significant improvement over an equivalent but unbounded baseline, suggesting that this arrangement may indeed confer a learning advantage.

1. Introduction

Chomsky (1965) postulates that as human children are naturally exposed to a language, the quantity and nature of the linguistic examples to which they are exposed is insufficient to fully explain the children's successful acquisition of the grammar of the language; Chomsky (1980) dubs this claim *the poverty of the stimulus*. Chomsky (1965) asserts that the space of possible human languages must therefore be constrained by a set of linguistic universals with which children's brains are innately primed, and that this biological fact is a necessary precondition for human language learning. Chomsky (1986) uses the term *Universal Grammar* to describe this proposed innate mental model that underlies human language acquisition.

The argument from the poverty of the stimulus and the associated claim of an innate Universal Grammar gained wide acceptance within the Chomskyan generative tradition. The specific details of exactly what aspects of language cannot be learned without Universal Grammar has not always been well defined; similarly, the nature of exactly what proposed linguistic universals constitute Universal Grammar have been widely debated. In striving to identify empirical mechanisms by which poverty of the stimulus claims might be rigorously tested, Pullum and Scholz (2002) conclude that although such claims could potentially be true, the linguistic examples most widely cited in support fail to hold up to close scrutiny.

Pullum and Scholz argue that mathematical learning theory and corpus linguistics have a key role to play in empirically testing poverty of the stimulus claims. Preliminary work along these lines using manually constructed grammars of child-directed speech was performed by Perfors, Tenenbaum, and Regier (2006), who demonstrate empirically that a basic learner, when presented with a corpus of child-directed speech, can learn to prefer a hierarchical grammar (a probabilistic context-free grammar) over linear and regular grammars using a simple Bayesian probabilistic measure of the complexity of a grammar.

However, full induction of probabilistic context-free grammars (PCFGs) has long been considered a difficult problem (Solomonoff 1964; Fu and Booth 1975; Carroll and Charniak 1992; Johnson, Griffiths, and Goldwater 2007; Liang et al. 2007; Tu 2012). Lack of success for direct estimation was attributed either to a lack of correlation between the linguistic accuracy and the optimization objective (Johnson, Griffiths, and Goldwater 2007), or the likelihood function or the posterior being filled with weak local optima (Smith 2006; Liang et al. 2007). The first contribution of this article is to describe a simple PCFG induction model with a fixed category domain that predicts a large majority of attested constituent boundaries, and predicts labels consistent with nearly half of attested constituent labels on the Eve corpus, a standard evaluation data set of child-directed speech.

But evidence suggests that children learn very constrained grammars (Lieven, Pine, and Baldwin 1997; Tomasello 2003, and more). These non-nativist models (Bannard, Lieven, and Tomasello 2009) usually assume that the grammar children first acquire is linear and templatic, consisting of multiword frames with slots to be filled in or just *n*-grams. The grammar may also include various kinds of rule-like probabilities for the frames or transition probabilities for the words or *n*-grams. Much work (Redington, Chater, and Finch 1998; Mintz 2003; Freudenthal et al. 2007; Thompson and Newport 2007) shows that syntactic categories and surface word order may be captured with these simple statistics without hypothesizing hierarchical structures. However, the transition between those linear or very shallow grammars and fully recursive grammars is never explicitly modeled; therefore, there is no empirical evidence from computational modeling about how easy this transition may be. The second contribution of this article is to explore the idea that this difference between shallow and fully recursive grammars is determined by working memory, so the shallow and recursive grammars are unified into different performance grammars sharing the same underlying competence grammar.

There has long been a distinction within the linguistic discipline of theoretical syntax between a hypothesized model of language that is posited to exist within in the brain of each speaker of that language and the phenomenon of language as it is actually spoken and encountered in the real world. The concept of a mental model of language has been described in terms of *langue* (de Saussure 1916), linguistic competence (Chomsky 1965), or simply as the grammar of the language, while the details of how language is actually spoken and used have been described as *parole* (de Saussure 1916), linguistic performance (Chomsky 1965), or sometimes as usage.

Chomsky (1965) argues that models of linguistic performance should be informed by models of linguistic competence, but that models of competence should not take performance into account: "Linguistic theory is concerned primarily with an ideal speaker-listener, in a completely homogeneous speech-community, who knows its language perfectly and is unaffected by such grammatically irrelevant conditions as memory limitations, distractions, shifts of attention and interest, and errors" (page 3). Within the Chomskyan generative tradition, this idea that syntactic theory should model an idealized grammar of linguistic competence (rather than one that incorporates performance) has remained dominant in the decades since (see Newmeyer 2003, for example). Others outside this tradition have criticized the Chomskyan position in part for its failure to connect idealized theories of competence to actual language usage (for example, see Pylyshyn 1973; Miller 1975; Kates 1976).

The framework for unsupervised grammar induction presented in this article is significant in that it represents a concrete discovery procedure that can produce both a competence grammar **G** (a PCFG in Chomsky normal form) and a corresponding formally defined performance grammar $\mathbf{G}_D$ (another PCFG defined to be sensitive to center-embedding depth). Although PCFGs in principle allow for unlimited recursion in the form of center-embedding (Chomsky and Miller 1963), evidence from corpus studies of spoken and written language use strongly indicates that such recursion essentially never extends beyond the limits of human cognitive memory constraints (Schuler et al. 2010; Noji, Miyao, and Johnson 2016). Given a cognitively motivated recursive depth bound D, performance grammar $\mathbf{G}_D$ can be viewed as a specific instantiation of competence grammar **G** that is guaranteed to never violate the depth bound. In this analysis of model behavior and depth-bounding (§8) we observe that by utilizing a depth bound, the grammar induction procedure is more consistent in discovering a highly accurate grammar than it is when inducing an unbounded grammar over the same corpus. This fact argues against Chomsky's assertion that memory limitations are an irrelevant consideration in the search for a grammar of a language.

This article is an extended presentation of Jin et al. (2018a) with additional evaluation and analyses of PCFG induction prior to depth bounding. These additional evaluations and analyses include quantitative analyses of effects of manipulation of hyperparameters, and quantitative and qualitative linguistic analyses of categories and rules in generated grammars for several languages. Code used in this work can be found at `https://github.com/lifengjin/pcfg_induction`.

The remainder of this article is organized as follows: Section 2 describes related work in unsupervised grammar induction. Section 3 describes an unbounded PCFG induction model based on Gibbs sampling. Section 4 describes a depth-bounded version of this model. Section 5 describes a method for evaluating labeled parsing accuracy for unsupervised grammar induction. Section 6 describes experiments to evaluate the unbounded PCFG induction model on synthetic data with a known solution. Section 7 describes experiments to evaluate the unbounded PCFG induction model on child-directed speech. Section 8 describes experiments to evaluate the depth-bounded PCFG induction model on child-directed speech. Section 9 describes experiments to explore the phenomena of natural bounding in induction on child-directed and adult language data. Section 10 describes replication of these results on newswire data. Finally, Section 11 provides some concluding remarks.

2. Related work

Unsupervised grammar inducers hypothesize hierarchical structures for strings of words. Using context-free grammars (CFGs) to define these structures with labels, previous attempts at either CFG parameter estimation (Carroll and Charniak 1992; Pereira and Schabes 1992; Johnson, Griffiths, and Goldwater 2007) or directly inducing a CFG as well as its probabilities (Liang et al. 2007; Tu 2012) have not achieved as much success as experiments with other kinds of formalisms that produce unlabeled constituents (Klein and Manning 2004; Seginer 2007a; Ponvert, Baldridge, and Erk 2011). The assumption has been made that the space of grammars is so big that constraints must be applied to the learning process to reduce the burden of the learner (Gold 1967; Cramer 2007; Liang et al. 2007).

Much of this grammar induction work used strong linguistically motivated constraints or direct linguistic annotation to help the inducer eliminate some local optima. Pereira and Schabes (1992) use bracketed corpora to provide extra structural information to the inducer. Use of part-of-speech (POS) sequences in place of word strings is popular in the dependency grammar induction literature (Klein and Manning 2002, 2004; Berg-Kirkpatrick et al. 2010; Jiang, Han, and Tu 2016; Noji, Miyao, and Johnson 2016). Combinatory Categorial Grammar (CCG) induction also relies on a limited number POS tags to assign basic categories to words (Bisk and Hockenmaier 2012; Bisk, Christodoulopoulos, and Hockenmaier 2015), among other constraints such as CCG combinators, to induce labeled dependencies. Other linguistic constraints and heuristics such as constraints of root nodes (Noji, Miyao, and Johnson 2016), attachment rules (Naseem et al. 2010), acoustic cues (Pate and Goldwater 2013), and punctuation as phrasal boundaries (Seginer 2007a; Ponvert, Baldridge, and Erk 2011) have also been used in induction. More recently, neural PCFG induction systems (Jin et al. 2019; Kim et al. 2019; Kim, Dyer, and Rush 2019) and unsupervised parsing models (Shen et al. 2018, 2019; Drozdov et al. 2019) have been shown to predict accurate syntactic structures. These more complex neural network models may not contain explicit biases, but may contain implicit confounding factors implemented during development on English or other natural languages, which may function like linguistic universals in constraining the search over possible grammars. Experiments described in this article use only Bayesian PCFG induction in order to eliminate these possible confounds and evaluate the hypothesis that grammar may be acquired using only event

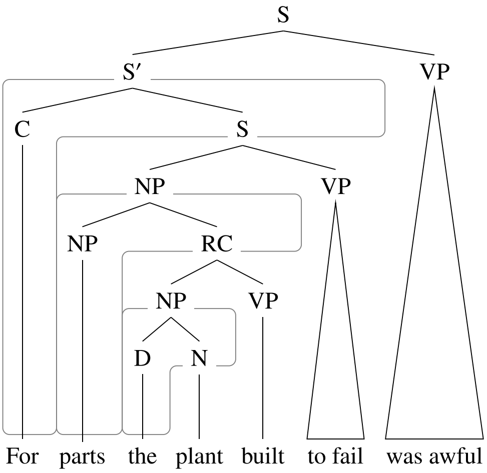


Figure 1
Stack elements after the word *the* in a left-corner parse of the sentence *For parts the plant built to fail was awful.*

categorization and decomposition into categorized sub-events using mathematically transparent parameters.[1]

Depth-like constraints have been applied in work by Seginer (2007a) and Ponvert, Baldridge, and Erk (2011) to help constrain the search over possible structures. Both of these systems are successful in inducing phrase structure trees from only words, but only generate unlabeled constituents. Center-embedding constraints on recursion depth have also been applied to parsing (Schuler et al. 2010; Ponvert, Baldridge, and Erk 2011; Shain et al. 2016; Noji, Miyao, and Johnson 2016; Jin et al. 2018b), motivated by human cognitive constraints on memory capacity (Chomsky and Miller 1963). Center-embedding recursion depth can be defined in a left-corner parsing paradigm (Rosenkrantz and Lewis 1970; Johnson-Laird 1983; Abney and Johnson 1991) as the number of left children of right children that occur on the path from a word to the root of a parse tree. Left-corner parsers require only minimal stack memory to process left-branching and right-branching structures, but require an extra stack element to process each center embedding in a structure. For example, a left-corner parser must add a stack element for each of the first three words in the sentence, *For parts the plant built to fail was awful*, shown in Figure 1. These kinds of depth bounds in sentence processing have been used to explain the relative difficulty of center-embedded sentences compared with more right-branching paraphrases like *It was awful for the plant's parts to fail*. However, depth-bounded grammar induction has never been compared against unbounded induction in the same system, in part because most previous depth-bounding models are built around sequence models, the complexity of which grows exponentially with the maximum allowed depth.

In order to compare the effects of depth-bounding more directly, this work extends a chart-based Bayesian PCFG induction model (Johnson, Griffiths, and Goldwater 2007) to include depth bounding, which allows both bounded and unbounded PCFGs to be induced from unannotated text. Experiments reported in this article confirm that

[1] It is also not straightforward to augment neural network models to test the contribution of depth bounds.

depth-bounding does empirically have the effect of significantly limiting the search space of the inducer. This work also shows that it is possible to induce an accurate unbounded PCFG from raw text with no strong linguistic constraints.

3. Unbounded Statistical Grammar Induction Model

Experiments described in this article evaluate the extent to which natural language grammars learned by humans may simply be those grammars with the highest posterior probability given the sentence data to which they are exposed. These posterior probabilities $\mathsf{P}(\text{grammar} \mid \text{sentences})$ are equivalent to the product of the probability of a grammar, multiplied by the probability of a set of trees given that (probabilistic) grammar, multiplied by the probability of the sentence data given those trees, summed over all possible trees, then divided by the probability of those sentences:

$$\begin{aligned}\mathsf{P}(\text{grammar} \mid \text{sentences}) &= \frac{\sum_{\text{trees}} \mathsf{P}(\text{grammar}, \text{trees}, \text{sentences})}{\mathsf{P}(\text{sentences})} \\ &= \frac{\sum_{\text{trees}} \mathsf{P}(\text{grammar}) \cdot \mathsf{P}(\text{trees} \mid \text{grammar}) \cdot \mathsf{P}(\text{sentences} \mid \text{trees})}{\mathsf{P}(\text{sentences})} \end{aligned} \tag{1}$$

This factoring suggests that a maximum over probabilistic grammars may be estimated by a process of randomly generating a large set of grammars and a large set of trees given each grammar, then calculating the fraction of generated trees whose words match the observed sentences.

More specifically, the generative induction model used in these experiments assumes a PCFG in Chomsky normal form (allowing only unary expansions at preterminals and binary expansions at non-preterminals) with a set C of category labels. This grammar is implemented as a matrix $\mathbf{G}$ of rule probabilities $\mathsf{P}(c \rightarrow a\ b \mid c)$ or $\mathsf{P}(c \rightarrow w \mid c)$ with one row for each of C parent symbols c and one column for each of $|C|^2 + |W|$ combinations of left and right child symbols a and b, which can be pairs of nonterminals or observed words from vocabulary W followed by null symbols $\perp$. For example, a grammar consisting of the probabilistic rules shown in Figure 2a can be represented by the matrix in Figure 2b. This grammar matrix can be defined using a Kronecker delta column vector δ_c (a vector with ones at index c and zeros elsewhere) to index parent categories as rows, and Kronecker product $\delta_a^\top \otimes \delta_b^\top$ of Kronecker delta row vectors $\delta_a^\top$ and $\delta_b^\top$ to index every combination of left child a and right child b categories in a single large vector, as columns (see Figure 3). Each vector of combinations of left and right

a)
$$\begin{aligned}
\mathsf{P}(\text{NP} \rightarrow \text{NP NP} \mid \text{NP}) &= 0.1\\
\mathsf{P}(\text{NP} \rightarrow \text{NP PP} \mid \text{NP}) &= 0.4\\
\mathsf{P}(\text{PP} \rightarrow \text{P NP} \mid \text{PP}) &= 0.8\\
\mathsf{P}(\text{PP} \rightarrow \text{P PP} \mid \text{PP}) &= 0.2\\
\mathsf{P}(\text{NP} \rightarrow \text{books} \perp \mid \text{NP}) &= 0.3\\
\mathsf{P}(\text{NP} \rightarrow \text{stories} \perp \mid \text{NP}) &= 0.2\\
\mathsf{P}(\text{P} \rightarrow \text{about} \perp \mid \text{P}) &= 1.0
\end{aligned}$$

b)
$$\begin{array}{c|cccccccccccc}
 & \text{NP NP} & \text{NP PP} & \text{NP P} & \text{PP NP} & \text{PP PP} & \text{PP P} & \text{P NP} & \text{P PP} & \text{P P} & \text{books} \perp & \text{stories} \perp & \text{about} \perp \\
\text{NP} & .1 & .4 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & .3 & .2 & 0 \\
\text{PP} & 0 & 0 & 0 & 0 & 0 & 0 & .8 & .2 & 0 & 0 & 0 & 0 \\
\text{P} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1
\end{array}$$

Figure 2
Example matrix representation (b) of a probabilistic context-free grammar (a).

$$\text{a)}\quad \begin{array}{r} \text{NP} \\ \text{PP} \\ \text{P} \end{array}\left[\begin{array}{c} 0 \\ 1 \\ 0 \end{array}\right]
\qquad
\text{b)}\ \underbrace{\left[\begin{array}{ccc} \text{NP} & \text{PP} & \text{P} \\ 0 & 1 & 0 \end{array}\right]}_{\delta^{\top}_{\text{PP}}} \otimes \underbrace{\left[\begin{array}{ccc} \text{NP} & \text{PP} & \text{P} \\ 0 & 0 & 1 \end{array}\right]}_{\delta^{\top}_{\text{P}}} = \left[\begin{array}{ccccccccc} \text{NP NP} & \text{NP PP} & \text{NP P} & \text{PP NP} & \text{PP PP} & \text{PP P} & \text{P NP} & \text{P PP} & \text{P P} \\ \multicolumn{3}{c}{\underbrace{0\quad 0\quad 0}_{(\delta^{\top}_{\text{PP}})_{[1]}\ \delta^{\top}_{\text{P}}}} & \multicolumn{3}{c}{\underbrace{0\quad 0\quad 1}_{(\delta^{\top}_{\text{PP}})_{[2]}\ \delta^{\top}_{\text{P}}}} & \multicolumn{3}{c}{\underbrace{0\quad 0\quad 0}_{(\delta^{\top}_{\text{PP}})_{[3]}\ \delta^{\top}_{\text{P}}}} \end{array}\right]$$

Figure 3
Indexing using a Kronecker delta (a), and a Kronecker product of Kronecker deltas (b).

child categories is then concatenated with a vector of probabilities over words w indexed by a Kronecker delta row vector $\delta_w^\top$, to compose each row of $\mathbf{G}$:[2]

$$\mathbf{G} = \sum_c \delta_c \left[\left(\sum_{a,b} \mathsf{P}(c \rightarrow a\ b \mid c)\ \delta_a^\top \otimes \delta_b^\top \right) \left(\sum_w \mathsf{P}(c \rightarrow w \mid c)\ \delta_w^\top \right) \right] \tag{2}$$

The P(grammar) term in Equation (1) is defined to be a Dirichlet distribution over expansions $\mathsf{P}(c \rightarrow a\ b \mid c)$ of each category c, with symmetric parameter β, and this is the distribution from which grammars are randomly sampled in a generative process:

$$\mathsf{P}(\text{grammar } \mathbf{G}) = \text{Dirichlet}(\mathbf{G}; \beta) \qquad\qquad \mathbf{G} \sim \text{Dirichlet}(\beta) \tag{3}$$

A Dirichlet distribution multiplies in a likelihood term for $\beta - 1$ hypothetical instances of each categorical outcome, and renormalizes over the size of the probability simplex (the space of well-formed categorical distributions). The symmetric parameter therefore biases the inducer to prefer (if high) more uniform distributions or (if low) distributions with a small number of high-probability expansions for each parent category.

Probabilities and random sampling processes for trees are defined recursively over expansions from each parent category to its left and right child category. Each tree τ is accounted here as a set $\{\tau_\epsilon, \tau_1, \tau_2, \tau_{11}, \tau_{12}, \tau_{21}, ...\}$ of category labels τ_η, where $\eta \in \{1,2\}^*$ is a Gorn address specifying a path of left (1) or right (2) branches from the root. The distribution over trees P(trees | grammar) in Equation (1) is then defined to be a product of probabilities of all grammar rule expansions in each tree, so trees are randomly sampled from the top down in our generative process, using a categorical distribution over pairs of left and right child category labels $\tau_{\eta 1}$ and $\tau_{\eta 2}$ (drawn from the union of $C \times C$ and $W \times \{\bot\}$) given each parent category label τ_η in each tree τ:

$$\mathsf{P}(\text{trees } \tau_{1..N} \mid \text{grammar } \mathbf{G}) = \prod_{\tau \in \tau_{1..N}} \prod_{\tau_\eta \in \tau} \delta_{\tau_\eta}{}^\top \mathbf{G}\, (\delta_{\tau_{\eta 1}} \otimes \delta_{\tau_{\eta 2}}) \qquad \tau_{\eta 1}, \tau_{\eta 2} \sim \text{Categorical}(\delta_{\tau_\eta}{}^\top \mathbf{G}) \tag{4}$$

where τ_η in W or $\{\bot\}$ are taken to be terminal and not expanded.

Finally, because each tree contains the words of a sentence, the probability of sentences given trees P(sentences | trees) in Equation (1) is simply one if the words in all the trees match the sentences in the corpus, and zero otherwise.

[2] A Kronecker product multiplies two matrices of dimension $m \times n$ and $o \times p$ (or vectors in case n and p equal one) into a matrix of dimension $mo \times np$ consisting of a copy of the first matrix with each element replaced by a copy of the second matrix multiplied by that element.

With unlimited resources, it would be possible to make claims about the posterior probabilities of CFGs given sentences in some training corpus by randomly generating a sufficiently large set of grammars and a sufficiently large set of trees using this sampling process, then calculating the fraction of generated trees that match the training corpus. Grammars that generate the corpus more frequently could then be said to have greater statistical support from the corpus, and thus be natural candidates for induction. However, because the space of probabilistic grammars and corpora is vast and resources are limited, a Gibbs sampling approach (Goodman 1998; Johnson, Griffiths, and Goldwater 2007) is instead used to estimate the posterior distribution over grammars given a corpus by randomly walking through this space, starting from some random probabilistic grammar and a random set of trees, then making a series of random changes to that grammar's rule probabilities in a way that is proportional to the posterior distribution over grammars given a corpus. This is done by randomly generating a new grammar at each step t given the observations of rules used in the previous set of trees:

$$\mathbf{G}^{(t)} \sim \text{Dirichlet}\left(\beta + \sum_{\tau\in\tau^{(t-1)}_{1..N}} \sum_{\tau_\eta\in\tau} \delta_{\tau_\eta} \left(\delta_{\tau_{\eta 1}} \otimes \delta_{\tau_{\eta 2}}\right)^\top\right) \tag{5}$$

then randomly generating a new set of trees $\tau^{(t)}_{1..N}$ for the corpus sentences given the current grammar. Each tree $\tau^{(t)}$ is sampled from the top down, for each span $\tau^{(t)}_\eta$ from word i to word j, by first choosing a split point $k_{i,j}$ such that $i < k_{i,j} < j$, then sampling a pair of category labels $c_{i,k_{i,j}}$ (for $\tau^{(t)}_{\eta 1}$) and $c_{k_{i,j},j}$ (for $\tau^{(t)}_{\eta 2}$) adjacent at this split point, both using vectors of likelihoods $\mathbf{v}_{i,j}$ for words i through j given each possible category label:

$$k_{i,j} \sim \text{Categorical}\left(\sum_{k\in\{i+1..j-1\}} \delta_k \, \delta^\top_{c_{i,j}} \mathbf{G}^{(t)} \left(\mathbf{v}_{i,k} \otimes \mathbf{v}_{k,j}\right)\right) \tag{6a}$$

$$c_{i,k}, c_{k,j} \sim \text{Categorical}\left(\delta^\top_{c_{i,j}} \mathbf{G}^{(t)} \, \text{diag}(\mathbf{v}_{i,k} \otimes \mathbf{v}_{k,j})\right) \tag{6b}$$

where diag($\mathbf{v}$) defines a matrix with elements of vector $\mathbf{v}$ along its diagonal and zeros elsewhere. The vector of likelihoods $\mathbf{v}_{i,j}$ for each span of words is defined recursively in terms of likelihood vectors for each possible left and right child span (if the span contains multiple words), concatenated with a Kronecker delta vector concentrated at the current word (if the span contains just one word):

$$\mathbf{v}_{i,j} = \mathbf{G}^{(t)} \begin{bmatrix} \sum_{k\in\{i+1..j-1\}} \mathbf{v}_{i,k} \otimes \mathbf{v}_{k,j} \\ [\![i+1=j]\!] \, \delta_{w_i} \end{bmatrix} \tag{7}$$

where $[\![\phi]\!]$ is one if ϕ is true and zero otherwise. Figure 4 shows the Gibbs sampling process for unbounded grammars.

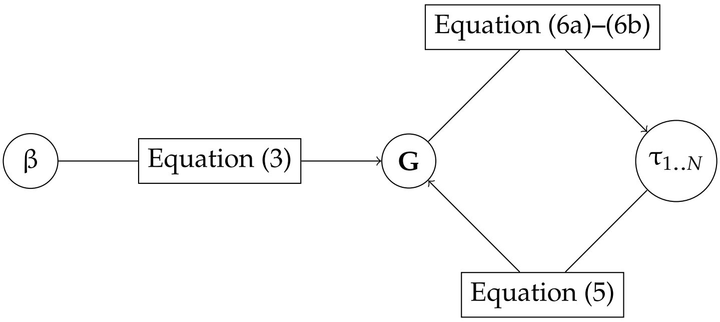


Figure 4
Process diagram of Gibbs sampler for unbounded grammars.

4. Bounded Statistical Grammar Induction Model

Experiments described in this article also evaluate the effect of constraints related to center-embedding depth on grammar induction. These constraints are implemented by defining a depth- and side-specific category set $C_D = \{1..D\} \times \{1, 2\} \times C$ for the constrained grammar, with indicators of depth $d \in \{1..D\}$ and side $s \in \{1, 2\}$ (where side 1 indicates a left sibling and side 2 indicates a right sibling). A bounded grammar $\mathbf{G}_D$ with this category set is then tiled together from depth- and side-specific grammars $\mathbf{G}_{d,s}$ using matrices $\mathbf{D}_{d,s}$ and $\mathbf{E}_{d,s}$ to map categories of parents and combinations of children to their depth- and side-specific counterparts in C_D:

$$\mathbf{G}_D = \sum_{d \in \{1..D\}} \sum_{s \in \{1,2\}} \mathbf{D}_{d,s}\, \mathbf{G}_{d,s}\, \mathbf{E}_{d,s}{}^{\top} \tag{8}$$

This depth-bounded grammar $\mathbf{G}_D$ is substituted for $\mathbf{G}$ in Equations (6a) and (6b) in a depth-bounded version of the Gibbs sampler. An example depth- and side-specific grammar matrix $\mathbf{G}_2$, based on the grammar in Figure 2 is shown in Figure 5.

Center embedding, which requires an additional embedding depth in a left-corner parser, is defined to occur at left children of right children, so left-sibling categories at depth d are defined to expand to left and right children at depth d, and right-sibling categories at depth d are defined to expand to a left child at depth $d + 1$ and a right child at depth d. The depth- and side-specific mapping matrices $\mathbf{D}_{d,s}$ and $\mathbf{E}_{d,s}$ therefore respect this division:

$$\mathbf{D}_{d,s} = \delta_d \otimes \delta_s \otimes \mathbf{I} \tag{9a}$$

$$\mathbf{E}_{d,1} = \delta_d \otimes \delta_1 \otimes \mathbf{I} \otimes \delta_d \otimes \delta_2 \otimes \mathbf{I} \tag{9b}$$

$$\mathbf{E}_{d,2} = \delta_{d+1} \otimes \delta_1 \otimes \mathbf{I} \otimes \delta_d \otimes \delta_2 \otimes \mathbf{I} \tag{9c}$$

where $\mathbf{I}$ is the identity matrix.

Using this definition of center embedding, when a depth constraint of D is applied, it excludes (eliminates the probability of) trees with non-terminal left siblings at depths deeper than D from the generative model's distribution over trees, so this distribution must be renormalized to account for these missing trees in a consistent probability distribution. The depth- and side-specific grammars $\mathbf{G}_{d,s}$ are therefore defined to reweight

$$\begin{array}{r cccccccccccc}
 & \text{NP}_{1,1}\,\text{NP}_{1,2} & \text{NP}_{1,1}\,\text{PP}_{1,2} & \text{NP}_{1,1}\,\text{P}_{1,2} & \dots & \text{NP}_{2,1}\,\text{NP}_{1,2} & \text{NP}_{2,1}\,\text{PP}_{1,2} & \text{NP}_{2,1}\,\text{P}_{1,2} & \dots & \text{P}_{2,2}\,\text{P}_{2,2} & \text{books}\ \bot & \text{stories}\ \bot & \text{about}\ \bot \\
\text{NP}_{1,1} & .09 & .38 & 0 & & 0 & 0 & 0 & & 0 & .3 & .2 & 0 \\
\text{PP}_{1,1} & 0 & 0 & 0 & & 0 & 0 & 0 & & 0 & 0 & 0 & 0 \\
\text{P}_{1,1} & 0 & 0 & 0 & & 0 & 0 & 0 & & 0 & 0 & 0 & 1 \\
\vdots & & & & \ddots & & & & \ddots & & & & \\
\text{NP}_{1,2} & 0 & 0 & 0 & & .07 & .34 & 0 & & 0 & .31 & .21 & 0 \\
\text{PP}_{1,2} & 0 & 0 & 0 & & 0 & 0 & 0 & & 0 & 0 & 0 & 0 \\
\text{P}_{1,2} & 0 & 0 & 0 & & 0 & 0 & 0 & & 0 & 0 & 0 & 1 \\
\vdots & & & & \ddots & & & & \ddots & & & & \\
\text{P}_{2,2} & 0 & 0 & 0 & & 0 & 0 & 0 & & 0 & 0 & 0 & 1
\end{array}$$

Figure 5
Example depth- and side-specific grammar matrix $\mathbf{G}_2$, based on the grammar in Figure 2.

and renormalize the original grammar $\mathbf{G}$ by a containment likelihood $\mathbf{h}^{(I)}_{d,s}$, which is a vector with one element for each category in C containing the probability of that category generating a complete yield within depth d as an s-side sibling:

$$\mathbf{G}_{d,1} = \frac{1}{\mathbf{h}^{(I)}_{d,1}} \,\mathbf{G}\,\text{diag}\begin{bmatrix} \mathbf{h}^{(I)}_{d,1} \otimes \mathbf{h}^{(I)}_{d,2} \\ \mathbf{1} \end{bmatrix} \tag{10a}$$

$$\mathbf{G}_{d,2} = \frac{1}{\mathbf{h}^{(I)}_{d,2}} \,\mathbf{G}\,\text{diag}\begin{bmatrix} \mathbf{h}^{(I)}_{d+1,1} \otimes \mathbf{h}^{(I)}_{d,2} \\ \mathbf{1} \end{bmatrix} \tag{10b}$$

Following van Schijndel, Exley, and Schuler (2013) and Jin et al. (2018b), the containment likelihood $\mathbf{h}^{(I)}_{d,s}$ is estimated iteratively over paths of length $i \in \{0..I\}$ as the probability of a randomly generated tree of height i with each category as its root fitting within center-embedding depth d:

$$\mathbf{h}^{(0)}_{d,s} = \mathbf{0} \tag{11a}$$

$$\mathbf{h}^{(i)}_{d,1} = [\![d \leq D+1]\!]\ \mathbf{G} \begin{bmatrix} \mathbf{h}^{(i-1)}_{d,1} \otimes \mathbf{h}^{(i-1)}_{d,2} \\ \mathbf{1} \end{bmatrix} \tag{11b}$$

$$\mathbf{h}^{(i)}_{d,2} = [\![d \leq D]\!]\ \mathbf{G} \begin{bmatrix} \mathbf{h}^{(i-1)}_{d+1,1} \otimes \mathbf{h}^{(i-1)}_{d,2} \\ \mathbf{1} \end{bmatrix} \tag{11c}$$

Following previous work, experiments described in this paper use $I = 20$.

Equations (8)–(11c) define a depth- and side-specific grammar $\mathbf{G}_D$ from a depth- and side-independent grammar $\mathbf{G}$, which is used in place of $\mathbf{G}$ in Equations (6a)–(6b) to generate depth- and side-specific trees $\tau_{1..N}$. A depth- and side-independent grammar **G** can then be sampled by aggregating over depth- and side-specific rule frequencies-$\mathbf{F}_D$ in these trees, to complete the cycle:

$$\mathbf{G} \sim \text{Dirichlet}\left(\beta + \sum_{d}\sum_{s} \mathbf{D}_{d,s}{}^{\top}\mathbf{F}_D\,\mathbf{E}_{d,s}\right) \tag{12}$$

These depth- and side-specific frequencies are calculated from sampled trees as in Equation (5):

$$\mathbf{F}_D = \sum_{\tau\in\tau_{1..N}} \sum_{\tau_\eta\in\tau} \delta_{\tau_\eta}\,(\delta_{\tau_{\eta 1}} \otimes \delta_{\tau_{\eta 2}})^{\top} \tag{13}$$

Figure 6 shows the complete Gibbs sampling process for bounded grammars.

5. Labeled Parsing Evaluation

Experiments described in this article evaluate the accuracy of parse trees $\tau_{1..N}$ hypothesized by induced grammars against attested trees $\tilde{\tau}_{1..N}$ in annotated corpora. In these evaluations it is straightforward to match the yields of constituents in hypothesized trees (the sequences of words from the first word position i to the last word position j of each constituent) against those of constituents in attested trees, but comparisons of category labels are complicated by the fact that labels $\tilde{\tau}_{n,i,j}$ and $\tau_{n,i,j}$ of attested and hypothesized constituents are drawn from different sets: The former from symbols like 'S' and 'NP,' and the latter from integers $1..|C|$. Fortunately, an induced grammar can still be considered successful to the degree that it produces trees whose constituents match in yield and whose labels predict the attested labels of constituents with corresponding yields (for example, if 1 usually corresponds to 'S,' 2 usually corresponds to 'NP,' etc.). This predictability can be quantified as category *homogeneity*, which is the relative increase in the log of the expected probability of the attested categories in

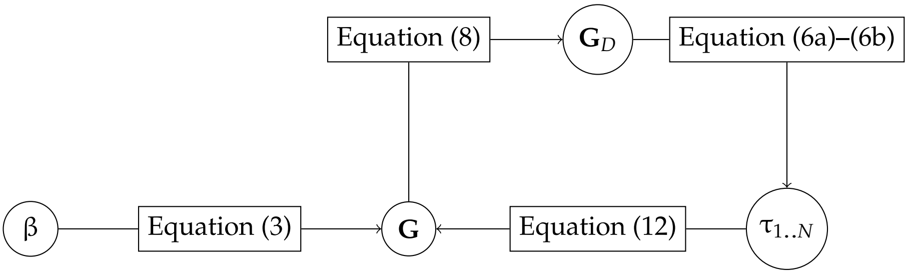


Figure 6
Process diagram of Gibbs sampler for bounded grammars.

the corpus due to conditioning each attested category $\tilde{\tau}_{n,i,j}$ on the category $\tau_{n,i,j}$ of the constituent with the same yield in the hypothesized tree:[3]

$$\mathsf{Hom}(\tilde{\tau}_{1..N}, \tau_{1..N}) = 1 - \frac{\sum_{\tilde{c}\in\tilde{C}} \sum_{c\in C} \mathsf{P}(\tilde{c}, c) \log \mathsf{P}(\tilde{c} \mid c)}{\sum_{\tilde{c}\in\tilde{C}} \mathsf{P}(\tilde{c}) \log \mathsf{P}(\tilde{c})} \tag{14a}$$

$$= 1 - \frac{\sum_{n\in 1..N} \sum_{i,j \text{ s.t. } \tilde{\tau}_{n,i,j}\in\tilde{C}, \tau_{n,i,j}\in C} \log \mathsf{P}(\tilde{\tau}_{n,i,j} \mid \tau_{n,i,j})}{\sum_{n\in 1..N} \sum_{i,j \text{ s.t. } \tilde{\tau}_{n,i,j}\in\tilde{C}, \tau_{n,i,j}\in C} \log \mathsf{P}(\tilde{\tau}_{n,i,j})} \tag{14b}$$

where $\mathsf{P}(\tilde{c}, c) = \sum_n \sum_{i,j} [\![\tilde{\tau}_{n,i,j} = \tilde{c}, \tau_{n,i,j} = c]\!] / \sum_n \sum_{i,j} [\![\tilde{\tau}_{n,i,j} \in \tilde{C}, \tau_{n,i,j} \in C]\!]$ is the probability of a span that is a constituent in both $\tau_{1..N}$ and $\tilde{\tau}_{1..N}$ being attested with category $\tilde{c}$ and assigned category c by the induced grammar, and $\mathsf{P}(\tilde{c}) = \sum_n \sum_{i,j} [\![\tilde{\tau}_{n,i,j} = \tilde{c}, \tau_{n,i,j} \in C]\!] / \sum_n \sum_{i,j} [\![\tilde{\tau}_{n,i,j} \in \tilde{C}, \tau_{n,i,j} \in C]\!]$ is the probability of a span that is a constituent in both $\tau_{1..N}$ and $\tilde{\tau}_{1..N}$ being attested with category $\tilde{c}$ but assigned any category by the induced grammar. This category homogeneity can then be weighted by unlabeled constituent *recall*, which is the fraction of constituents in attested trees that also appear in the hypothesized tree, to give a *recall homogeneity* (RH) measure:

$$\mathsf{RH}(\tilde{\tau}_{1..N}, \tau_{1..N}) = \frac{\sum_{n\in 1..N} \sum_{i,j} [\![\tau_{n,i,j} \in C, \tilde{\tau}_{n,i,j} \in \tilde{C}]\!]}{\sum_{n\in 1..N} \sum_{i,j} [\![\tilde{\tau}_{n,i,j} \in \tilde{C}]\!]} \cdot \mathsf{Hom}(\tilde{\tau}_{1..N}, \tau_{1..N}) \tag{15}$$

Both recall and the subtracted term of homogeneity can be allocated to individual sentences (using the term inside the sum $\sum_n$ over sentences) for significance testing via permutation sampling.

Note that this use of recall and homogeneity is distinct from commonly used F-score and V-measure for hypothesized constituents and category labels, respectively. F-score is the harmonic mean of recall and precision (which has the same form as recall but with τ and $\tilde{\tau}$ reversed), and V-measure is the harmonic mean of homogeneity and completeness (which has the same form as homogeneity but with τ and $\tilde{\tau}$ reversed). These aggregated measures are usually used as checks on evaluated models that can generate unlimited numbers of hypotheses or hypotheses of unlimited granularity. However, in the present application, hypothesized constituents in parse trees are limited by the number of words in each sentence, and hypothesized category labels are limited to a constant set of categories of size C, so checks on the number and granularity of hypotheses are not necessary. Moreover, the use of recall rather than F-score in these evaluations assumes the decision to suppress annotation of constituents to make flatter trees is motivated by expediency on the part of the annotators, rather than linguistic theory, so extra constituents in binary-branching trees that are not present in attested trees are not counted against induced grammars unless they interfere with the recall of other attested constituents. Likewise, the use of homogeneity rather than V-measure in these evaluations assumes the decision to suppress annotation of information about case or subcategorization information in category labels is motivated by expediency rather than linguistic theory, so the use of categories to make such additional distinctions is not counted against induced grammars unless it interferes with the homogeneity of predictions of other attested categories from hypothesized categories.

[3] Here, $\tau_{n,i,j} \notin C$ if no constituent in τ yields words i to j, and $\tilde{\tau}_{n,i,j} \notin \tilde{C}$ if no constituent in $\tilde{\tau}$ yields words i to j.

Experiments in Section 7.2 show that even without completeness as a check on the size of the category label set, results peak at $C = 45$ and decline thereafter.

Notwithstanding this use of RH in tuning and internal evaluations, comparisons of models proposed in this article to other existing models do use F-score, in order to ensure a fair comparison using the same measure to which these other models have been optimized.

Significance testing with the RH measure adopts the conventional permutation testing in supervised parsing, where trees from two induced grammars are randomly permuted in order to calculate the probability of the difference between the two candidate grammars in terms of the chosen evaluation metric. Scores for permuted samples are calculated by summing per-sentence-recall (PSR) and per-sentence-heterogeneity (the unit complement of homogeneity; PSH) scores for each sentence, then subtracting the summed heterogeneity from one to get homogeneity, and multiplying by recall to get RH:

$$\mathsf{RH}(\tilde{\tau}_{1..N}, \tau_{1..N}) = \left(\sum_{n \in 1..N} \mathsf{PSR}(n, \tilde{\tau}_{1..N}, \tau_{1..N}) \right) \cdot \left(1 - \sum_{n \in 1..N} \mathsf{PSH}(n, \tilde{\tau}_{1..N}, \tau_{1..N}) \right) \tag{16}$$

Per-sentence recall and heterogeneity are then calculated by pulling out these summations from the fractional terms in Equations (15) and (14b), respectively.

$$\mathsf{PSR}(n, \tilde{\tau}_{1..N}, \tau_{1..N}) = \frac{\sum_{i,j} [\![\tau_{n,i,j} \in C, \tilde{\tau}_{n,i,j} \in \tilde{C}]\!]}{\sum_{n \in 1..N} \sum_{i,j} [\![\tilde{\tau}_{n,i,j} \in \tilde{C}]\!]} \tag{17}$$

$$\mathsf{PSH}(n, \tilde{\tau}_{1..N}, \tau_{1..N}) = \frac{\sum_{i,j \text{ s.t. } \tilde{\tau}_{n,i,j} \in \tilde{C}, \tau_{n,i,j} \in C} \log \mathsf{P}(\tilde{\tau}_{n,i,j} \mid \tau_{n,i,j})}{\sum_{n \in 1..N} \sum_{i,j \text{ s.t. } \tilde{\tau}_{n,i,j} \in \tilde{C}, \tau_{n,i,j} \in C} \log \mathsf{P}(\tilde{\tau}_{n,i,j})} \tag{18}$$

6. Experiment 1: Evaluation of Unbounded PCFG Induction on Synthetic Data

The unbounded model described in Section 3 is evaluated first on synthetic data (Jin et al. 2018b) to determine whether it can reliably learn a recursive grammar from data with a known optimum solution. The symmetric concentration hyper-parameter β is set to be 0.2, following Jin et al. (2018b). The corpus consists of 50 sentences each of the form *a b c*; *a b b c*; *a b a b c*; and *a b b a b b c*, which has optimal tree structures as shown in Figure 7.[4] The (b) and (d) trees require the system to hypothesize depth 2 structures. The system was able to recall all optimal tree structures with an equivalent category allocation.

The accuracy of the unbounded model was also compared against that of existing induction models by Seginer (2007a), [5] Ponvert, Baldridge, and Erk (2011), [6] Shain et al. (2016), [7] as well as [Kim, Dyer, and Rush 2019]. [8] The two models from Kim, Dyer, and Rush (2019) differ in that the model with z induces sentence-specific grammars, and

[4] The tokens a, b, and c are randomly chosen uniformly from $\{a_1, \ldots, a_{50}\}$, $\{b_1, \ldots, b_{50}\}$ and $\{c_1, \ldots, c_{50}\}$, respectively.

[5] `https://github.com/DrDub/cclparser`.

[6] `https://github.com/eponvert/upparse`.

[7] `https://github.com/tmills/uhhmm/tree/coling16`.

[8] `https://github.com/harvardnlp/compound-pcfg`.

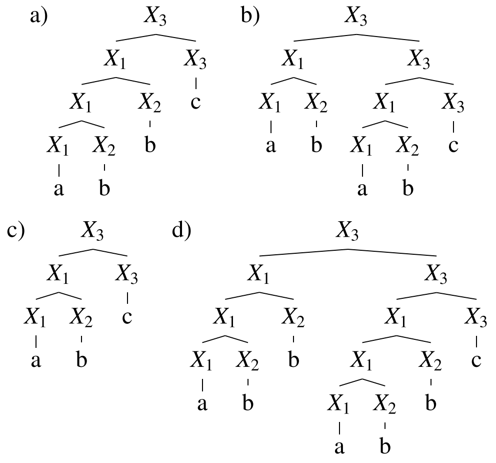


Figure 7
Synthetic center-embedding structure. Note that tree structures (b) and (d) have depth 2 because they have complex sub-trees spanning *a b* and *a b b*, respectively, embedded in the center of the yield of their roots.

the model without z induces one grammar for all sentences. The results are shown in Table 1. No other system was able to recall all optimal tree structures.

7. Experiment 2: Evaluation of *Unbounded* PCFG Induction on Child-Directed Speech

Observing that the model is able to correctly identify known grammars from data, we then evaluate the unbounded PCFG inducer on a corpus of child-directed speech from the Adam and Eve sections of the Brown corpus (Brown 1973) of CHILDES (Macwhinney 1992). The Adam data set consists of transcripts of interactions between Adam and his caregivers recorded at ages ranging from 2 years 3 months to 5 years 2 months. Eve is similar, with interactions recorded between age 1 year 6 months and 2 years 3 months. Penn Treebank–style syntactic annotation for the child-directed utterances is provided by Pearl and Sprouse (2013) using an automatic parser (Charniak and Johnson 2005) and human annotators. There are 28,779 sentences in the annotated

Table 1
The oracle best accuracy scores of unlabeled parse evaluation of different systems on synthetic data.

System	Recall	Precision	F1	RH
Seginer (2007a)	0.71	0.83	0.77	–
Ponvert, Baldridge, and Erk (2011)	0.81	0.91	0.86	–
Shain et al. (2016)	0.38	0.38	0.38	–
Kim, Dyer, and Rush (2019) without z	0.73	0.73	0.73	0.73
Kim, Dyer, and Rush (2019) with z	0.73	0.73	0.73	0.73
Unbounded PCFG §3	**1.00**	**1.00**	**1.00**	**1.00**

Adam corpus, with average sentence length of 6 words. There are 67 unique syntactic categories used in the data set. *N*-ary branching is not binarized in the human annotation, but unary branching chains are collapsed and the topmost category in the chain is used as the category for the constituent. The Eve section has 14,251 sentences, with 64 unique syntactic categories, and the average sentence length is 5.6 words. The number of unique phrasal categories after unary chain collapse is 25 and 21, respectively.

Hyperparameters β and *C* are set to optimize accuracy on the Adam section. Several analyses are performed using grammars and trees induced using Adam. Finally held-out evaluation is performed on the Eve section.

Following previous work, these experiments leave all punctuation in the input for learning as a proxy for prosodic cues about phrasal boundaries (Seginer 2007b). Punctuation is then removed in all evaluations on development and test data. All results reported for each condition include induced grammars and trees from running the system with 10 random seeds. Each run contains 700 sampling cycles, and the final sampled grammar is used to generate the final parses of the corpus. Accuracy is evaluated by comparing optimal (Viterbi) parses instead of sampled parses. These evaluated parses are strictly binary-branching trees, although annotations may contain flatter *n*-ary trees. Results include all runs for each condition, shown in plots as boxes with boundaries at the first and the third quartiles, with medians as green lines inside, and with upper and lower whiskers showing the minimum and maximum of each set of data points. Circles are used for outliers, which are data points with values more extreme than 1.5 times of the interquartile range, the distance between the first and the third quartile.

7.1 Optimization of Concentration Parameter on Exploratory Partition

In Bayesian induction, the Dirichlet concentration hyperparameter β controls the probability of a sampled multinomial distribution, with high values yielding more uniform distributions over expansion rules in the grammar and with low values concentrating the probability mass on only a few expansions. Figure 8 shows RH scores for runs with

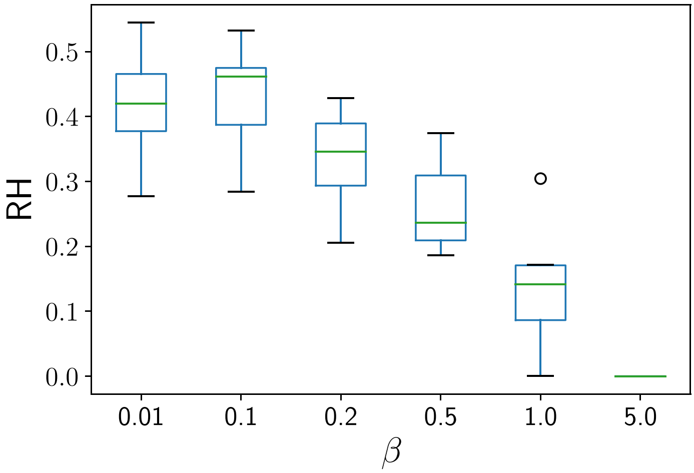


Figure 8
RH scores for various β values on exploratory partition (Adam).

different β values on Adam with the number of syntactic categories $C = 30$. Results show a peak at $\beta = 0.1$, indicating a preference for sparse, highly concentrated probabilities over a few expansion rules.

Indeed, human grammars are generally sparse in this way (Johnson, Griffiths, and Goldwater 2007; Goldwater and Griffiths 2007). For example, in the Penn Treebank (Marcus, Santorini, and Marcinkiewicz 1993), there are 73 unique nonterminal categories. In theory, there can be more than 28 million possible unary, binary, and trinary branching rules in the grammar. However, there are only 17,020 unique rules found in the corpus, showing the high sparsity of attested rules in the grammar. In other frameworks like CCG (Steedman 2002), where lexical categories can be in the thousands, the number of attested lexical categories is still small compared to all possible lexical categories.

The sparsity also shows up in POS assignments of words. Usually the number of POS tags a word can have is very small. For words with low-frequency and hapax legomena, β has a particularly strong influence on their posterior uniformity of POS assignment, with natural language grammars clearly preferring low uniformity.

Constituency grammar induction is often measured using F1 scores over unlabeled spans (Seginer 2007a; Ponvert, Baldridge, and Erk 2011, inter alia). Figure 9 shows unlabeled F1 scores with different β values on Adam. Contrary to the prediction, grammar accuracy peaks at high values for β when measured using unlabeled F1. However, upon close inspection, these grammars with high unlabeled F1 are almost purely right-branching grammars, which does indeed perform very well on English child-directed speech in unlabeled parsing evaluation, but the right-branching grammars have phrasal labels that do not correlate with human annotation when evaluated with RH. This indicates that instead of capturing human intuitions about syntactic structure, such grammars have only captured broad branching tendencies.

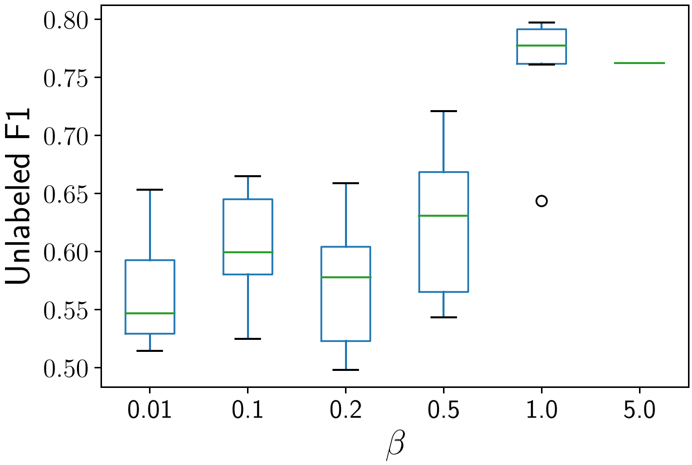


Figure 9
Unlabeled F1 scores for various β values on exploratory partition (Adam).

7.2 Optimization of Category Domain Size on Exploratory Partition

Previous work on PCFG induction usually has used fewer than 20 syntactic categories (Shain et al. 2016; Jin et al. 2018b). This number is substantially smaller than the number of categories in human annotations, but it may be expected because there may not be enough statistical clues in the data for the inducer to distinguish some categories from other ones. For example, determiners and cardinal numbers may appear to be very similar distributionally, because they usually occur before bare nouns and bare noun phrases. However, the labeled evaluation with the sparsity parameter in Section 7.1 indicates that unlabeled evaluation is not informative enough about the accuracy of induced grammars, as it includes no measure of accuracy for induced constituent labels. Figure 10 shows induction results on the Adam data set with several category domain sizes given this optimal value of β. Results show a peak of RH at $C = 45$. This suggests that the inducer may have insufficient categories to use at lower domain sizes, yielding much lower RH values at $C = 15$. The accuracy at $C = 45$ is a well-formed peak with the smallest variance among all experimental settings, but there is a secondary peak at $C = 75$, with some induced grammars as accurate as induced grammars with 45 categories. This may indicate some statistical evidence in the data for further subcategorization of the grammars with 45 categories, but such evidence may not be strong enough to reduce posterior multimodality.

7.3 Correlation of Model Fit and Parsing Accuracy

Model fit, or data likelihood, has been reported not to be correlated or to be correlated only weakly with parsing accuracy for some unsupervised grammar induction models when the model has converged to a local maximum (Smith 2006; Johnson, Griffiths, and Goldwater 2007; Liang et al. 2007). Figure 11 shows the correlation between data likelihood and RH at convergence for all 70 runs with $\beta = 0.1$. There is a significant ($p < 0.001$) positive correlation (Pearson's $r = 0.737$) between data likelihood and RH at convergence for our model. This indicates that although noisy and unreliable, the

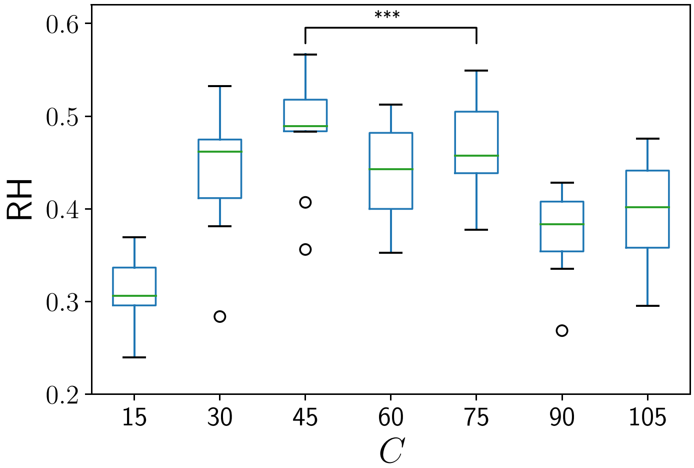


Figure 10
RH scores for various C values and $\beta = 0.1$ on exploratory partition (Adam) (***: $p < 0.001$).

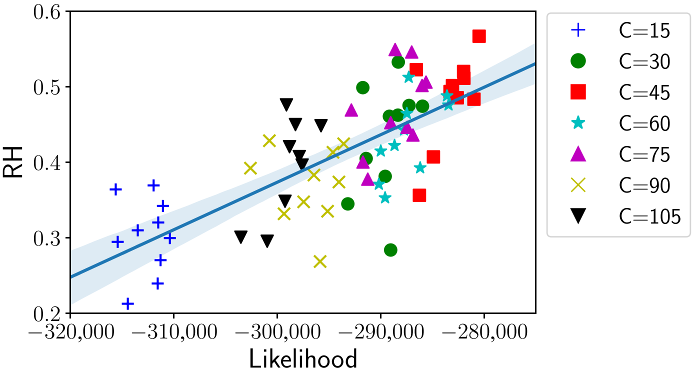


Figure 11
The correlation between likelihood and RH on Adam over various C values for $\beta = 0.1$.

data likelihood can be used as a metric to do preliminary model selection. The figure also shows that the distribution of likelihoods from various C values also indicates the correlation between likelihood and model performance, with most of the induced grammars with high performing C values such as 45 or 75 in the region of the highest likelihoods, and most of the low performing C values such as 15 or 90 in the region of the lowest likelihoods. The difference between this significant correlation of parsing accuracy and data likelihood and previous results of weak or no correlation may be due to the use here of labeled (RH) accuracy as a more natural measure of parsing accuracy than unlabeled (F1) accuracy. It may also be due to the simpler language used in Adam compared to that of newswire data sets used in previous work. Finally, the discrepancy may be due to the use of Expectation Maximization in previous work, which may overfit a grammar to a data set, and could give unrealistically high likelihood to grammars that are too specific for a particular set of sentences.

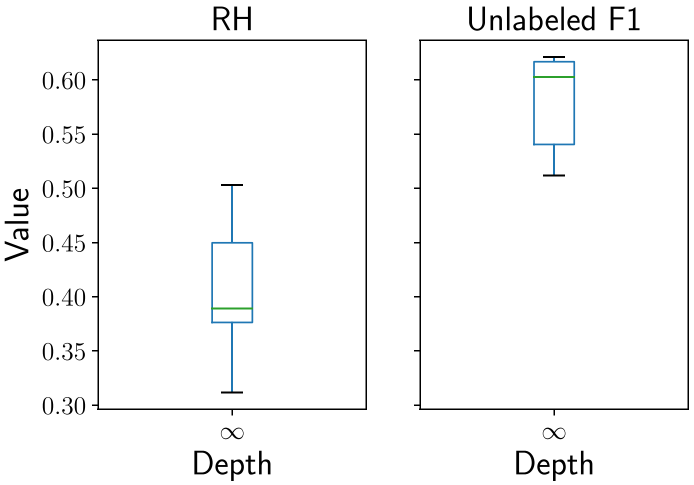


Figure 12
Unbounded induction experiment on the held-out partition (Eve) with $\beta = 0.1$ and $C = 45$.

Table 2
PARSEVAL scores on Eve data set with previously published induction systems.

System	F1	RH
Seginer (2007a)	0.52	–
Ponvert, Baldridge, and Erk (2011)	0.56	–
Shain et al. (2016)	0.66	–
Kim, Dyer, and Rush (2019) without z	0.51	**0**.44
Kim, Dyer, and Rush (2019) with z	0.31	0.39
this work (D=∞, C=45)	0.62	**0**.44
Right-branching	**0.76**	0.00

7.4 Results for Unbounded Induction on Held-Out Partition

With the hyperparameters tuned on Adam, experiments are run on the held-out section of Eve. Results are shown in Figure 12. The median unlabeled F1 score is around 0.6, and the median RH score is 0.38. The RH of the highest-likelihood run is 0.44. Table 2 shows the unlabeled F1 and labeled RH scores for published systems, using the induced grammar for this work from the run with the highest likelihood on the whole section. The inducer optimized for RH still achieves good unlabeled parsing accuracy, although the unlabeled F1 score is still lower than that of a purely right-branching baseline. Figure 9 shows that $\beta = 1.0$ does help induce grammars that are mostly right-branching but still retain some linguistically meaningful constituents, which push the unlabeled F1 score above the right-branching baseline accuracy at 0.75 on the Adam section. It is reasonable to assume that using $\beta = 1.0$ on Eve will also achieve the same result. However, the deterioration of the quality of constituent labeling at high βs makes optimizing for unlabeled F1 much less attractive. For some of the published systems, there is no way to produce labeled trees, therefore the labeled evaluation is not applicable to them. For the right-branching baseline, because there is no trivial and automatic way to assign different category labels to constituents, its RH score is 0.0.

7.5 Analysis of Learned Syntactic Categories and Grammatical Rules

We are interested in examining the learned categories and rules and compare them to annotation. Many of the most common induced rules look linguistically sensible. The twenty most frequent rules generated by the run of the unbounded inducer with the highest likelihood probability using optimal $\beta = 0.1$ and $C = 45$ parameters on the Adam data set are shown in Table 3. Each rule is followed by the most common attested rule for the same decomposition (on the upper line), and some randomly sampled examples (on the lower line, with a vertical bar showing the split point between left and right child spans).[9] The recall homogeneity for this run is 0.57. Of these twenty most frequent rules, only six (the first, seventh, eighth, ninth, nineteenth, and twentieth) do not seem to correspond to any linguistically recognizable syntactic analysis. Those that do are the following:

[9] Question marks ('??') for parent, left child, or right child indicate no constituent was attested at that location.

Table 3
The most frequent rules induced in Adam ($\beta = 0.1, C = 45$) and their correspondences in the attested trees. The examples are randomly sampled from the induced trees. Question marks ('??') for parent, left child, or right child indicate no constituent was attested at that location.

Rank Rule	Corresponding gold rules, counts and examples
1. 4 → 37 30	?? → NP COP (0.53); ?? → NP AUX (0.09); ?? → NP VBZ (0.08); ?? → VB NP (0.05)
(5,229)	dirt | is ; he | 's ; it | 's ; he | 's
2. 24 → 5 25	?? → ?? ?? (0.57); VP → MD VP (0.18); ?? → MD ?? (0.07); ?? → ?? VB (0.04)
(4,613)	will n't | step on your candy ; do n't | know ; 'm | afraid you 'll forget ; do n't | want to play
3. 42 → 33 25	?? → NP VP (0.53); ?? → NP ?? (0.25); S → NP VP (0.10); ?? → NP VB (0.06)
(4,294)	you | do ; you | show him ; you | do in the kitchen ; these | things to ride on
4. 36 → 11 8	?? → VB NP (0.53); VP → VB NP (0.13); ?? → VB ?? (0.08); ?? → VBP NP (0.05)
(4,155)	ask | ursula ; have | a bump ; put | the pillows ; see | that
5. 23 → 38 27	ROOT → WHNP SQ (0.47); ROOT → WHADVP SQ (0.19); ?? → ?? ?? (0.04)
(4,126)	that 's a train part | is n't it ; who | 's there ; what | for ; you got your fingers in it | did n't you
6. 25 → 36 13	?? → ?? ?? (0.49); VP → VB PP (0.10); ?? → VP ?? (0.05); VP → VB ADVP (0.03)
(3,642)	eat yourself | up ; do | when you go to school ; draw | on it ; play | with the record
7. 23 → 34 17	?? → ?? ?? (0.71); ?? → ?? PP (0.06); ?? → ?? SBAR (0.04); ?? → S ?? (0.04)
(3,514)	paul stay away | away away from there ; no i do n't know | what delfc means
8. 34 → 4 43	?? → ?? ?? (0.57); ?? → ?? NP (0.17); ?? → ?? VBG (0.04); ?? → ?? JJ (0.03)
(3,227)	they 're | in your box ; they 're just | playing ; those are | stamps you use ; it says | here
9. 23 → 4 43	?? → ?? ?? (0.85); ?? → ?? NP (0.02); ?? → ?? VP (0.01); ROOT → VB NP (0.01)
(3,087)	just like | adam ; you told | the carpenter you had a big burp ; they are | taking baths
10.23 → 6 34	ROOT → INTJ S (0.28); ?? → ?? ?? (0.19); ?? → INTJ ?? (0.19); ROOT → INTJ FRAG (0.06)
(3,005)	because | you 'll break it ; because | you 're still there ; oh | hurry up
11. 8 → 0 32	NP → DT NN (0.50); ?? → ?? ?? (0.14); NP → PRP$ NN (0.10); NP → DT NNS (0.07)
(2,931)	any | noise ; the little | boy ; any | more ; the | policeman
12. 7 → 0 10	NP → DT NN (0.55); NP → PRP$ NN (0.17); ?? → ?? ?? (0.10); ?? → DT NN (0.03)
(2,899)	your | wrist ; our | rug ; the | toy ; the other | side
13.43 → 0 32	NP → DT NN (0.45); ?? → ?? ?? (0.23); NP → PRP$ NN (0.07); ?? → DT NN (0.06)
(2,776)	any | more ; cowboy | hat ; morning or | afternoon ; a | lobster
14.27 → 35 42	?? → ?? ?? (0.64); ?? → AUX ?? (0.24); ?? → MD ?? (0.06); SQ → COP NP (0.02)
(2,631)	are | you going to do ; do | n't you tell ursula what you have ; did | you hurt yourself
15.25 → 11 8	VP → VB NP (0.60); ?? → VB NP (0.07); VP → VBP NP (0.05); ?? → VB ?? (0.05)
(2,461)	close | it ; want | some more paper ; seen | everything ; like | it
16. 5 → 35 31	?? → AUX NOT (0.80); ?? → MD NOT (0.17); ?? → COP NOT (0.01); VP → AUX NOT (0.01)
(2,387)	do | n't ; did | n't ; do | n't ; does | n't
17.27 → 30 37	SQ → COP NP (0.51); ?? → COP NP (0.24); ?? → AUX NP (0.05); ?? → COP ?? (0.03)
(2,278)	about | the treasure house ; is | it ; is | that ; is | it
18.13 → 40 7	PP → IN NP (0.67); ?? → IN ?? (0.08); ADVP → RB RB (0.07); ?? → IN NP (0.04)
(2,236)	in | yours ; of | those ; around | here ; at | paul
19.23 → 12 9	?? → ?? ?? (0.78); ?? → SBARQ ?? (0.05); ?? → SQ ?? (0.03); ?? → ?? PP (0.03)
(2,228)	is she dancing | on the horse 's back ; what kind | of paper ; what happens | when you press it
20. 0 → 0 1	?? → DT JJ (0.56); ?? → DT NN (0.15); ?? → ?? JJ (0.04); ?? → PRP$ JJ (0.03)
(1,844)	a | nice ; a | dozen ; one | half ; a | few

- The second most frequent and the sixteenth most frequent rules seem to simply undo the programmatic tokenization of contractions of modals and negation adverbs (e.g., *is — n't*) which is common in Penn Treebank annotations (Marcus, Santorini, and Marcinkiewicz 1993).
- The third rule, the fourth and fifteenth rules, and the eighteenth rule fairly selectively attach subjects to verb phrases, direct objects to transitive verbs, and complements to prepositions, respectively.
- The fifth rule decomposes content questions into question words followed by sentences containing gaps (but also conflates these with sentences followed by echo questions).

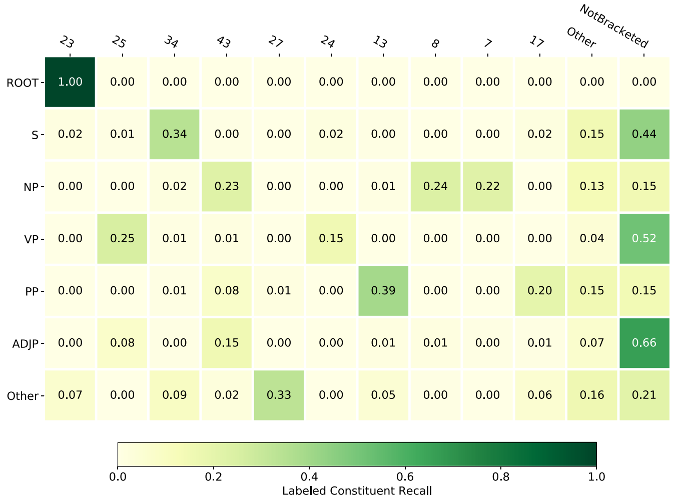
- The sixth most common rule right-adjoins particles and adverbial modifiers onto verb phrases.
- The tenth rule left-adjoins interjections onto sentences.
- Rules eleven through thirteen decompose noun phrases into determiners followed by common nouns. It is also interesting to note that the model reliably distinguishes subjects (category 33 in this run) from direct objects, (category 8) and complements of prepositions (category 7).[10] This suggests that case systems, which treat subjects, direct objects, and oblique objects as different categories, might naturally arise from distributions of words in sentences, rather than from a biological bias. Rules eleven and thirteen have the same children categories but different parent categories. Further inspection of the rules using the parent categories shows that these two types of noun phrases are distinguished by whether the main verb needs further complements or adjuncts.
- Rules fourteen and seventeen perform subject-auxiliary inversion by attaching subjects to auxiliaries below the complements. This kind of structure is unusual in movement-based analyses, but is a common feature of categorial grammar analyses because it allows both the subject and the complement to be adjacent to the auxiliary as its arguments.

Figure 13a shows a confusion matrix for this same highest-likelihood run on Adam with $\beta = 0.1, C = 45$, showing percentages of several common attested non-preterminal categories that are hypothesized as one of the ten most common induced categories. Preterminal categories are not included because their boundaries as single words are trivially induced. This run correctly recalled most noun phrases and prepositional phrases, but missed a large proportion of clauses and a majority of verb phrases. Figure 13b shows the same confusion matrix with percentages of hypothesized categories that correspond to each attested category. This shows that the hypothesized categories that correspond to noun phrases, verb phrases and prepositional phrases mostly exclusively represent these categories.

Table 4 shows the 20 most frequent induced syntactic categories at the preterminal positions and the corresponding human annotated POS tags, showing the percentage of each induced category attested with each tag. Attested POS tags that have fewer than 100 word tokens or fewer than 5% of the induced category instances are not included in the table. This time all but two induced categories (the eighteenth and twentieth) seem linguistically meaningful:

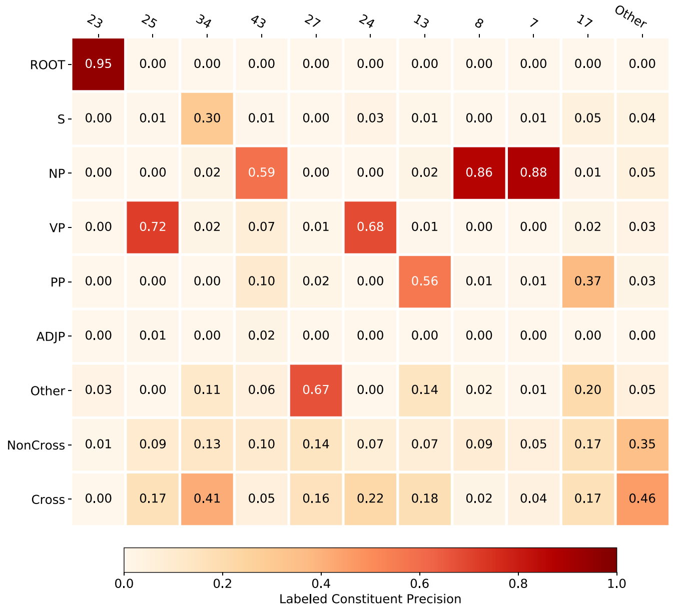
- 93% of the most common induced preterminal category correspond to attested determiners or possessive pronouns.
- 99% of the second and twelfth most common induced preterminal categories (categories 33 and 28) and 86% and 89% of the sixth and tenth most common categories (categories 37 and 8) correspond to attested pronouns and other single-word noun phrases. Table 5 lists examples for these four induced categories found in the Viterbi parses. Category 37 is

[10] Category 43 is used for complements of merged contractions. This analysis appears to be consistent, but is not a linguistically familiar analysis.



	23	25	34	43	27	24	13	8	7	17	Other	NotBracketed
ROOT	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
S	0.02	0.01	0.34	0.00	0.00	0.02	0.00	0.00	0.00	0.02	0.15	0.44
NP	0.00	0.00	0.02	0.23	0.00	0.00	0.01	0.24	0.22	0.00	0.13	0.15
VP	0.00	0.25	0.01	0.01	0.00	0.15	0.00	0.00	0.00	0.00	0.04	0.52
PP	0.00	0.00	0.01	0.08	0.01	0.00	0.39	0.00	0.00	0.20	0.15	0.15
ADJP	0.00	0.08	0.00	0.15	0.00	0.00	0.01	0.01	0.00	0.01	0.07	0.66
Other	0.07	0.00	0.09	0.02	0.33	0.00	0.05	0.00	0.00	0.06	0.16	0.21

(a) Recall of syntactic categories on Adam with the run with the highest likelihood.



	23	25	34	43	27	24	13	8	7	17	Other
ROOT	0.95	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
S	0.00	0.01	0.30	0.01	0.00	0.03	0.01	0.00	0.01	0.05	0.04
NP	0.00	0.00	0.02	0.59	0.00	0.00	0.02	0.86	0.88	0.01	0.05
VP	0.00	0.72	0.02	0.07	0.01	0.68	0.01	0.00	0.00	0.02	0.03
PP	0.00	0.00	0.00	0.10	0.02	0.00	0.56	0.01	0.01	0.37	0.03
ADJP	0.00	0.01	0.00	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Other	0.03	0.00	0.11	0.06	0.67	0.00	0.14	0.02	0.01	0.20	0.05
NonCross	0.01	0.09	0.13	0.10	0.14	0.07	0.07	0.09	0.05	0.17	0.35
Cross	0.00	0.17	0.41	0.05	0.16	0.22	0.18	0.02	0.04	0.17	0.46

(b) Precision of syntactic categories on Adam with the run with the highest likelihood.

Figure 13
Categories induced on Adam with $\beta = 0.1$ and $C = 45$.

usually a third-person singular subject (that, it, he), category 33 is almost always a plural or second-person subject (you, they, we), category 8's most common instances are accusative pronouns (it, them, me), and category 28 is mostly the nominative first person pronoun (*I*) occurring in the subject position. Word order information and subject–verb agreement seem to drive this subcategorization, which resembles a combination of case and number. Because the inducer has no sub-word phonological information,

Table 4
Recall of gold POS tags in the top 20 most frequent induced syntactic categories at the preterminal positions. Note that because of unary chain collapse, phrasal tags like NP can appear at preterminal positions.

Rank	Induced category	Category count	Attested category and relative frequency
1.	0	11,327	DT (0.77); PRP$ (0.16)
2.	33	9,983	NP (0.99)
3.	11	8,853	VB (0.71); VBP (0.12); VBD (0.07)
4.	30	8,031	COP (0.64); AUX (0.09); VBZ (0.07); VP (0.05)
5.	32	7,865	NN (0.81); NNS (0.10)
6.	37	7,402	NP (0.86)
7.	35	7,333	AUX (0.72); MD (0.17)
8.	38	6,900	WHNP (0.54); WHADVP (0.23); WP (0.07)
9.	40	6,712	IN (0.80); RB (0.05)
10.	8	6,013	NP (0.89)
11.	6	5,424	INTJ (0.60); ADVP (0.10); NP (0.09); CC (0.05)
12.	28	4,004	NP (0.99)
13.	10	3,880	NN (0.88); NNS (0.08)
14.	43	3,171	ADJP (0.27); NP (0.22); VP (0.13); JJ (0.10); VBG (0.07)
15.	31	3,086	NOT (0.99)
16.	36	3,043	VB (0.60); VP (0.18); VBP (0.05)
17.	13	2,705	PRT (0.32); ADVP (0.32); NP (0.12)
18.	1	2,483	JJ (0.66); NN (0.20)
19.	3	2,388	TO (0.80); IN (0.15)
20.	18	2,220	RB (0.20); NOT (0.19); VBG (0.13); IN (0.11)

Table 5
Recall of top 3 most frequent words in the four induced categories that correspond to noun phrases.

Rank	Induced category	Category count	Attested words and relative frequency
1.	33	9,983	you (0.86); they (0.05); we (0.02)
2.	37	7,402	that (0.38); it (0.27); he (0.07)
3.	8	6,013	it (0.36); them (0.06); me (0.06)
4.	28	4,004	I (0.70); he (0.10); it (0.06)

it relies solely on word order information to distinguish nominative and accusative cases, which is especially important for pronouns like *it* and common nouns in English when the two cases are syncretic.

- 100% of the third most common induced preterminal category (category 11) and 83% of the sixteenth most common category (category 36) correspond to attested verbs. It is also interesting to note that at least half of category 11 appear as the left child in the fourth and fifteenth most common induced rules, generally in a transitive verb context (an attested verb followed by a noun phrase), and many of category 36 appear as the left child in the sixth most common rule, often in a non-transitive verb context (an attested verb followed by a particle or prepositional phrase). This suggests that the inducer distinguishes transitive and intransitive verbs. This is especially interesting because the inducer does not appear to distinguish base, participial, and past-tense forms of verbs, presumably deriving a higher overall posterior probability from subcategorization distinctions.

- There are also common homogenous categories corresponding to auxiliaries (73% of the fourth most common induced preterminal and 89% of the seventh most common), common nouns (91% of the fifth most common induced preterminal and 96% of the thirteenth most common), interrogative pronouns (84% of the eighth most common preterminal), prepositions (80% of the ninth most common preterminal), and others.

8. Experiment 3: Evaluation of Bounded PCFG Induction on Child-Directed Speech

The Adam and Eve sections from the Brown Corpus are then used to evaluate the depth-bounded model defined in Section 4. Transcribed child-directed speech data in Chinese Mandarin (Tong; Deng et al. 2018) and German (Leo; Behrens 2006) are also collected from the CHILDES corpus with reference trees automatically generated using the state-of-the-art Kitaev and Klein (2018) supervised parser trained with the Chinese (Xia et al. 2000; The Chinese Treebank) and German (Skut et al. 1998; NEGRA) treebanks. They are used as held-out data sets for the bounded grammar induction experiments, using cross-linguistic hyperparameters tuned on English. There are 19,541 sentences in the Tong data set being recorded between age 1 year 0 months and 4 years 5 months, with an average sentence length of 5.7 and 55 unique syntactic categories. The Leo data set contains 20,000 child-directed utterances randomly sampled from the original Leo corpus, as the original corpus contains records of interactions between Leo and the caregivers between age 1 year 11 months and 4 years 11 months with high frequency. There are 72 unique syntactic categories in the parsed data set with an average sentence length of 6.7 words. Disfluencies in all corpora are removed, and only sentences spoken by caregivers are kept in the data.

The hyperparameter $\beta = 0.1$ is used for all experiments as it is found to be optimal in experiments described in Section 7. The optimal C was found to be 45 with Adam, but all depth-bounding experiments described here use $C = 30$ because the memory of available graphics processing units is not sufficient to contain depth-specific grammars at higher values of C. All the other settings follow the unbounded experiments: 10 randomly seeded runs for each experimental setting with results reported using box and whisker plots, induction with punctuation and evaluation without punctuation, and labeled evaluation with Viterbi parses.

8.1 Optimization of Depth on Exploratory Partition

The exploratory partition (Adam) is first used to determine an optimal depth bound D. Figure 14 shows the interaction between depth and RH scores on Adam at $\beta = 0.1$ and $C = 30$. The RH peaks at $D = 3$, which is consistent with previous results showing that three levels of nested center-embeddings appear to be the maximum in natural language text in many languages (Karlsson 2007, 2010; Schuler et al. 2010), which is

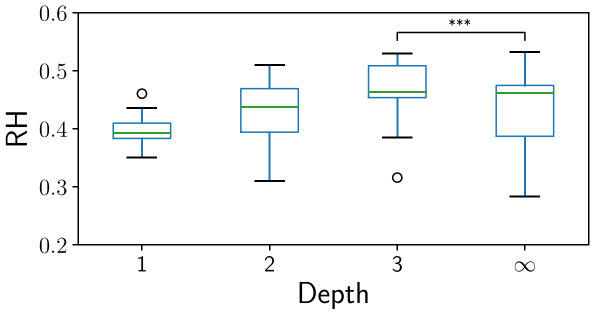


Figure 14
Depth-bounding on Adam with β = 0.1 and $C = 30$.

in turn caused by limited amount of working memory. Induced grammars at $D = 1$ appear to be inadequate for capturing child-directed speech. $D = \infty$ grammars show great variance in accuracy, with induced grammars among the most accurate and least accurate. This shows the value of depth-bounding: The process of depth-bounding acts as an inductive bias, removing possible grammars as posterior modes with low accuracy such that the inducer is more likely to find grammars that are high in data likelihood and also consistent with human memory constraints.

Significance testing with the labeled evaluation metric RH, described in Section 5, is used in all experiments reporting significance levels. The parses from all 10 runs for each experiment condition are concatenated, and random permutations between parses from the two experimental conditions are carried out to calculate the probability of the observed accuracy difference between these two conditions. Results show that the difference between $D = 3$ and $D = \infty$ is highly significant ($p < 0.001$) on Adam, showing that depth-bounding significantly improves the chance of inducing more accurate grammars.

Figure 15 shows two trees from the two runs with $D = \infty$ and $D = 3$ with the highest likelihood on Adam. The analysis of the unbounded grammar (a) has a depth of 5, shown by the deeply nested center embedding analysis of the span *scratch or cut you can clean it with a ball of cotton*, which does not resemble any kind of linguistic analysis. Using depth-bounding, such analyses will never be entertained by the inducer, even when in this case the unbounded grammar may have a higher likelihood than the bounded grammar. The analysis of the bounded grammar (b) is closer to linguistic annotation, where the *if* clause is separated from the main clause. Some of the noun and prepositional, phrases are also clearly identified, and the depth of the tree is 3.

8.2 Results for Bounded Induction on Held-Out Partition

The bounded induction model with β = 0.1, $C = 30$, and $D = 3$ is then evaluated on held-out data sets in three languages: Eve in English, Tong in Mandarin Chinese, and Leo in German. Figure 16 shows that the models bounded at depth 3 are more accurate than unbounded models with both unlabeled and labeled evaluation metrics for all data sets, similar to what has been observed in Adam. Significance testing with item-level permutation with unlabeled F1 shows the accuracy differences across three data sets are all highly significant ($p < 0.001$).

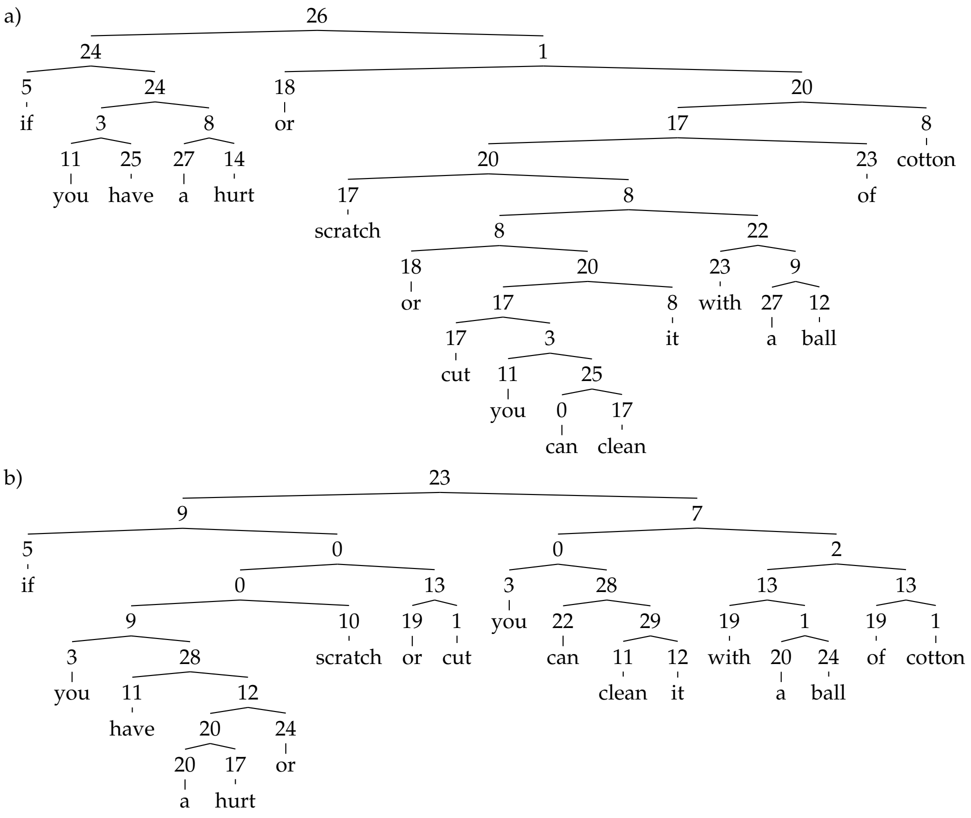
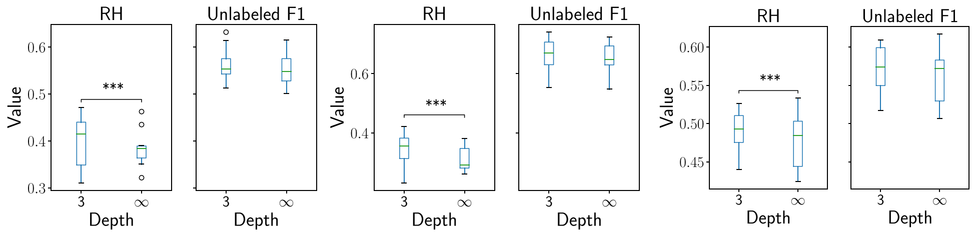


Figure 15
Example syntactic analyses from $D = \infty$ and $D = 3$ runs on Adam with the highest likelihood.



(a) Depth-bounding on Eve (b) Depth-bounding on Tong (c) Depth-bounding on Leo

Figure 16
Comparison of labeled and unlabeled evaluation of grammars bounded at depth 3 and unbounded grammars on English (Eve), Chinese Mandarin (Tong), and German (Leo) data sets from CHILDES ($\beta = 0.1, C = 30$).

8.3 Analysis of Learned Syntactic Categories and Grammatical Rules on Chinese

Table 6 shows the top 20 most frequent induced rules and annotated rules found in the automatically parsed data. Induced rules that capture linguistic phenomena that are different from English are described below.

Table 6
The most frequent rules induced in Tong and their correspondences in the gold annotation. The examples are randomly sampled from the bounded induced trees.

Rank	Rule	Corresponding gold rules, counts and examples
1.	17→3 12 (7,593)	?? →?? ?? (0.40); IP →NP VP (0.37); IP →ADVP VP (0.06); IP →INTJ VP (0.05) 来 | 给你吧 ; 你 | 怎么认识龙的 ; 他 | 今天过生日啦 ; 哎 | 游泳
2.	12→6 12 (7,056)	?? →?? ?? (0.24); VP →ADVP VP (0.17); ?? →ADVP ?? (0.14); ?? →VV ?? (0.06) 才 | 能配起来 ; 还 | 得火车头拉着这个车厢才会走啊 ; 会 | 说成语了 ; 问你 | 你喝不喝
3.	8→11 24 (4,217)	VP →VV NP (0.23); ?? →?? ?? (0.22); IP →VV NP (0.05); VP →VV VV (0.05) 写 | 着 ; 涂 | 到外头去 ; 坐 | 在上边 ; 涂 | 完
4.	17→13 26 (4,010)	CP →IP SP (0.52); ?? →?? ?? (0.29); CP →NP SP (0.03); IP →IP VP (0.02) 起床了 | 就放学啦 ; 同同你还记得你上次这画的 | 是什么吗 ; 你又到我屋里去干吗 | 呀
5.	13→3 12 (3,892)	IP →NP VP (0.29); ?? →?? ?? (0.17); ?? →NP VP (0.16); ?? →NP ?? (0.14) 这 | 故事你都很熟悉了 ; 哦这 | 像一朵花 ; 看小羊 | 在画什么 ; 小朋友 | 就把这个蜡烛吹灭了
6.	12→8 14 (3,673)	?? →VP SP (0.37); ?? →?? ?? (0.20); VP →VV AS (0.09); VP →VRD AS (0.06) 找不着2 | 了 ; 填那个亲子阅读手册 | 呢 ; 救同同 | 了 ; 脏 | 不脏
7.	12→21 15 (3,407)	VP →VC NP (0.26); VP →VV NP (0.14); ?? →?? ?? (0.13); VP →VE NP (0.11) 有 | 几个海豚 ; 学 | 熊大熊二说话 ; 是 | 胜利 ; 有 | 谁
8.	6→4 3 (2,366)	PP →P NP (0.26); ?? →VV NP (0.24); ?? →BA NP (0.12); ?? →ADVP NP (0.07) 离 | 爸爸 ; 给 | 爸爸 ; 刚才 | 这下两个脚 ; 穿 | 你
9.	12→6 8 (2,173)	VP →ADVP VP (0.23); VP →VV VP (0.20); ?? →?? ?? (0.12); ?? →ADVP VP (0.09) 给同同 | 看一看 ; 可以 | 做个桥 ; 都 | 坐好嘞 ; 用手 | 画画
10.	0→2 1 (1,828)	QP →CD CLP (0.41); DP →DT CLP (0.37); ?? →NT NT (0.02); ?? →CD CD (0.02) 这 | 个 ; 还有 | 半 ; 两 | 个 ; 六 | 月
11.	8→6 8 (1,547)	?? →?? ?? (0.21); VP →VV VP (0.13); VP →ADVP VP (0.08); IP →ADVP VP (0.07) 能 | 点歌 ; 能 | 逃跑 ; 那你 | 找一找 ; 不 | 能按
12.	3→0 29 (1,454)	NP →DP NP (0.22); NP →ADJP NP (0.18); NP →DNP NP (0.12); ?? →?? ?? (0.11) 汽车 | 书 ; 哪些 | 小动物 ; tom | 猫 ; 你的 | 鞋子
13.	15→0 29 (1,426)	NP →DP NP (0.17); NP →QP NP (0.17); ?? →?? ?? (0.16); NP →ADJP NP (0.09) 高的 | 楼 ; 白色的 | 呵 ; 屋顶的 | 话 ; 三根 | 蜡烛
14.	3→23 22 (1,356)	?? →INTJ NP (0.28); ?? →ADVP NP (0.25); ?? →IP NP (0.07); ?? →NP NP (0.06) 哎呀 | 那 ; 哦 | 同同 ; 嗯 | 我 ; 同同 | 这
15.	17→8 14 (1,281)	CP →IP SP (0.38); ?? →?? ?? (0.32); IP →VV AS (0.07); IP →VP SP (0.04) 在这儿 | 呢 ; 荡秋千 | 咯 ; 给你吃 | 好吃的 ; 就在那个厨房里 | 啊
16.	24→10 9 (1,074)	?? →VV VV (0.45); ?? →VP VV (0.06); ?? →VRD VV (0.05); ?? →PP VP (0.05) 起 | 来 ; 过 | 一加 ; 进 | 去 ; 过 | 鱼
17.	17→6 8 (763)	?? →?? ?? (0.37); IP →ADVP VP (0.36); IP →VV VP (0.10); IP →PP VP (0.07) 快 | 走 ; 不 | 让我坐在这儿 ; 屋顶 | 放在上面 ; 要 | 等一会
18.	3→2 1 (693)	DP →DT CLP (0.63); QP →CD CLP (0.13); NP →DP NP (0.05); DP →DT QP (0.03) 这 | 个 ; 一 | 眼 ; 这 | 个 ; 这 | 个
19.	18→20 18 (626)	LCP →NP LC (0.45); ?? →?? ?? (0.10); ?? →NN LC (0.09); NP →NN NN (0.06) 腿 | 上 ; 妈妈 | 干净的枕头上 ; 桶 | 里 ; 嘴 | 底下
20.	12→19 7 (595)	VP →ADVP VP (0.46); ?? →VP DEC (0.08); ?? →ADVP VP (0.08); ?? →VP SP (0.05) 开 | 的 ; 好 | 高 ; 特别 | 猛 ; 这么 | 远

- The first and fourth rule show two different ways to form sentences, with the first rule used mainly for declarative sentences, and the fourth rule for questions. The fourth rule splits a question into an ordinary sentence with a sentence-final particle, which includes 吗 (*ma*, the particle for a yes–no question) and 好不好 (*good or not*, a phrase to turn a declarative sentence into a question). The fifteenth rule also splits a sentence into a sentence and a particle, but the whole sentence is declarative. The sentence-final punctuation helps the inducer to distinguish these two types of sentences, but it must rely on statistics to split the particle off of the rest of the sentence, because the particle in many cases is not present. It is worth noting that the two most frequent rules are also the rules with the largest number of unattested constituents, because these rules are for sentence-level constituents, and bracketing error at any nodes below may cause the top level rule to have unattested constituents.

- Chinese Mandarin is a classifier language. A classifier is usually needed when a determiner or a number occurs before a noun. The twelfth and thirteenth rules are used for combining noun phrases with prenominal modifiers like a determiner phrase or a quantifier phrase, which in turn is formed by the eighteenth rule. The twelfth rule constructs noun phrases at the subject position, and the thirteenth rule is for objects. Similar to English, the nominative-accusative case distinction is still induced in the grammar although there is no morphological marker or lexical distinction for them.
- There appear to be statistical cues for semantics of nouns, too. Table 7 shows the four main induced categories in this grammar. The distinction is clear, the first category is for personal pronouns and relative terms, the second for general nouns, the third for location terms (they are annotated as LC in Penn Chinese Treebank, but show up as NPs because of unary chain removal) which are used as nouns in Chinese, and finally wh-pronouns. These four classes of nouns seem to have distinct statistical properties. For example, personal pronouns and relative terms almost never appear with determiner phrases and quantifier phrases, but general nouns almost always do. The location terms most of the time are modified by noun phrases and wh-pronouns appear in questions.
- *Ba* in Chinese Mandarin takes the object of the verb and moves it to the preverbal position, making a normally SVO sentence SOV. It can be considered as a light verb (Ding 2001; Duan and Schuler 2015) or a preposition and a case marker (Ye, Zhan, and Zhou 2007). Figure 17 shows the automatic annotation is consistent with annotation guidelines for Penn Chinese Treebank. The induced analysis seems to support *ba* as a preposition: The *ba* phrase combines with the VP after it and the result is also a VP.

9. Experiment 4: Natural Bounding in Child-Directed and Adult Language Data

It seems likely that children and adults have different working memory capabilities, and therefore different capacities for center embedding. This section describes experiments

Table 7
Recall of the top 3 most frequent words in the four induced categories that correspond to nouns in Tong.

Rank	Induced category	Category count	Attested words and relative frequency
1.	3	12,425	你(you, 0.28); 我(I, 0.07); 妈妈(mom, 0.06)
2.	29	4,073	车(car, 0.03); 人(people, 0.02); 东西(thing, 0.03)
3.	18	1,991	上(up, 0.16); 这里(this place, 0.13); 里(inside, 0.05)
4.	15	1,707	什么(what, 0.38); 谁(who, 0.03); 几(how many, 0.02)

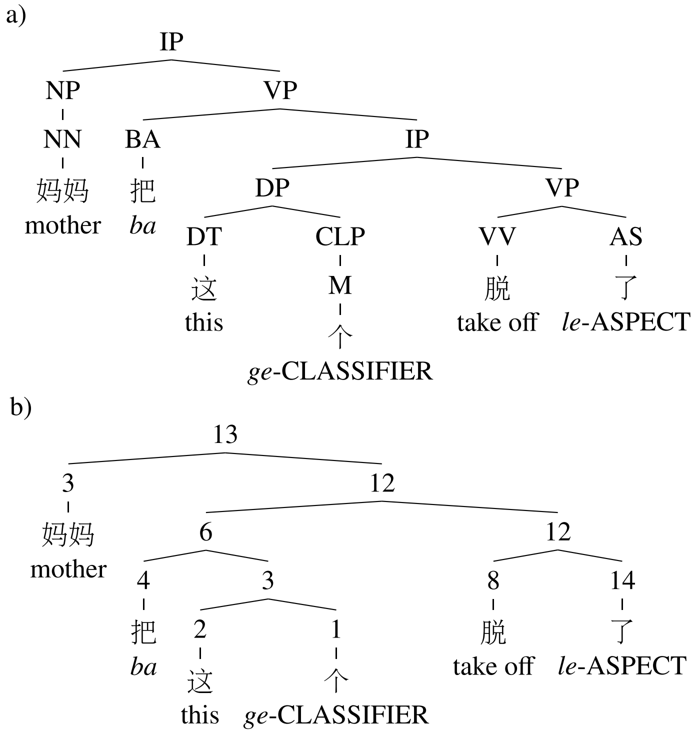


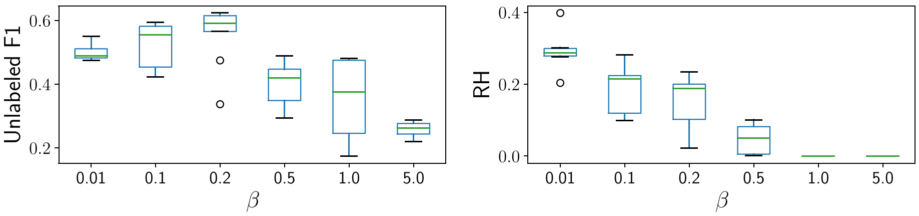
Figure 17
Example syntactic analyses for a *ba* construction in Tong: 妈妈把这个脱了(Mom takes off this).

to measure the difference between adult and child-directed speech during statistical grammar induction.

Following earlier work (Klein and Manning 2004; Seginer 2007a; Ponvert, Baldridge, and Erk 2011; Shain et al. 2016), these experiments use the Wall Street Journal section of the Penn Treebank (Marcus, Santorini, and Marcinkiewicz 1993) as adult language data. Sentences shorter than or equal to 20 words are used in the experiments, partitioned into a development set, WSJ20Dev, and a held-out set, WSJ20Test, following Jin et al. (2018a).

We are first interested in how the sparse preference related to the hyperparameter β behaves on adult-directed newswire data. Figure 18 shows the interaction between β values and the evaluation metrics. The right-branching bias contributed by high beta can still be seen on this data set. The unlabeled F1 peaks at $\beta = 0.2$, but when labels are taken into account, the evaluation metric RH peaks at a much lower $\beta = 0.01$, indicating again the preference of sparse priors in labeled grammar induction. The observation that the optimal β on WSJ20Dev is lower than on child-directed speech may point at the possibility that the right-branching bias from high β on child-directed speech provides an advantage to evaluation of structures in RH such that a certain level of that right branching bias is favored. On WSJ20Dev, however, the advantage from the right-branching bias is outweighed by the disadvantage it brings to labeling accuracy, therefore the optimal β appears to be low. Experiments with WSJ20 data sets use $\beta = 0.01$.

Because the sentences in both transcribed child-directed speech corpora and adult-directed newswire corpora are produced by humans, several predictions can be made



(a) Unlabeled F1 on WSJ20Dev

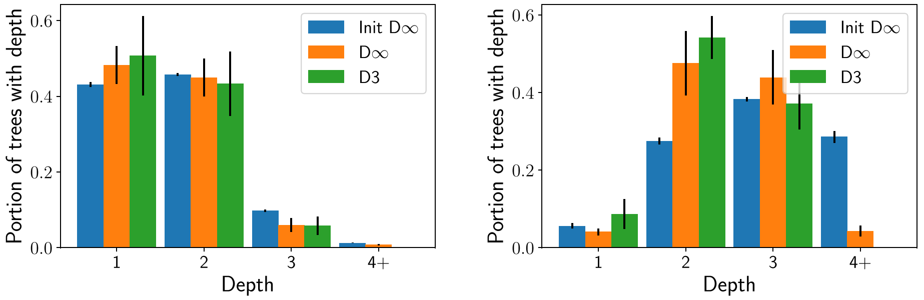
(b) RH on WSJ20Dev

Figure 18
The effect of β on the WSJ20Dev data set.

for the distribution of tree depths, which is the proportion of sentences whose induced tree has a certain maximum depth, from the induced trees on both kinds of data sets. First, the distribution of tree depths at initialization should show a wide range, because at this time the tree depths are only correlated to sentence length. Second, the distribution of tree depths for unbounded models should substantially narrow after training, where the induced grammar implicitly learns the human memory limits from data, but imperfect learning is also expected, shown by a small number of sentences with trees with depth 4 or higher. Third, compared to child-directed data, adult-directed newswire data have more complicated structure, so the expected tree depth on adult-directed data should be higher than child-directed data. Note that the child-directed data are still generated by adults; therefore the maximum memory depth is still expected to be similar to other adult-generated data. However, the trees generated by the grammar may reflect the fact that the sentences are relatively short and simple in structure in child-directed data.

Figure 19 shows the distribution of tree depths on the development sets of child-directed data, Adam, and adult-directed data, WSJ20dev. The three bars represent the induced trees from initialized unbounded grammars (blue), unbounded grammars after training (orange), and grammars bounded at depth 3 (green). The predictions listed above bear out in the results. The expected tree depth on Adam is 1.59 for the unbounded models, and 1.55 for the bounded models, which is a significant difference ($p < 0.001$ using a permutation test). The expected tree depth on WSJ20Dev is 2.49 for the unbounded models, and 2.28 for the bounded models, which is also a significant difference ($p < 0.001$ using a permutation test), confirming that depth-bounding leads to trees with lower usage of stack elements. Also notably, the percentage of trees with depth 4 or higher from unbounded models on both Adam and WSJ20dev is very small, indicating that the unbounded models are able to learn implicitly the human memory constraints from data. However, the unbounded grammars face a larger search space of possible grammars, and they may need to allocate rules, categories, or probability mass for deep tree structures, which make them less accurate, as shown in previous experiments.

This experiment points to a potential unsupervised method for determining the optimal depth limit for a data set, although the number of potential candidate values for the depth limit parameter is already very small. The optimal maximum depth on Adam appears to be 3. On WSJ20Dev it appears to be 3 or 4, taking into account results from psycholinguistic literature (Karlsson 2007, 2010; Schuler et al. 2010). The following experiments choose depth 3 as the maximum allowed depth for the depth-bounding model. We leave the empirical investigation of accuracy of grammars bounded at depth 4 for future work, because of its extensive requirements of resources and runtime.



(a) Distribution of tree depths on Adam

(b) Distribution of tree depths on WSJ20dev

Figure 19
Distribution of tree depths on child-directed and adult-directed sentences. The difference of expected tree depths of bounded and unbounded models is significant on both data sets.

10. Experiment 5: Replication of Depth Bound Effects in Newswire Corpora

Independent of its value for modeling child language acquisition, there may also be engineering benefits to applying center-embedding depth bounds during grammar induction on newswire corpora. The proposed bounded and unbounded models are run with 10 random seeds, with the mean accuracy and standard deviation shown in Table 8. Both models achieve similar unlabeled accuracy, but the significant difference between labeled evaluation accuracy ($p < 0.001$) indicates depth-bounding facilitates discovery of grammars with better labeling accuracy, leading to overall better accuracy when labels are taken into consideration. This shows that depth-bounded grammar induction models as a human language acquisition model also works with more syntactically complex newswire text.

Results of induced grammars with highest likelihoods from several induction systems are presented in Table 9 for comparison. Neural induction systems achieve higher accuracy than the pure statistical systems in this work, but are not as easy to augment with depth bounds. The larger accuracy difference between statistical models and neural models on WSJ compared with child-directed data may indicate that more categories are required to capture relatively complex syntactic structures, of which the neural models have 90 but the statistical models have 30. We therefore leave integration of depth bounding into neural induction systems for future work.

11. Conclusion

This article describes unbounded and depth-bounded PCFG induction models, intended to represent something akin to a competence grammar and a performance model

Table 8
Mean and standard deviation of scores on WSJ20Test data set with proposed models on 10 runs. The difference of RH between the two models is significant ($p < 0.001$).

System	F1	RH
this work (D=∞, C=30)	0.49 ± 0.02	0.28 ± 0.03
this work (D=3, C=30)	0.49 ± 0.02	0.31 ± 0.02

Table 9
Accuracy scores on WSJ20Test data set with previously published induction systems.

System	F1	RH
Seginer (2007a)	**0.61**	–
Ponvert, Baldridge, and Erk (2011)	0.44	–
Jin et al. (2018b)	**0.61**	–
Jin et al. (2019)	0.51	–
Kim, Dyer, and Rush (2019) without z	0.52	0.35
Kim, Dyer, and Rush (2019) with z	0.54	**0.37**
this work (D=∞, C=30)	0.51	0.28
this work (D=3, C=30)	0.51	0.32

that takes working memory and processing constraints into account, and evaluates them on transcribed corpora of child-directed speech.

Results from Section 7 show that the model predicts 44%—nearly half—of labeled attested constituents in the held-out partition, according to the RH measure. It is interesting that so much linguistic structure can be predicted from words alone, without semantics, and without universal linguistic constraints. It is anticipated that this assessment will encourage future research to discover how the other half can be predicted. Moreover, properties of the induced structures might be used to guide future linguistic analysis—for example, incorporating readily induced properties of number, case, and subcategorization into standard part-of-speech tag sets.

Results in Section 8 also show a statistically significant positive effect for depth bounding on grammar induction. This suggests a natural explanation for simplified grammatical behaviors observed in child production, as due to a basic memory-bounding mechanism that facilitates acquisition.

Acknowledgments

The authors would like to thank the anonymous reviewers and the editor for their helpful comments. Computations for this project were partly run on the Ohio Supercomputer Center. This research was funded by Defense Advanced Research Projects Agency award HR0011-15-2-0022 and by National Science Foundation grant 1816891. 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

Abney, Steven P. and Mark Johnson. 1991. Memory Requirements and local ambiguities of parsing strategies. *Journal of Psycholinguistic Research*, 20(3):233–250. DOI: `https://doi.org/10.1007/BF01067217`

Bannard, Colin, Elena Lieven, and Michael Tomasello. 2009. Modeling children's early grammatical knowledge. In *Proceedings of the National Academy of Sciences of the United States of America*, 106(41):17284–17289. DOI: `https://doi.org/10.1073/pnas.0905638106`, PMID: 19805057, PMCID: PMC2765208

Behrens, Heike. 2006. The input-output relationship in first language acquisition. *Language and Cognitive Processes*, 21(1–3):2–24. DOI: `https://doi.org/10.1080/01690960400001721`, PMID: 19805057, PMCID: PMC2765208

Berg-Kirkpatrick, Taylor, Alexandre Bouchard-Côté, John DeNero, and Dan Klein. 2010. Painless unsupervised learning with features. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 582–590, Los Angeles, CA.

Bisk, Yonatan, Christos Christodoulopoulos, and Julia Hockenmaier. 2015. Labeled grammar induction with minimal supervision. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 870–876, Beijing. DOI: https://doi.org/10.3115/v1/P15-2143

Bisk, Yonatan and Julia Hockenmaier. 2012. Simple robust grammar induction with combinatory categorial grammars. In *Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence*, 26(1):1643–1649.

Brown, Roger. 1973. *A First Language: The Early Stages*, Harvard University Press, Cambridge, MA.

Carroll, Glenn and Eugene Charniak. 1992. Two experiments on learning probabilistic dependency grammars from corpora. *Working Notes of the Workshop on Statistically-Based NLP Techniques*, (March):1–13.

Charniak, Eugene and Mark Johnson. 2005. Coarse-to-fine n-best parsing and MaxEnt discriminative reranking. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 173–180, Ann Arbor, MI. DOI: https://doi.org/10.3115/1219840.1219862

Chomsky, Noam. 1965. *Aspects of the Theory of Syntax*, MIT Press, Cambridge, MA. DOI: https://doi.org/10.21236/AD0616323, PMID: 14125365

Chomsky, Noam. 1980. On cognitive structures and their development: A reply to Piaget, Piattelli-Palmarini, Massimo, editor, *Language and Learning: The Debate Between Jean Piaget and Noam Chomsky*, Harvard University Press, pages 751–755, Cambridge, MA.

Chomsky, Noam. 1986. *Knowledge of Language: Its Nature, Origin, and Use*, Praeger, New York.

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

Cramer, Bart. 2007. Limitations of current grammar induction algorithms. In *Proceedings of the ACL 2007 Student Research Workshop*, pages 43–48, Prague. DOI: https://doi.org/10.3115/1557835.1557845, PMID: 17194528

de Saussure, Ferdinand. 1916. *Cours de linguistique générale*. Bally, Charles and Sechehaye, Albert, editors.

Deng, Xiangjun, Virginia Yip, Brian Macwhinney, Stephen Matthews, Mai Ziyin, Zhong Jing, and Hannah Lam, 2018. A multimedia corpus of child Mandarin: The Tong corpus. *Journal of Chinese Linguistics* 46(1):69–92. DOI: https://doi.org/10.1353/jcl.2018.0002

Ding, Picus Sizhi. 2001. Semantic change versus categorical change: A study of the development Of BA in Mandarin. *Journal of Chinese Linguistics*, 29(1):102–128.

Drozdov, Andrew, Patrick Verga, Mohit Yadav, Mohit Iyyer, and Andrew McCallum. 2019. Unsupervised latent tree induction with deep inside-outside recursive auto-encoders. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1129–1141, Minneapolis, MN. DOI: https://doi.org/10.18653/v1/N19-1116

Duan, Manjuan and William Schuler. 2015. Parsing Chinese with a generalized categorial grammar. In *Proceedings of the Grammar Engineering Across Frameworks (GEAF) 2015 Workshop*, pages 25–32, Beijing. DOI: https://doi.org/10.18653/v1/W15-3304

Freudenthal, Daniel, Julian M. Pine, Javier Aguado-Orea, and Fernand Gobet. 2007. Modeling the developmental patterning of finiteness marking in English, Dutch, German, and Spanish using MOSAIC. *Cognitive Science*, 31(2):311–341.

Fu, King Sun and Taylor L. Booth. 1975. Grammatical inference: Introduction and survey. *IEEE Transactions on Systems, Man, and Cybernetics*, SMC-5(1,4): 95–111,409–423. DOI: https://doi.org/10.1109/TSMC.1975.5408432

Gold, Mark E. 1967. Language identification in the limit. *Information and Control*, (10):447–474. DOI: https://doi.org/10.1016/S0019-9958(67)91165-5

Goldwater, Sharon and Tom Griffiths. 2007. A fully Bayesian approach to unsupervised part-of-speech tagging. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 744–751, Prague.

Goodman, Joshua. 1998. Parsing Inside-Out. arXiv preprint cmp-lg/9805007.

Jiang, Yong, Wenjuan Han, and Kewei Tu. 2016. Unsupervised neural dependency parsing. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 763–771, Austin, TX.

Jin, Lifeng, Finale Doshi-Velez, Timothy Miller, William Schuler, and Lane Schwartz. 2018a. Depth-bounding is effective: Improvements and evaluation of unsupervised PCFG induction. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2721–2731, Brussels. DOI: https://doi.org/10.18653/v1/D18-1292

Jin, Lifeng, Finale Doshi-Velez, Timothy Miller, William Schuler, and Lane Schwartz. 2018b. Unsupervised grammar induction with depth-bounded PCFG. *Transactions of the Association for Computational Linguistics*, 6:211–224. DOI: https://doi.org/10.1162/tacl_a_00016

Jin, Lifeng, Finale Doshi-Velez, Timothy Miller, Lane Schwartz, and William Schuler. 2019. Unsupervised learning of PCFGs with normalizing flow. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2442–2452, Florence. DOI: https://doi.org/10.18653/v1/P19-1234, PMID: 30482428

Johnson, Mark, Thomas Griffiths, and Sharon Goldwater. 2007. Bayesian inference for PCFGs via Markov chain Monte Carlo. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 139–146, Rochester, NY.

Johnson-Laird, Philip N. 1983. *Mental Models: Towards a Cognitive Science of Language, Inference, and Consciousness*. Harvard University Press, Cambridge, MA.

Karlsson, Fred. 2007. Constraints on multiple center-embedding of clauses. *Journal of Linguistics*, 43:365–392. DOI: https://doi.org/10.1017/S0022226707004616

Karlsson, Fred. 2010. Working memory constraints on multiple center-embedding. In *Proceedings from the 32nd Annual Meeting of the Cognitive Science Society*, pages 2045–2050, Portland, OR.

Kates, Carol. 1976. A critique of Chomsky's theory of grammatical competence, *Forum Linguisticum*, 1(1):15–24.

Kim, Yoon, Chris Dyer, and Alexander Rush. 2019. Compound probabilistic context-free grammars for grammar induction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2369–2385, Florence. DOI: https://doi.org/10.18653/v1/P19-1228, PMID: 31697821

Kim, Yoon, Alexander Rush, Lei Yu, Adhiguna Kuncoro, Chris Dyer, and Gábor Melis. 2019. Unsupervised recurrent neural network grammars. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1105–1117, Minneapolis, MN. DOI: https://doi.org/10.18653/v1/N19-1114

Kitaev, Nikita and Dan Klein. 2018. Constituency parsing with a self-attentive encoder. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2676–2686, Melbourne. DOI: https://doi.org/10.18653/v1/P18-1249

Klein, Dan and Christopher 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-04)*, pages 478–485, Barcelona. DOI: https://doi.org/10.3115/1218955.1219016

Klein, Dan 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*, pages 128–135, Philadelphia, PA. DOI: https://doi.org/10.3115/1073083.1073106

Liang, Percy, Slav Petrov, Michael Jordan, and Dan Klein. 2007. The infinite PCFG using hierarchical Dirichlet processes. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 688–697, Prague.

Lieven, Elena V. M., Julian M. Pine, and Gillian Baldwin. 1997. Lexically-based learning and early grammatical development. *Journal of Child Language*, 24(1):187–219. DOI: https://doi.org/10.1017/S0305000996002930, PMID: 9154014

Macwhinney, Brian. 1992. *The CHILDES Project: Tools for Analyzing Talk*, third edition, Lawrence Elrbaum Associates, Mahwah, NJ.

Marcus, Mitchell P., Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

Miller, George. 1975. Some comments on competence and performance. In Doris Aaronson and Robert Rieber, editors, *Developmental Psycholinguistics and Communication Disorders*, volume 263 of *Annals of the New York Academy of Sciences*. The New York Academy of Science, New York, pages 201–204. DOI: https://doi.org/10.1111/j.1749-6632.1975.tb41583.x

Mintz, Toben H. 2003. Frequent frames as a cue for grammatical categories in child directed speech. *Cognition*, 90(1):91–117. DOI: https://doi.org/10.1016/S0010-0277(03)00140-9.

Naseem, Tahira, Harr Chen, Regina Barzilay, and Mark Johnson. 2010. Using universal linguistic knowledge to guide grammar induction. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 1234–1244, Cambridge, MA.

Newmeyer, Frederick. 2010. Grammar is grammar and usage is usage. *Language*, 79(4):682–707. DOI: https://doi.org/10.1353/lan.2003.0260

Noji, Hiroshi, Yusuke Miyao, and Mark Johnson. 2016. Using left-corner parsing to encode universal structural constraints in grammar induction. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 33–43, Austin, TX. DOI: https://doi.org/10.18653/v1/D16-1004

Pate, John K. and Sharon Goldwater. 2013. Unsupervised dependency parsing with acoustic cues. *Transactions of the Association for Computational Linguistics*, 1:63–74. DOI: https://doi.org/10.1162/tacl_a_00210

Pearl, Lisa and Jon Sprouse. 2013. Syntactic islands and learning biases: Combining experimental syntax and computational modeling to investigate the language acquisition problem. *Language Acquisition*, 20(1):23–68. DOI: https://doi.org/10.1080/10489223.2012.738742

Pereira, Fernando and Yves Schabes. 1992. Inside-outside reestimation from partially bracketed corpora. In *30th Annual Meeting of the Association for Computational Linguistics*, pages 128–135, Newark, DE. DOI: https://doi.org/10.3115/981967.981984

Perfors, Amy, Joshua B. Tenenbaum, and Terry Regier. 2006. Poverty of the stimulus? A rational approach. In *Proceedings of the 28th Annual Conference of the Cognitive Science Society*, pages 663–668, Vancouver.

Ponvert, Elias, Jason Baldridge, and Katrin Erk. 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: Human Language Technologies*, pages 1077–1086, Portland, OR.

Pullum, Geoffrey K. and Barbara C. Scholz. 2002. Empirical assessment of stimulus poverty arguments. *Linguistic Review*, 18:9–50. DOI: https://doi.org/10.1515/tlir.19.1-2.9

Pylyshyn, Zenon W. 1973. The role of competence theories in cognitive psychology. *Journal of Psycholinguistic Research*, 2(1):21–50. DOI: https://doi.org/10.1007/BF01067110, PMID: 24197794

Redington, Martin, Nick Chater, and Steven Finch. 1998. Distributional information: A powerful cue for acquiring syntactic categories. *Cognitive Science*, 22(4):425–469. DOI: https://doi.org/10.1207/s15516709cog2204_2

Rosenkrantz, D. J. and P. M. Lewis. 1970. Deterministic left corner parsing. In *11th Annual Symposium on Switching and Automata Theory*, pages 139–152, Santa Monica, CA. DOI: https://doi.org/10.1109/SWAT.1970.5

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

Seginer, Yoav. 2007a. Fast unsupervised incremental parsing. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 384–391, Prague. DOI: https://doi.org/10.1162/coli.2010.36.1.36100

Seginer, Yoav. 2007b. *Learning Syntactic Structure*. Ph.D. thesis, University of Amsterdam.

Shain, Cory, William Bryce, Lifeng Jin, Victoria Krakovna, Finale Doshi-Velez, Timothy Miller, William Schuler, and Lane Schwartz. 2016. Memory-bounded left-corner unsupervised grammar induction on child-directed input. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 964–975, Osaka.

Shen, Yikang, Zhouhan Lin, Chin-Wei Huang, and Aaron C. Courville. 2018. Neural language modeling by jointly learning syntax and lexicon. In *6th International Conference on Learning Representations, ICLR 2018, Conference Track Proceedings*, Vancouver.

Shen, Yikang, Shawn Tan, Alessandro Sordoni, and Aaron C. Courville. 2019. Ordered neurons: Integrating tree structures into recurrent neural networks. In *7th International Conference on Learning Representations, ICLR 2019*, New Orleans, LA.

Skut, Wojciech, Thorsten Brants, Brigitte Krenn, and Hans Uszkoreit. 1998. A linguistically interpreted corpus of German newspaper text. In *Proceedings of the ESSLLI Workshop on Recent Advances in Corpus Annotation*, page 7, Saarbrücken.

Smith, Noah Ashton. 2006. *Novel Estimation Methods for Unsupervised Discovery of Latent Structure in Natural Language Text*. PhD Thesis, Johns Hopkins University.

Solomonoff, Ray J. 1964. A formal theory of inductive inference. *Information and Control*, 7(1–2):1–22, 224–254. DOI: https://doi.org/10.1016/S0019-9958(64)90131-7

Steedman, Mark. 2002. Formalizing affordance. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, pages 834–839, Fairfax, VA.

Thompson, Susan P. and Elissa L. Newport. 2007. Statistical learning of syntax: The role of transitional probability. *Language Learning and Development*, 3(1):1–42. DOI: https://doi.org/10.1080/15475440709336999.

Tomasello, Michael. 2003. *Constructing a Language: A Usage-Based Theory of Language Acquisition*, Harvard University Press, Cambridge, MA.

Tu, Kewei. 2012. *Unsupervised Learning of Probabilistic Grammars*. Ph.D. thesis, Iowa State University.

van Schijndel, Marten, Andy Exley, and William Schuler. 2013. A model of language processing as hierarchic sequential prediction. *Topics in Cognitive Science*, 5(3):522–540. DOI: https://doi.org/10.1111/tops.12034, PMID: 23765642

Xia, Fei, Martha Palmer, Nianwen Xue, Mary Ellen Okurowski, John Kovarik, Fu-Dong Chiou, Shizhe Huang, Tony Kroch, and Mitch Marcus. 2000. Developing guidelines and ensuring consistency for Chinese text annotation. In *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC'00)*, Athens.

Ye, Zheng, Weidong Zhan, and Xiaolin Zhou. 2007. The semantic processing of syntactic structure in sentence comprehension: An ERP study. *Brain Research*, 1142(1):135–145. DOI: https://doi.org/10.1016/j.brainres.2007.01.030, PMID: 17303093