

# Universal-Guided Grammar Induction

Yuan Gao<sup>1,2</sup>, Weiwei Sun<sup>1</sup>,

{yg386, ws390}@cam.ac.uk

<sup>1</sup>Department of Computer Science & Technology, University of Cambridge

<sup>2</sup>ALTA Institute

## Abstract

We study whether Universal Grammar syntactic principles can serve as explicit inductive biases for unsupervised grammar induction. We introduce a neuro-symbolic PCFG framework that converts hypothesized linguistic universals into rule schemata, and then learns rule probabilities from child-directed speech. We investigate three assumptions: 1) extended projection, 2) prototype theory of part-of-speech, and 3) the final-over-final constraint. Experiments show that with a much smaller hypothesis space, the universal-prior model achieves comparable performance to state-of-the-art neural grammar induction models, and yields more interpretable functional category assignment. The full dataset and code are available at [https://github.com/yuan-w-gao/ug\\_bias](https://github.com/yuan-w-gao/ug_bias).

## 1 Introduction

Grammar induction is a fundamental task in natural language processing and computational linguistics (Clark, 2002; Clark et al., 2012; Klein and Manning, 2004). Given a corpus of unannotated sentences as input, the goal is to recover the underlying hierarchical grammatical structure. Computationally, a grammar induction system can be broken down into two components, as illustrated in Figure 1. While modern neural approaches do well at learning distributed parameterizations from large-scale data (Shen et al., 2017; Kim et al., 2019; Yang et al., 2021b), they typically operate over a dense, unconstrained rule space (Figure 1, left panel). Therefore, their induced structures are often difficult to interpret and weakly constrained.

When grammar induction is used as a computational model of language acquisition (Johnson, 2013), however, the learned structures are expected not only to support prediction, but also to provide cognitively and linguistically meaningful explanations (Buttery, 2004; Kwiatkowski et al., 2012;

Perfors et al., 2011a). Children acquire highly complex grammatical systems from limited and noisy input, which strongly suggests that human learning relies on prior structural constraints rather than an unconstrained search over arbitrary grammar rules (Chomsky, 1965; Perfors et al., 2011a; Berwick et al., 2011). Previous computational models have attempted to constrain the learner in various ways. Some works use linguistically structured formalisms, such as Combinatory Categorical Grammar (CCG; Steedman, 1996, 2001). Others impose broader structural constraints motivated primarily by computational efficiency or mathematical bounds, such as left-corner parsing constraints (Noji et al., 2016) and derivational depth bounding (Jin et al., 2018).

This paper studies whether hypotheses inspired by Universal Grammar (UG; Chomsky, 1993, 2005, 2014) can be encoded as structural constraints on the space of grammar available to an unsupervised machine learner. To separate innate structural biases from statistical learning, we operationalize candidate universal principles as explicit rule schemata within a grammar induction system. As illustrated in Figure 1, our model is neither given a manually specified exact grammar (right panel) nor left to search a dense, unconstrained space (left panel). Instead, the UG-inspired schemata define a restricted structural superset (middle panel). This design explicitly dictates which grammatical structures are permissible before learning begins, and a neural estimator then maximizes the marginal log-likelihood of unannotated sentences to learn the rule probabilities. Within this approach, we operationalize three specific theoretical principles: the prototype theory of parts-of-speech (Croft, 2001), extended projection (EPP; Grimshaw, 2000), and the Final-over-Final Constraint (FOFC; Biberauer et al., 2014)

First, we incorporate the prototype theory of parts-of-speech (PT; Croft, 2001). PT hypothesizes

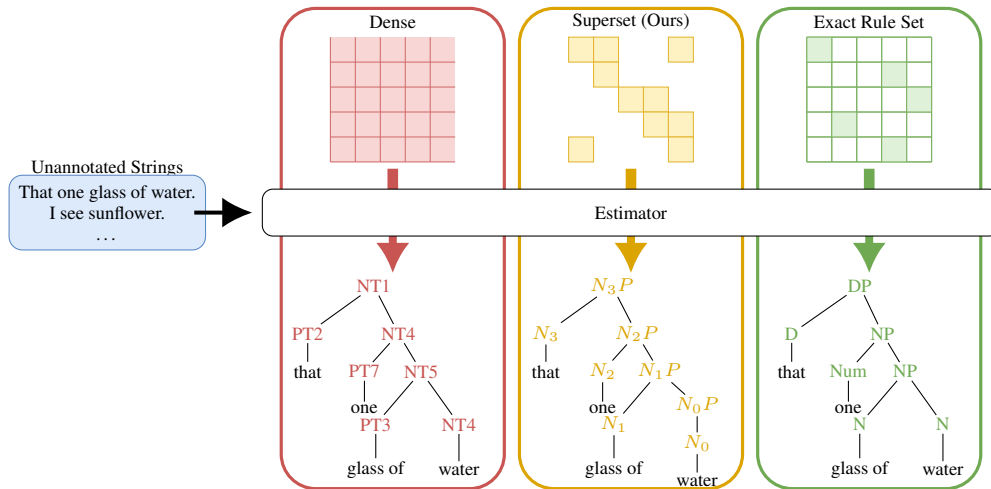


Figure 1: Architectural comparison of hypothesis spaces for unsupervised grammar induction. (Left) Standard neural approaches operate over a dense, unconstrained rule space where any combination of nonterminals is initially permissible. (Right) An exact, manually specified symbolic grammar. (Middle) Our Universal Grammar-inspired rule schemata constrain the structural hypothesis space a priori to a permissible superset. A neural estimator then learns the rule probabilities over this restricted space directly from unannotated strings.

that the fundamental communicative functions of human language are referentiality, predication, and modification. By structuring the grammar’s primitive domains around these three cognitive functions rather than traditional, language-specific lexical categories, the model avoids relying on grammatical labels.

Second, we constrain the internal organization of these domains using extended projection. This principle proposes that a lexical head is not an isolated category, but is embedded in a sequence of increasingly abstract functional layers within a single grammatical domain (Grimshaw, 2000, 2005; Corver, 2013). Rather than depending on a specific sequence of part-of-speech tags (Cinque, 1999; Rizzi, 1997), grammatical structure operates as a layered projection that expands from a base lexical unit into a fully realized syntactic phrase. The tree in Figure 1 central panel illustrates this concept. The lexical item *water* occupies the base of the projection ( $N_0$ ). As the phrase builds upward, successive layers ( $N_1$  through  $N_3$ ) introduce increasingly abstract grammatical functions, subordinating the lexical head under functional heads (like *glass of*, and *that*). All intermediate projects, i.e.  $N_1P$ ,  $N_2P$ , and  $N_3P$ , are all *noun-y* phrases.

Third, to govern how these functional layers are linearly ordered, we adopt the Final-over-Final Constraint (Biberauer et al., 2014). Linguistic studies show that even though human languages often have ‘disharmonic’ word orders, the structural

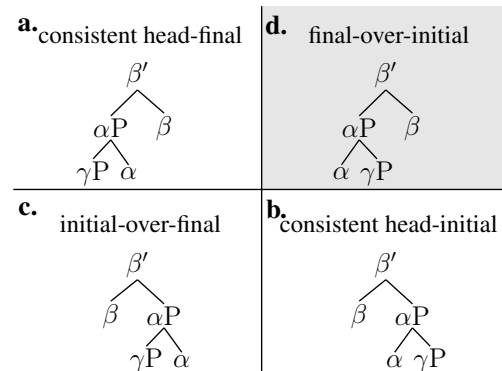


Figure 2: Illustration of the Final-over-Final Constraint (FOFC) on branching directionality. The constraint permits consistently harmonic structures, including head-final (a) and head-initial (b) configurations, as well as initial-over-final structures (c). However, FOFC explicitly bans final-over-initial configurations (d).

variation is often not random (Greenberg, 1963; Biberauer et al., 2008). FOFC acts as a strong syntactic universal that restricts the geometric shape of phrase structures by enforcing a strict monotonicity on branching directions. As illustrated in Figure 2, FOFC allows consistently harmonic structures (a, b) as well as initial-over-final configurations (c), while explicitly banning initial-over-head phrases (d) where a head-initial phrase is embedded as a non-head within a head-final one. By adopting this constraint directly into our rule schemata, the model is prevented from searching trees with head-over-initial structures, which are unattested in a wide range of human languages.

Importantly, these theoretical principles make no language-specific assumptions. We therefore treat this work as a foundational pilot study, using unsupervised grammar induction as a computational testbed to evaluate the functional consequences of these priors on naturalistic input. We evaluate on English child-directed speech from the CHILDES database (MacWhinney, 2000).

Our empirical evaluation moves beyond basic parsing accuracy to assess both the structural and cognitive plausibility of the induced grammar. In terms of structural recovery, unlabelled F1 scores show that the schema-prior model achieves highest bracketing accuracy (0.7384) using an extremely efficient grammar space of only 91 nonterminals, compared to the massive capacity scaling required by dense neural baselines such as Compound PCFG (Kim et al., 2019) and TN-PCFG (Yang et al., 2021b). More importantly, this structural restriction forces the model to induce linguistically and developmentally coherent representations. When evaluated across category purity, construction recall, and acquisition gradients, the model achieves near-perfect morpheme attachment (0.99) and reliably isolates closed-class functional structures from open-class content words, accurately mirroring human developmental patterns. Finally, component ablations validate our theoretical design, showing that PT drives category purity, cross-domain composition enables referential structure, and directionality constraints ensure stable hierarchical composition.

## 2 Related Work

### 2.1 Unsupervised Grammar Induction

Grammar induction (GI) learns the latent hierarchical structure of language from unannotated data. Computationally, this requires defining a hypothesis space of grammatical rules and an inference algorithm to estimate it. There are three main ways to specify this hypothesis space, as illustrated in Figure 1: exact symbolic grammars, dense grammars, and rule supersets. The majority of prior work falls under the dense grammar paradigm. Take scalar parameterization as a motivating example — that is, parameterizing rules as discrete scalars: assigning  $N$  preterminals to  $M$  unique tokens induces an  $N \times M$  parameter matrix for lexical rules, while for binary phrasal rules, given  $L$  phrasal nonterminals, the number of possible rules grows as  $L \times (L + N) \times (L + N)$  (one nonterminal on the left-

hand side combined with any pair of nonterminals on the right-hand side). A fundamental challenge for such dense grammars is its search space, if all the parameters are learned as independent scalars (Pate and Johnson, 2016).

Modern neural GI bypasses the lexical sparsity through distributed representations, sharing information across the vocabulary. However, structurally, they operate over a fully connected, dense rule space where any combination of nonterminals is initially permissible (Kim et al., 2019; Drozdov et al., 2019; Zhu et al., 2020; Yang et al., 2021b). Other approaches include bi-lexicalized grammar (Yang et al., 2021a), character-based modeling (Jin et al., 2021a), and discontinuous parsing (Yang et al., 2022, 2023). While these models achieve high unlabelled bracketing accuracy, their unconstrained structural density comes at the cost of interpretability and high data demand for training. They often fail to identify cognitively meaningful syntax, and their lack of explicit control can lead to ambiguous optimization (Park and Kim, 2024).

**Superset Grammars** Another line of work treats grammar induction as a process of inference over an intentionally restricted rule superset. These models define a broad but constrained set of permissible structures, allowing the learning algorithm to prune unneeded rules (Johnson, 2013). For example, CCG-based approaches restrict the grammar space through a bounded linguistically structured formalism (Bisk and Hockenmaier, 2012b,a, 2015). Other models inject unambiguous structural biases (Tu et al., 2012), or combine strict categorical constraints with modern neural parameterization to process raw lexical emissions efficiently (Clark and Schuler, 2023). Building on this, our neuro-symbolic model adopts the superset strategy. We adopt neural parameterization to mitigate the lexical sparsity issue, but explicitly restrict the structural hypothesis prior using UG-inspired rule schemata.

### 2.2 Modeling Language Acquisition

GI is a standard computational proxy for syntactic acquisition (Buttery, 2004; Kwiakowski et al., 2012). By framing language learning as a process of statistical inference (Johnson, 2013), it enables controlled experiments on what types of abstract structural biases make complex grammars learnable from realistic, stimulu-poor linguistic inputs (Chomsky, 1965; Gold, 1967; Crain, 1991; Pullum

and Scholz, 2002; Perfors et al., 2011b).

However, standard GI is necessarily an idealization. Human language learning is deeply interactive and multimodal. Models that learn strictly from strings ignore the semantic, visual, and multimodal cues (Shi et al., 2019; Jin and Schuler, 2020; Zhang et al., 2021). Therefore, the computational task of inducing a static grammar from text should not be treated as equal to the process of acquiring a language (Dupre, 2024).

All models are wrong, but some are useful (Box, 1979). We therefore treat GI not as a fine-grained model of acquisition per se, but as a testbed for the functional consequences of candidate priors under naturalistic input.

### 2.3 Linguistic Universals as Inductive Biases

Computational methods often use linguistic knowledge to regularize language processing, with both strict architectural universals and language-specific generalizations (O’Horan et al., 2016; Bender, 2016). While explicitly incorporating typological features can result in performance gains (Ponti et al., 2019), these benefits are not always consistent (Bjerva, 2024). Much of the work treats cross-linguistic generalizations as information for parameter sharing across languages (Naseem et al., 2012; Ammar et al., 2016; Scholivet et al., 2019). When universals are used explicitly to constrain a structure, it typically did so via soft preferences, such as universal dependency templates (Cohen and Smith, 2009; Naseem et al., 2010), structural sparsity regularization (Ganchev et al., 2010; Gillenwater et al., 2010), or cognitive bounds (Jin et al., 2018, 2021b). These methods encode universals as auxiliary penalties. Very few models enforce hard structural bounds, such as early Principles-and-Parameter frameworks (Gibson and Wexler, 1994) or left-corner stack depth restrictions (Noji et al., 2016). In this work, we operationalize formalized syntactic principles directly into our rule generator. We encode these linguistic universals as hard structural constraints rather than soft probabilistic penalties, empirically test them for unsupervised induction.

## 3 Neuro-Symbolic PCFG Induction

Our neuro-symbolic framework consists of two co-operating components. The symbolic component is a structurally constrained context-free grammar generated by applying linguistic universals through

a rule schema, detailed in Section 4. The neural component then assigns probabilities to these productions, as described in Section 3.2

### 3.1 Probabilistic Context-Free Grammars

A probabilistic context-free grammar is a tuple  $G = (\mathcal{S}, \mathcal{N}, \mathcal{P}, \Sigma, \mathcal{R}, \theta)$ , with start symbol  $\mathcal{S}$ , non-terminals  $\mathcal{N}$ , preterminals  $\mathcal{P}$ , vocabulary  $\Sigma$ , productions  $\mathcal{R}$ , and rule probabilities  $\theta$ . A production takes one of three shapes, a unary rule  $A \rightarrow B$ , a binary rule  $A \rightarrow BC$ , and a lexical rule  $T \rightarrow w$  with  $T \in \mathcal{P}$  and  $w \in \Sigma$ . Each entry of  $\theta$  is normalized over the productions that share a left-hand side nonterminal. The probability of a tree is the product over its productions, and the probability of an unannotated utterance is the sum over the trees that yield the observed string.

### 3.2 Estimating Rule Probabilities

While classical grammar induction updates independent scalar parameters using expectation maximization from a uniform start, our framework adopts the Neural PCFG parameterization (Kim et al., 2019). Under this parameterization, each left-hand side nonterminal  $A$  is associated with a learned embedding  $e_A$ , and each specific production  $r$  carries a learned vector representation  $g_r$ . The probability of a rule is computed by passing the concatenated representations through a multi-layer perceptron (MLP) and applying a softmax across all rules within the left-hand side:

$$p_\theta(r|A) = \text{softmax}_A(\text{MLP}([e_A; g_r]))$$

With shared embeddings, the neural estimator allows rare productions to borrow statistical strength from related syntactic structures.

To optimize these distributed parameters, the model is trained to maximize the marginal log-likelihood of the unannotated sentences in the training corpus. For a given sentence  $x$ , the marginal probability is defined as the sum over the probabilities of all valid parse trees  $t$  that yield the sentence:

$$p_\theta(x) = \sum_{t \in \mathcal{T}(x)} p_\theta(t)$$

where  $\mathcal{T}(x)$  represents the set of all possible trees for  $x$  permitted by the restricted structural superset  $\mathcal{R}$ , and  $p_\theta(t)$  is the product of its constituent rule probabilities. Because calculating this sum directly is intractable over an exponentially large tree space,

$p_\theta(x)$  and its gradients are computed efficiently using the inside algorithm.

### 3.3 Efficient Inference and Training

To optimize the marginal likelihood efficiently, we decouple the symbolic structural generation from the neural parameter updates, as summarized in Algorithm 1. Because our rule schemata strictly restrict the hypothesis space  $\mathcal{R}$  prior to learning, we can parse each unannotated utterance exactly once before neural training begins.

The symbolic parser processes each utterance into a packed derivation forest, a compact data structure that implicitly represents the exponentially large set of valid trees  $\mathcal{T}(x)$ . During training, the estimator scores the rules and runs the inside algorithm directly over these fixed forests to compute the marginal log-likelihood and its gradients. This exact recurrence allows gradients to backpropagate directly into the neural scorer and its embeddings. At evaluation, a Viterbi pass over the forest extracts the most probable hierarchical derivation.

---

#### Algorithm 1 Neuro-Symbolic PCFG Induction

---

**Require:** utterances  $\{w^{(m)}\}_{m=1}^M$ , schema configuration  $\Phi$

- 1:  $\mathcal{R} \leftarrow \text{EMITGRAMMAR}(\Phi, \Sigma)$
- 2: **for** each  $w^{(m)}$  **do**
- 3:    $F^{(m)} \leftarrow \text{PARSE}(\mathcal{R}, w^{(m)}) \triangleright$  packed forest
- 4: **end for**
- 5: **repeat**
- 6:   **for** each parsed  $w^{(m)}$  **do**
- 7:     score rules via grouped softmax
- 8:      $\ell^{(m)} \leftarrow \text{INSIDE}(F^{(m)}, \log p_\theta)$
- 9:   **end for**
- 10:   update the scorer by gradient ascent on  $\sum_m \ell^{(m)}$
- 11: **until** convergence
- 12: **return** decoded MAP brackets and weights

---

## 4 UG Principles as Rule Schemata

The symbolic prior in this framework is a rule schema. A schema is a deterministic function that maps a theoretical configuration and a vocabulary to a set of context-free productions:

$$\text{CONFIG}(\Phi, \Sigma) \rightarrow \mathcal{R} \quad (1)$$

where  $\Phi$  holds the structural settings and  $\Sigma$  is the observed vocabulary. In practice, CONFIG acts as

a combinatorial engine by taking abstract configurational templates and generates exhaustive set of permissible rules  $\mathcal{R}$ . Table 1 summarizes how each assumptions constrains the rule space.

### 4.1 Prototypes of Part-of-Speech

The first assumption, derived from the prototype theory of parts-of-speech (PT), is that grammar can be organized around three basic functions of language: referring, predicating, and modifying. A referential expression introduces an entity. A predicative expression says something about an entity or event. A modifying expression restricts the interpretation of another expression. These functions are more abstract than part-of-speech labels, but they impose more structure than an unlabeled set of nonterminals. The schema represents these functions as primitive domains,

$$D \in \{R, P, M\}.$$

### 4.2 Extended Projection

The second assumption is that each functional domain is internally organized as a layered projection.

To formalize this, the schema assigns a maximum depth  $N_D$  to each domain  $D$ . The domain is populated by a sequence of heads: the lexical head  $D_0$  and the functional heads  $D_1, \dots, D_{N_D}$ . Any head within this hierarchy can emit a terminal word from the vocabulary:

$$D_i \rightarrow w, \text{ for all } w \in \Sigma$$

To begin building the phrasal spine, the lexical head first projects into a base phrase via a unary rule  $D_0P \rightarrow D_0$ . Once this base phrase is built, the schema builds the hierarchy by allowing a higher functional head  $D_i$  to take a lower phrase  $D_jP$  (where  $j < i$ ) as its complement, forming a new extended phrase  $D_iP$ . To allow for flexibility in linear word order, this binary projection rule is emitted in both head-initial and head-final configurations:

$$D_iP \rightarrow D_i D_jP \quad \text{or} \quad D_iP \rightarrow D_jP D_i.$$

The FOFC constraint in Section 4.4 determines which of these directions is available at each level.

To constrain the search space while maintaining robustness, the gap between levels ( $i - j$ ) is bounded by a predefined skip-window parameter. This ensures that the model learns a structured functional spine without needing to rigidly enforce every single intermediate layer for every utterance.

| Principle                    | Schematic Effect                                                                                           | Rules in Example Grammar                                                                                                 |
|------------------------------|------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------|
| <b>Prototype Theory (PT)</b> | Creates primitive domains (R, P, M) by defining base phrasal rules and open lexical emissions.             | $R_0 P \rightarrow R_0$<br>$P_0 P \rightarrow P_0$<br>$D_i \rightarrow w$ for all $w, (D, i)$                            |
| <b>Extended Projection</b>   | Builds hierarchical spines within domains, connecting lower phrases to higher functional heads.            | $R_1 P \rightarrow R_0 P \quad R_1$<br>$R_2 P \rightarrow R_2 \quad R_1 P$                                               |
| <b>Cross-Domain</b>          | Links domains via predicative saturation and optional modifier attachment.                                 | $P_0 P \rightarrow P_0 \quad R_2 P \mid R_2 P \quad P_0$<br>$X_i P \rightarrow X_i P \quad M_0 P \mid M_0 P \quad X_i P$ |
| <b>FOFC</b>                  | Enforces monotonic branching (e.g., final then initial) and explicitly bans final-over-initial structures. | $R_1 P \rightarrow R_0 P \quad R_1$ (head-final)<br>$R_2 P \rightarrow R_2 \quad R_1 P$ (head-initial)                   |
| <b>Start</b>                 | Allows ROOT to rewrite to the projections of domains at any level.                                         | $\text{ROOT} \rightarrow R_i P$ for all $i \in [0, 2]$<br>$\text{ROOT} \rightarrow P_i P$ for all $i \in [0, 1]$         |

Table 1: UG-inspired principles, their schematic effects, and example rules emitted at depths  $R = 2$ ,  $P = 1$ ,  $M = 0$  with skip window 1.  $R$ ,  $P$ , and  $M$  represent the referential, predicative, and modifying domains.

This matters because the same schema should cover both rich phrases, such as *that one glass of water*, and shorter phrases that express only part of the projected structure.

### 4.3 Cross-Domain Composition

While extended projection builds the internal hierarchical spine of a single domain, cross-domain composition defines how the referential, predicative, and modifying domains interact to form complete clauses.

The predicative saturation rule is the main cross-domain rule,

$$P_i P \rightarrow P_k R_j P \mid R_j P P_k.$$

This rule encodes the idea that predication is relational. A predicative base is licensed through composition with a referential expression. The skip parameter allows saturation to occur even if the referential phrase is not maximally projected.

Unlike predicative saturation, a modifying phrase  $M_k P$  can attach to any existing phrase  $X_i P$  (where  $X \in \{R, P, M\}$ ) without altering the core category or projection level of the phrase it modifies. This recursive attachment is defined as:

$$X_i P \rightarrow X_i P M_{d_M} P \mid M_{d_M} P X_i P,$$

where  $X \in \{R, P\}$ .

### 4.4 FOFC and Head Directionality

The rules generated by extended projection and cross-domain composition permit both head-initial and head-final branching. However, allowing unconstrained branching at every layer of the hierarchy expands the hypothesis space and permits

structures that are typologically unattested. To restrict this geometric space, the schema applies the FOFC as a strict structural filter.

For each domain  $D$ , the schema sets a split point along the projection spine. On one side of the split, the schema emits projection rules in one direction. On the other side, it emits projection rules in the other direction. For example, a projection structure may begin with head-final rules,

$$D_i P \rightarrow D_j P D_i,$$

and then switch to head-initial rules,

$$D_i P \rightarrow D_i D_j P.$$

The reverse is also possible. What the schema excludes is repeated alternation across layers. The constraint therefore prunes directionally inconsistent projection rules.

### 4.5 Compositional Construction

The full configuration is

$$\Phi = (\mathcal{D}, \{d_D\}_D, k, \{s_D\}_D, \{b_D\}_D, \mathcal{E}, \mathcal{R}^{\text{root}}), \quad (2)$$

with  $\mathcal{D} = \{R, P, M\}$ , projection depths  $\{d_D\}_D$ , skip parameter  $k$ , FOFC split points  $\{s_D\}_D$ , transition bands  $\{b_D\}_D$ , embedding set  $\mathcal{E}$  fixed by PMR, and root targets

$$\mathcal{R}^{\text{root}} \subseteq \bigcup_{D \in \{R, P\}} \bigcup_{i=0}^{d_D} \{D_i P\}.$$

### Example Derivation

To see how the schema’s constraints interact to form a complete tree, consider an example derivation for the clause *the dog s chase ed*, illustrated in Figure 3.

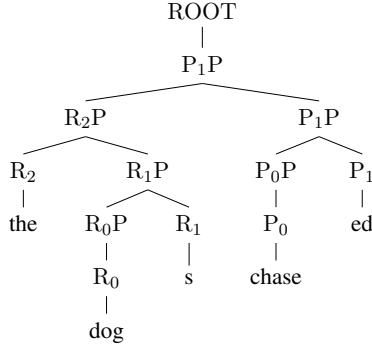


Figure 3: Derivation of the clause *the dog s chase ed*.

First, the schema constructs the referential subject. The lexical head “dog” ( $R_0$ ) forms a base phrase ( $R_0P$ ) and projects upward by taking the plural functional affix “s” ( $R_1$ ) and the determiner “the” ( $R_2$ ) to complete the fully saturated referential spine ( $R_2P$ ). In parallel, the schema constructs the predicative spine. The lexical verb “chase” ( $P_0$ ) projects into a base phrase ( $P_0P$ ) and takes the past tense affix “ed” ( $P_1$ ) as a functional head, forming  $P_1P$ .

Finally, cross-domain composition unites the two spines. The resulting structure represents a typologically sound, hierarchically strict derivation that perfectly respects the PT, EPP, and FOFC priors, all without relying on a single language-specific part-of-speech tag.

## 5 Experiments

We evaluate on English child-directed speech using the morpheme-segmented CHILDES subset from [Marcheva et al. \(2025\)](#) with 100,000 training utterances and 12,615 test utterances. Our schema-prioried PCFG uses projection depth  $d = 20$ , skip window  $k = 5$ , modifier depth of 2, and an FOFC split at level 2. We compare against state-of-the-art neural baselines TN-PCFG ([Yang et al., 2021b](#)) and Compound PCFG ([Kim et al., 2019](#)). Results are averaged over three random seeds unless otherwise stated. Full training and hardware details are provided in Appendix A.

### 5.1 Evaluation Metrics

We evaluate the induced trees against gold annotations across four dimensions, measuring both raw structural accuracy and linguistic interpretability.

To evaluate overall syntactic structure, we measure unlabelled F1 and recall for constituent boundary recovery. With reference spans  $G$  and induced spans  $I$ , F1 is the harmonic mean of  $|G \cap I|/|I|$

and recall  $|G \cap I|/|G|$ , computed per sentence and averaged across the test split. We include a trivial right-branching baseline as a performance floor and a binary ceiling as the maximum attainable score for strictly binary trees against reference trees that include higher-arity nodes.

Beyond basic bracketing, we evaluate the interpretability of the induced grammars through category purity, which measures whether words with similar syntactic roles map to the same induced preterminal. Each word is assigned its modal preterminal,  $\hat{P}(w) = \arg \max_P p(P | w)$ , from lexical-rule weights. For each reference POS, top-cluster concentration is the share of its tokens assigned to the most common induced preterminal. We report pooled scores for function POS and content POS. Acquisition-based scores apply the same top-cluster concentration measure to nominal-word slices defined by concreteness ratings ([Brysbaert et al., 2014](#)) and age-of-acquisition ratings ([Kuperman et al., 2012](#)). Finally, construction recall evaluates the model’s ability to identify theory-relevant multi-word spans. We report recall for determiner plus noun, auxiliary or copula plus verb, and subject-verb-object constructions.

### 5.2 Results

| System         | NT   | F1                     | Recall                 |
|----------------|------|------------------------|------------------------|
| RB floor       | -    | 0.66                   | 0.75                   |
| Binary ceiling | -    | 0.85                   | 1.00                   |
| Schema (ours)  | 91   | <b>0.74</b> $\pm$ 0.00 | <b>0.84</b> $\pm$ 0.00 |
| C PCFG         | 31   | 0.43 $\pm$ 0.05        | 0.44 $\pm$ 0.06        |
| C PCFG         | 61   | 0.43 $\pm$ 0.03        | 0.44 $\pm$ 0.04        |
| C PCFG         | 121  | 0.64 $\pm$ 0.00        | 0.72 $\pm$ 0.00        |
| TN-PCFG        | 512  | 0.51 $\pm$ 0.06        | 0.54 $\pm$ 0.07        |
| TN-PCFG        | 2048 | 0.53 $\pm$ 0.04        | 0.56 $\pm$ 0.06        |

Table 2: English constituency results by symbol count.

We first evaluate whether the UG-inspired schemata improves induction accuracy compared to unconstrained dense grammars, reported in Table 2. On the English dataset, the schema-prioried model achieves a F1 of 0.74 using a highly restricted set of only 91 nonterminals. The neural baselines show a heavy reliance on capacity scaling and fail to match this performance. This extreme parameter efficiency confirms that the UG-inspired schemata successfully prune improbable trees from the search space, allowing the model to learn latent hierarchical structure.

| Family  | Purity      |             | Concreteness |             | AoA         |      | Morph       | Construction recall |             |             |
|---------|-------------|-------------|--------------|-------------|-------------|------|-------------|---------------------|-------------|-------------|
|         | Func        | Cont        | hi           | lo          | early       | late |             | DET+N               | AUX+V       | SVO         |
| Schema  | <b>0.84</b> | <b>0.71</b> | 0.93         | <b>0.74</b> | <b>0.78</b> | 0.77 | <b>0.99</b> | 0.53                | <b>0.87</b> | 0.52        |
| C PCFG  | 0.67        | 0.67        | 0.93         | 0.52        | 0.55        | 0.75 | 0.38        | 0.55                | 0.64        | 0.55        |
| TN-PCFG | 0.25        | 0.55        | 0.22         | 0.18        | 0.55        | 0.70 | 0.84        | <b>0.94</b>         | 0.25        | <b>0.65</b> |

Table 3: English category purity, acquisition gradients, morpheme attachment accuracy, and construction recall.

### 5.3 Category Interpretability and Developmental Plausibility

Beyond raw bracketing, Table 3 suggests that the schema’s priors meaningfully shape how the model organizes the lexicon. The schema-prior grammar is the only model whose induced categories follow all three expected asymmetries: function words cluster more tightly than content words (0.84 vs 0.71), concrete nouns more tightly than abstract nouns (0.93 vs 0.74), and early-acquired words more tightly than late-acquired words (0.78 vs 0.77). The neural baselines show less coherent structure. Both models cluster late-acquired words more cleanly than early-acquired ones, suggesting that without explicit structural priors to separate lexical bases from functional layers, purely distributional learning treat words too uniformly to capture the developmental reality of the lexicon.

Furthermore, these representational differences directly affect how the models handle functional composition. The schema’s near-perfect morpheme attachment accuracy (0.99) strongly suggests that it reliably groups the morphemes with their corresponding lexical heads. This extends to multi-word functional structures, where our model achieves the highest AUX/COP+V recall (0.87).

In contrast, the neural baselines show a stronger bias towards local lexical spans. TN-PCFG achieves the highest recall for DET+N and SVO constructions, but struggle to organize functional elements with low AUX+V recall. The Compound PCFG has a similar difficulty with functional structures, particularly in its low morpheme attachment accuracy. This shows that while dense grammar effectively captures linear lexical dependencies, the schema’s priors are better suited for identifying hierarchical functional layers.

### 5.4 Capacity Scaling

To understand how nonterminal vocabulary size affects performance, Table 4 compares model performance across different capacity thresholds.

The schema-prior grammar gains most of its

| Family  | NT   | F1          |
|---------|------|-------------|
| Schema  | 67   | 0.67        |
| Schema  | 91   | <b>0.74</b> |
| Schema  | 136  | 0.74        |
| C-PCFG  | 31   | 0.43        |
| C-PCFG  | 61   | 0.43        |
| C-PCFG  | 121  | 0.64        |
| TN-PCFG | 512  | 0.51        |
| TN-PCFG | 2048 | 0.52        |

Table 4: Capacity scaling on English.

bracketing accuracy before the grammar becomes large. Increasing projection depth beyond the mid-sized setting gives little return, which suggests that the useful constraint comes from the organization of the rule space, not from adding nonterminals.

The neural baselines show the opposite pattern. Their F1 improves as capacity grows, but the improvement requires a much larger nonterminal inventory. Capacity helps the neural baselines approximate useful structure, while the schema builds part of that structure into the hypothesis space before estimation begins.

### 5.5 Component Ablation

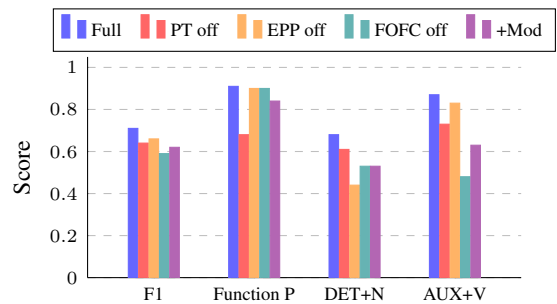


Figure 4: Component ablation.

The ablation study tests which part of the schema is responsible for each structural effect (Figure 4). The full model projects the referential and predicative domains by default. In the PT-off condition, the three-domain functional set is removed. EPP-off condition removes hierarchical projection structures. In the FOFC-off condition, projection

direction is no longer constrained. The +Mod condition extends projected layers to the modifier.

The results show a clear division of labor among the principles.

Prototypes of POS primarily drives category interpretability, as expected. Removing the three-domain separation collapses function word purity and AUX+V recall. This shows that the drop in F1 is primarily from the model losing its ability to separate functional layers from content words.

Disabling EP hierarchy specifically degrades DET+N recovery without affecting function purity. This suggests that the low-level referential constituents heavily relies on argument requirement, rather than on the three-domain inventory alone.

FOFC enforces stable composition. When projection direction is left unconstrained, the model can still identify functional words, but fails to compose it correctly, resulting in a plummeting F1 and AUX+V recall. This shows that functional categories require stable directional rules to form accurate hierarchical structures.

Finally, the +Mod condition shows that adding projected structure is not always beneficial. Extending hierarchical projection to the modifying domain lowers both F1 and construction recall. The added modifier spine competes destructively with core functional projections.

## 5.6 Category-level Recovery

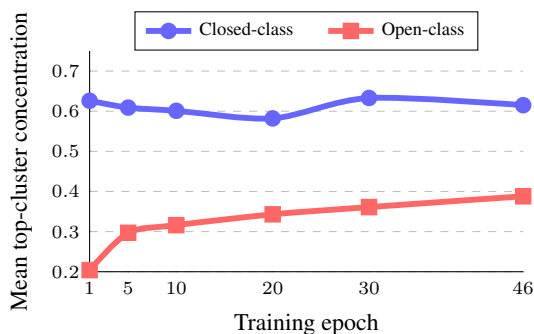


Figure 5: Category concentration over training.

To further investigate how the schema acquires different word types, we track how word categories form during training. Figure 5 shows the learning curves for closed-class functional words and open-class content words. Closed-class functional words, such as determiners and auxiliaries, achieve high concentration almost immediately, stabilizing by the first epoch and remaining consistent throughout training. Open-class content words, on the other

hand, begin with diffuse, low-concentration clusters and converge only gradually, never reaching the purity levels of the functional words.

The learning curve directly reflects the inductive biases of the schema. Because the grammar rules require specific functional elements to construct valid hierarchical trees, the model is forced to rapidly identify these closed-class items to parse the training data effectively. Open-class words, which usually occupy flexible positions within these trees, lack the same structural constraints.

This curve indicates that the schema does not attempt to solve general POS induction. Instead, it prioritizes the rapid acquisition of the grammatical markers necessary for syntactic parsing.

## 6 Conclusion

This paper investigates whether syntactic theories can be operationalized as explicit inductive biases for unsupervised grammar induction. We translated three Universal Grammar hypothesis, namely Prototype Theory (PT), Extended Projection (EPP), and the Final-over-Final Constraint (FOFC), into rule schema. By using a neural estimator to learn probabilities over this restricted grammar space, our design cleanly separates the symbolic prior from the statistical learning algorithm.

The empirical results demonstrate that explicitly constraining the hypothesis space fundamentally changes the nature of what the model learns. Rather than relying on the massive capacity scaling required by dense neural baselines, our schema-prior model achieves superior structural recovery using a highly efficient set of just 91 nonterminals. More importantly, the model induces highly interpretable, developmentally plausible categories. It naturally isolates closed-class functional material from open-class content words and achieves near-perfect morphological attachment, structural features that purely distributional representations fail to capture. Furthermore, our ablation studies show a clear division of labor among the theoretical principles.

Ultimately, our findings support a strong computational interpretation of syntactic universals. The proposed priors do not merely shrink the search space to make learning faster. Rather, they actively shape the topology of the space, ensuring that grammars with layered, linguistically meaningful functional structures are the most natural representations to recover from naturalistic input.

## 7 Limitations

This study is limited in scope. The experiments use English child-directed speech, which provides a controlled testbed with reference parses and established baselines, but does not establish cross-linguistic generality. The evaluation also depends on CTB-derived reference trees, which are useful for consistent comparison but do not directly encode the syntactic theory that motivates the schema. The model also learns from strings alone and abstracts away from semantic, pragmatic, prosodic, and interactional cues available to children. The results show that UG-inspired rule schemata can improve grammar induction under controlled input conditions, but they do not show that the proposed schema is psychologically real or sufficient as a model of child language acquisition.

## 8 Acknowledgements

This paper reports on research partly supported by Cambridge University Press & Assessment.

We thank the anonymous reviewers and the area chair for their useful feedback and suggestions, which greatly improved the manuscript.

## References

- Waleed Ammar, George Mulcaire, Miguel Ballesteros, Chris Dyer, and Noah A. Smith. 2016. [Many languages, one parser](#). *Transactions of the Association for Computational Linguistics*, 4:431–444.
- Emily M Bender. 2016. Linguistic typology in natural language processing. *Linguistic Typology*, 20(3):645–660.
- Robert C Berwick, Paul Pietroski, Beracah Yankama, and Noam Chomsky. 2011. Poverty of the stimulus revisited. *Cognitive science*, 35(7):1207–1242.
- Theresa Biberauer, Anders Holmberg, and Ian Roberts. 2008. Structure and linearization in disharmonic word orders. In *Proceedings of the 26th West Coast Conference on Formal Linguistics*, pages 96–104, Somerville, MA. Cascadia Proceedings Project.
- Theresa Biberauer, Anders Holmberg, and Ian Roberts. 2014. A syntactic universal and its consequences. *Linguistic Inquiry*, 45(2):169–225.
- Yonatan Bisk and Julia Hockenmaier. 2012a. [Induction of linguistic structure with Combinatory Categorical Grammars](#). In *Proceedings of the NAACL-HLT Workshop on the Induction of Linguistic Structure*, pages 90–95, Montréal, Canada. Association for Computational Linguistics.
- Yonatan Bisk and Julia Hockenmaier. 2012b. Simple robust grammar induction with combinatory categorical grammars. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 26, pages 1643–1649.
- Yonatan Bisk and Julia Hockenmaier. 2015. [Probing the linguistic strengths and limitations of unsupervised grammar induction](#). In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1395–1404, Beijing, China. Association for Computational Linguistics.
- Johannes Bjerva. 2024. [The role of typological feature prediction in NLP and linguistics](#). *Computational Linguistics*, 50(2):781–794.
- G.E.P. Box. 1979. [Robustness in the strategy of scientific model building](#). In ROBERT L. LAUNER and GRAHAM N. WILKINSON, editors, *Robustness in Statistics*, pages 201–236. Academic Press.
- Marc Brysbaert, Amy Beth Warriner, and Victor Kuperman. 2014. Concreteness ratings for 40 thousand generally known english word lemmas. *Behavior research methods*, 46(3):904–911.
- Paula Buttery. 2004. [A quantitative evaluation of naturalistic models of language acquisition; the efficiency of the triggering learning algorithm compared to a categorical grammar learner](#). In *Proceedings of the Workshop on Psycho-Computational Models of Human Language Acquisition*, pages 3–10, Geneva, Switzerland. COLING.
- Noam Chomsky. 1965. *Aspects of the Theory of Syntax*. MIT press.
- Noam Chomsky. 1993. *Lectures on government and binding: The Pisa lectures*. 9. Walter de Gruyter.
- Noam Chomsky. 2005. Three factors in language design. *Linguistic inquiry*, 36(1):1–22.
- Noam Chomsky. 2014. *The minimalist program*. MIT press.
- Guglielmo Cinque. 1999. *Adverbs and functional heads: A cross-linguistic perspective*. Oxford University Press.
- Alexander Clark. 2002. Unsupervised language acquisition: Theory and practice. *arXiv preprint cs/0212024*.
- Alexander Clark, Chris Fox, and Shalom Lappin. 2012. *The handbook of computational linguistics and natural language processing*. John Wiley & Sons.
- Christian Clark and William Schuler. 2023. [Categorical grammar induction from raw data](#). In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 2368–2379, Toronto, Canada. Association for Computational Linguistics.

- Shay B. Cohen and Noah A. Smith. 2009. [Variational inference for grammar induction with prior knowledge](#). In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, pages 1–4, Suntec, Singapore. Association for Computational Linguistics.
- Norbert Corver. 2013. Lexical categories and (extended) projection. *The cambridge handbook of generative syntax*, pages 353–424.
- Stephen Crain. 1991. Language acquisition in the absence of experience. *Behavioral and Brain Sciences*, 14(4):597–612.
- William Croft. 2001. *Radical construction grammar: Syntactic theory in typological perspective*. OUP Oxford.
- Andrew Drozdov, 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, Minnesota. Association for Computational Linguistics.
- Gabe Dupre. 2024. [Acquiring a language vs. inducing a grammar](#). *Cognition*, 247:105771.
- Kuzman Ganchev, Joao Graça, Jennifer Gillenwater, and Ben Taskar. 2010. Posterior regularization for structured latent variable models. *The Journal of Machine Learning Research*, 11:2001–2049.
- Edward Gibson and Kenneth Wexler. 1994. Triggers. *Linguistic inquiry*, 25(3):407–454.
- Jennifer Gillenwater, Kuzman Ganchev, João Graça, Fernando Pereira, and Ben Taskar. 2010. [Sparsity in dependency grammar induction](#). In *Proceedings of the ACL 2010 Conference Short Papers*, pages 194–199, Uppsala, Sweden. Association for Computational Linguistics.
- E Mark Gold. 1967. Language identification in the limit. *Information and control*, 10(5):447–474.
- Joseph H. Greenberg. 1963. Some universals of grammar with particular reference to the order of meaningful elements. In Joseph H. Greenberg, editor, *Universals of Language*, pages 73–113. MIT Press, Cambridge, MA.
- Jane Grimshaw. 2000. Locality and extended projection. *AMSTERDAM STUDIES IN THE THEORY AND HISTORY OF LINGUISTIC SCIENCE SERIES 4*, pages 115–134.
- Jane B. Grimshaw. 2005. *Extended Projection*. CSLI Lecture Notes, No. 151. CSLI Publications, Stanford, CA.
- Lifeng Jin, Finale Doshi-Velez, Timothy Miller, William Schuler, and Lane Schwartz. 2018. [Unsupervised grammar induction with depth-bounded PCFG](#). *Transactions of the Association for Computational Linguistics*, 6:211–224.
- Lifeng Jin, Byung-Doh Oh, and William Schuler. 2021a. [Character-based PCFG induction for modeling the syntactic acquisition of morphologically rich languages](#). In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4367–4378, Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Lifeng Jin and William Schuler. 2020. [Grounded PCFG induction with images](#). In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 396–408, Suzhou, China. Association for Computational Linguistics.
- Lifeng Jin, Lane Schwartz, Finale Doshi-Velez, Timothy Miller, and William Schuler. 2021b. [Depth-bounded statistical PCFG induction as a model of human grammar acquisition](#). *Computational Linguistics*, 47(1):181–216.
- Mark Johnson. 2013. Language acquisition as statistical inference. *The language-cognition interface*, pages 109–134.
- Yoon Kim, Chris Dyer, and Alexander M 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.
- Dan Klein 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, Spain.
- Victor Kuperman, Hans Stadthagen-Gonzalez, and Marc Brysbaert. 2012. Age-of-acquisition ratings for 30,000 english words. *Behavior research methods*, 44(4):978–990.
- Tom Kwiatkowski, Sharon Goldwater, Luke Zettlemoyer, and Mark Steedman. 2012. [A probabilistic model of syntactic and semantic acquisition from child-directed utterances and their meanings](#). In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 234–244, Avignon, France. Association for Computational Linguistics.
- Brian MacWhinney. 2000. *The CHILDES Project: Tools for Analyzing Talk, Volume II: The Database*, 3rd edition. Psychology Press, New York.
- Mila Marcheva, Theresa Biberauer, and Weiwei Sun. 2025. [Profiling neural grammar induction on morphemically tokenised child-directed speech](#). In *Proceedings of the Workshop on Cognitive Modeling*

- and *Computational Linguistics*, pages 47–54, Albuquerque, New Mexico, USA. Association for Computational Linguistics.
- Tahira Naseem, Regina Barzilay, and Amir Globerson. 2012. [Selective sharing for multilingual dependency parsing](#). In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 629–637, Jeju Island, Korea. Association for Computational Linguistics.
- Tahira Naseem, 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. Association for Computational Linguistics.
- Hiroshi Noji, 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, Texas. Association for Computational Linguistics.
- Helen O’Horan, Yevgeni Berzak, Ivan Vulić, Roi Reichart, and Anna Korhonen. 2016. [Survey on the use of typological information in natural language processing](#). In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1297–1308, Osaka, Japan. The COLING 2016 Organizing Committee.
- Jinwook Park and Kangil Kim. 2024. [Structural optimization ambiguity and simplicity bias in unsupervised neural grammar induction](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 15124–15139, Bangkok, Thailand. Association for Computational Linguistics.
- John K Pate and Mark Johnson. 2016. [Grammar induction from \(lots of\) words alone](#). In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 23–32, Osaka, Japan. The COLING 2016 Organizing Committee.
- Amy Perfors, Joshua B. Tenenbaum, and Terry Regier. 2011a. [The learnability of abstract syntactic principles](#). *Cognition*, 118(3):306–338.
- Amy Perfors, Joshua B Tenenbaum, and Terry Regier. 2011b. The learnability of abstract syntactic principles. *Cognition*, 118(3):306–338.
- Edoardo Maria Ponti, Helen O’horan, Yevgeni Berzak, Ivan Vulić, Roi Reichart, Thierry Poibeau, Ekaterina Shutova, and Anna Korhonen. 2019. Modeling language variation and universals: A survey on typological linguistics for natural language processing. *Computational Linguistics*, 45(3):559–601.
- Geoffrey K Pullum and Barbara C Scholz. 2002. Empirical assessment of stimulus poverty arguments. *The linguistic review*, 19(1-2):9–50.
- Luigi Rizzi. 1997. The fine structure of the left periphery. In *Elements of grammar: Handbook in generative syntax*, pages 281–337. Springer.
- Manon Scholivet, Franck Dary, Alexis Nasr, Benoit Favre, and Carlos Ramisch. 2019. [Typological features for multilingual delexicalised dependency parsing](#). 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 3919–3930, Minneapolis, Minnesota. Association for Computational Linguistics.
- Yikang Shen, Zhouhan Lin, Chin-Wei Huang, and Aaron Courville. 2017. Neural language modeling by jointly learning syntax and lexicon. *arXiv preprint arXiv:1711.02013*.
- Haoyue Shi, Jiayuan Mao, Kevin Gimpel, and Karen Livescu. 2019. [Visually grounded neural syntax acquisition](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1842–1861, Florence, Italy. Association for Computational Linguistics.
- Mark Steedman. 1996. *Surface Structure and Interpretation*. MIT Press.
- Mark Steedman. 2001. *The Syntactic Process*. MIT Press.
- Zhaopeng Tu, Yang Liu, Yifan He, Josef van Genabith, Qun Liu, and Shouxun Lin. 2012. Combining multiple alignments to improve machine translation. In *Proceedings of COLING 2012: Posters*, pages 1249–1260.
- Songlin Yang, Roger Levy, and Yoon Kim. 2023. [Unsupervised discontinuous constituency parsing with mildly context-sensitive grammars](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5747–5766, Toronto, Canada. Association for Computational Linguistics.
- Songlin Yang, Wei Liu, and Kewei Tu. 2022. Dynamic programming in rank space: Scaling structured inference with low-rank hmms and pcfgs. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4797–4809.
- Songlin Yang, Yanpeng Zhao, and Kewei Tu. 2021a. [Neural bi-lexicalized PCFG induction](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2688–2699, Online. Association for Computational Linguistics.
- Songlin Yang, Yanpeng Zhao, and Kewei Tu. 2021b. [PCFGs Can Do Better: Inducing Probabilistic Context-Free Grammars with Many Symbols](#). In *Proceedings of the 2021 Conference of the North*

*American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1487–1498, Online. Association for Computational Linguistics.

Songyang Zhang, Linfeng Song, Lifeng Jin, Kun Xu, Dong Yu, and Jiebo Luo. 2021. [Video-aided unsupervised grammar induction](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1513–1524, Online. Association for Computational Linguistics.

Hao Zhu, Yonatan Bisk, and Graham Neubig. 2020. The return of lexical dependencies: Neural lexicalized pcfgs. *Transactions of the Association for Computational Linguistics*, 8:647–661.

## A Experimental Setup and Reproducibility

### A.1 Model Configurations

The primary model evaluated in this work is the schema-prioried Neural PCFG. Across all main experiments, the symbolic rule generator was configured with a maximum projection depth of  $d = 20$  and a skip window of  $k = 5$ . The Final-over-Final Constraint (FOFC) direction switch was permitted at level 2.

We compare our framework against two state-of-the-art dense neural baselines: TN-PCFG (Yang et al., 2021b) and Neural PCFG (Kim et al., 2019). Both baselines were trained on the exact same morpheme-segmented corpus until convergence to ensure a fair comparison.

**Computational Efficiency:** All neural runs use Adam with early stopping on a held-out development split. The compound PCFG baseline adds a linear KL warm-up on its encoder over the first epochs of training. Results are averaged over three random seeds unless stated otherwise. The schema runs converge on a single CPU. The neural baselines are trained on a single A100 GPU.