Noise-tolerant learning as selection among deterministic grammatical hypotheses

Laurel Perkins
Department of Linguistics
University of California Los Angeles
Los Angeles, CA 90095-1543
perkinsl@ucla.edu

Tim Hunter
Department of Linguistics
University of California Los Angeles
Los Angeles, CA 90095-1543
timhunter@ucla.edu

Abstract

Children acquire their language's canonical word order from data that contains a messy mixture of canonical and non-canonical clause types. We model this as noise-tolerant learning of grammars that deterministically produce a single word order. In simulations on English and French, our model successfully separates signal from the noise introduced by non-canonical clause types, in order to identify that both languages are SVO. No such preference for the target word order emerges from a comparison model which operates with a fully-gradient hypothesis space and an explicit numerical regularization bias. This provides an alternative general mechanism for regularization in various learning domains, whereby tendencies to regularize emerge from a learner's expectation that the data are a noisy realization of a deterministic underlying system.

1 Introduction

Children at early stages of language acquisition draw accurate grammatical generalizations from incomplete, immature, and variable representations of their input. For example, infants learn their language's basic word order despite immature abilities to identify clause arguments, and despite non-canonical constructions that disrupt this basic word order (e.g., *wh*-questions, passives) (Hirsh-Pasek and Golinkoff, 1996; Perkins and Lidz, 2020, 2021). This is one of many ways in which learners draw generalizations that are more regular or deterministic than the variable data that they are learning from. What kind of mechanisms allow for learning to abstract away from messiness in (the learner's representation of) the data?

One potential answer emerges from studies of learning in the context of unpredictable variability, for example in the context of acquiring language from non-native speakers. This approach posits a general learning bias to *regularize* inconsistent variability (Hudson Kam and Newport, 2005, 2009; Reali and Griffiths, 2009; Culbertson et al., 2013; Ferdinand et al., 2019). Learners consider hypotheses that closely match the statistical distributions in their input, but in some circumstances they are biased to "sharpen" those distributions, pushing them towards more systematic extremes.

Implicit in this account is a hypothesis space that can accommodate the full variability of the data. For instance, when exposed to an artificial language in which determiners occur inconsistently with nouns, children are equipped to consider that the language allows determiners with any probability, but nonetheless prefer to use particular determiners all of the time or not at all (Hudson Kam and Newport, 2005, 2009). The literature takes this as evidence for a regularization bias operating within a learner's fully-flexible hypothesis space, pushing learners to prefer probabilities closer to 0 or 1 and producing near-categorical learning outcomes. This idea could be applied to the learning of basic word order in infancy— for example, learning that English is canonically SVO. Children who encounter a messy mixture of canonical and non-canonical sentences would be equipped to consider that clause arguments can flexibly occur in multiple orders in the language, but prefer hypotheses that are skewed heavily towards one consistent order.

Here, we explore a different approach. We propose that in certain circumstances, learners face a choice among discrete hypotheses, each of which is deterministic in a way that is incompatible with the full variability of the observed data. Learners expect that their data result from an opaque interaction between (i) one of the deterministic hypotheses currently under consideration, and (ii) various other processes that might introduce "noise" into the data. For a child learning an artificial determiner system, the data might reflect a combination

Proceedings of the Society for Computation in Linguistics (SCiL) 2023, pages 186-198.
Amherst, Massachusetts, June 15-17, 2023

of signal for deterministic rules, and noise coming from unknown grammatical or extra-grammatical processes. For a child learning the syntax of basic clauses, the data reflects a combination of signal for the target language's basic word order and noise introduced by non-canonical sentence types. Regularization emerges when learners are able to successfully identify signal for a deterministic hypothesis within their noisy data (Perkins et al., 2022; Schneider et al., 2020).

We introduce a general computational framework for performing this inference. A learner of the sort we describe below expects that its data are generated by a complex system: a core deterministic component that the learner is attempting to acquire, operating alongside a "noise" component whose properties are currently unknown. Using the case study of basic word order acquisition, we show that our model can learn to separate evidence for a deterministic grammar of canonical word order from the distorting effects of non-canonical noise processes. It does so without knowing ahead of time how much noise there is, or what its properties are. Moreover, we show that our approach fares better in this learning problem than the more common approach to regularization described above. This suggests that in certain domains, successful learning from noisy data is enabled by a hypothesis space comprising restrictive grammatical options.

2 The intuition behind our approach

Suppose that a bag contains coins of two types: Type A coins, which always come up heads, and Type B coins, which all have some single unknown probability θ of coming up heads. We know nothing about how many of each type are in the bag. We observe ten coin flips, producing eight heads and two tails. How many of these flips might we guess came from Type A coins, and how many from Type B coins? There is a wide range of options, including the possibility that all ten flips came from Type B coins; but given the observed skew towards heads, there is a clear intuition that Type A coins were probably responsible for a significant portion of the observations. Why is this?

Under the hypothesis that all ten flips came from Type B coins, eight of those flips would need to come up heads and two to come up tails in order to generate the observed data. Contrast this with the (more intuitively plausible) hypothesis that there were six Type A and four Type B flips. Under this hypothesis, the six Type A flips need to come up heads, which is guaranteed to happen; so, generating the observed data just amounts to having the four Type B flips produce two heads and two tails. This is clearly less "costly" than the first hypothesis's requirement that ten Type B flips produce eight heads and two tails. By positing six Type A flips, six of the heads that we need to generate "come for free"; with only Type B flips, however, we get no such head start.

More precisely, the *likelihood* of the observed data, under the hypothesis that relies on only four Type B flips, is $\binom{4}{2}\theta^2(1-\theta)^2$. Under the hypothesis that leaves all the work to ten Type B flips, this likelihood is $\binom{10}{8}\theta^8(1-\theta)^2$. It is the exponents that matter: the ten-flip likelihood is smaller than the four-flip likelihood whenever $\theta < 0.71$, so for most values of θ. This is one way to understand our intuitive preference for hypotheses that invoke Type A flips. We can make this even more precise by *marginalizing* over θ; see Appendix A for details. These details make clear that *all* that matters about a particular hypothesis is how many Type B flips it must appeal to. We've seen that four Type B flips is better than ten, but two is even better: the very best hypothesis is that there were eight Type A flips and two Type B flips (likelihood $(1-\theta)^2$).

Suppose now that, as well as the bag with two-headed coins and head-tail coins (call this Bag H), there is a bag with two-*tailed* coins and head-tail coins (Bag T). We again see 10 coin flips, 8 heads and 2 tails. We know that they all came from one of the two bags, and we have to guess which one.

We have seen that Bag H makes available "good" explanations of the data, which exploit the presence of two-headed coins to minimize the crucial number of uncertain head-tail flips. With Bag T, however, the available "known outcome" coins produce tails; so the best we can do is to suppose that both of the two observed tails came from the two-tailed coins, and rely on eight uncertain flips to do the rest of the work (likelihood θ^8). Since there is no way for the two-tailed coins to contribute to a good explanation of the observed high proportion of heads, Bag H is a better guess than Bag T.

This choice between Bag H and Bag T will correspond to the choice between competing restricted hypotheses in the learners we describe below. It will be useful to think of this as essentially a choice between the two-headed coin and the two-tailed coin, where either choice (since it's accompanied

by head-tail coins) is embedded in a system that also produces some "noise", i.e. divergences from what would be generated by these core mechanisms alone. When comparing such composite systems, our learner will prefer the one whose core mechanisms predict the skew in the data; this will provide the least costly solution, even though the shared noise possibilities ensure that all the competing systems can account for the data as a whole. And the proposed learner will do this without knowing *a priori* how much of the data is noise (i.e. how much of the data came from the head-tail coins) or what the contribution of noise looks like (i.e. the probability θ of noise contributing a head).

Perkins et al. (2022) applied this approach to model how learners might identify the core transitivity properties of verbs in their language, despite "noise" from non-canonical clause types. This type of noise might arise when a young child encounters an obligatorily-transitive verb in a sentence with a displaced object (e.g., *What did you bring?*) but is unable to parse it as such. By hypothesizing that unknown noise processes cause the data to be a distorted reflection of verbs' core argument-taking properties, their model was able to successfully identify that certain verbs were deterministically transitive and intransitive— for roughly the same reason that Bag H above provides a good explanation for data that does not consist entirely of heads.

Here, the basic syntax we consider generates subjects and objects according to some canonical order (SVO, SOV, etc.), yielding surface strings of verbs and noun phrases. And just like in Perkins et al., unknown grammatical processes— for instance, argument movement or ellipsis— operate alongside this basic syntax, with the result that the observed strings of verbs and noun phrases are a distorted reflection of canonical word order.

3 Applying this to PCFGs

We now turn to situations where a learner's core hypotheses take the form of grammars — specifically, probabilistic context-free grammars (PCFGs). The learner will observe some collection of strings, and in general none of the core grammars under consideration will be consistent with all of the observed strings. One way to apply the idea from above would be to suppose that some of the observed strings were generated by a separate "noise grammar" — just as some of the coin flips above were generated by the head-tail coin. But this would mean that every string is analyzed as either all signal (i.e. informative about the core grammar) or all noise, and so the learner would not be able to extract useful information from subparts of sentences.

Instead, we allow the signal-or-noise choice to be made at a finer-grained level: each derivational step might be contributed either by the core hypothesized grammar or by noise processes. Either way, each step is licensed by a CFG-style rewrite rule; in other words, the noise is itself characterized by particular rules for expanding nonterminals that sit alongside the rules of the core grammar. The overall system therefore consists of rules of two sorts, which we'll call *core rules* and *noise rules*.

Framed slightly more generally: we formulate a generative process for strings that we call a *Mixture PCFG*. A Mixture PCFG uses rules built out of terminal and nonterminal symbols in the manner of a standard PCFG. But whereas defining a standard PCFG involves identifying just a single set of rules, defining a Mixture PCFG involves identifying two sets of rules. For the moment we will simply call them ϕ-rules and ψ-rules, but in the case study below these will correspond to core rules and noise rules respectively. A particular candidate rewriting step, e.g., 'S $\rightarrow$ NP VP', might be included in the set of ϕ-rules, in which case it will have some non-zero probability $\phi_{\text{S}\rightarrow\text{NP VP}}$ associated with it; and independently might be included in the set of ψ-rules, in which case it will have some non-zero probability $\psi_{\text{S}\dashrightarrow\text{NP VP}}$ associated with it. In addition to these ϕ parameters and ψ parameters, a Mixture PCFG has one additional parameter ϵ^A associated with each nonterminal symbol A, which controls the choice between using a ϕ-rule or a ψ-rule to expand an occurrence of A.

To illustrate, an example Mixture PCFG is shown in Fig. 1. In this grammar and all those in the case studies below, NP is deterministically realized as np and V as v; we abstract away from these steps in all the discussion that follows.[1] We write ϕ-rules with standard arrows and ψ-rules with dashed arrows. Notice that the ϕ-probabilities associated with the expansions of a particular nonterminal symbol sum to one, as do the ψ-probabilities. Roughly foreshadowing the grammars we use in the case study below, the ϕ-rules in Fig. 1 encode the basic clause structure of an SVO language, and the ψ-rules generate "noise" that diverges from this canonical word order in various ways.

[1] This is just $\epsilon^{\text{NP}} = \epsilon^{\text{VP}} = 0$ and $\phi_{\text{NP}\rightarrow\text{np}} = \phi_{\text{V}\rightarrow\text{v}} = 1$.

ϕ-rules	ψ-rules	ϵ probabilities
1.0 S $\rightarrow$ NP VP	0.3 S $\dashrightarrow$ VP NP	$\epsilon^{\text{S}} = 0.2$
	0.5 S $\dashrightarrow$ NP S	
	0.2 S $\dashrightarrow$ VP	
0.4 VP $\rightarrow$ V	0.7 VP $\dashrightarrow$ NP V	$\epsilon^{\text{VP}} = 0.3$
0.6 VP $\rightarrow$ V NP	0.3 VP $\dashrightarrow$ NP	

Figure 1: An example Mixture PCFG.

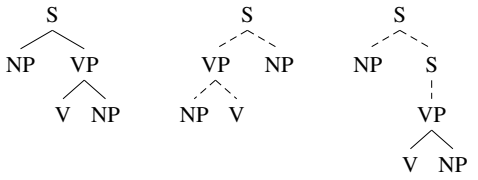

Figure 2: The three possible analyses of np v np (supressing NP$\rightarrow$np and V$\rightarrow$v rewrites).

To calculate the probability of a string under this Mixture PCFG, we sum over all possible ways it can be generated. For the string np v np, for example, there are three possibilities, shown in Fig. 2; solid lines represent expansions using ϕ-rules, and dashed lines expansions using ψ-rules.

The first tree represents one way of generating np v np that uses only ϕ-rules: ϵ^{S} is the probability of using a ψ-rule rather than a ϕ-rule to expand an occurrence of S, and so the probability of expanding the root S node as shown in this first tree is the product of $(1 - \epsilon^{\text{S}})$ and the corresponding ϕ-probability. The probability of the entire tree is the product of two such rewrites, as in (1); similarly, the probability of the second tree is given in (2). The third tree's probability, in (3), uses a more interesting combination of ϕ-rules and ψ-rules.

$$(1)\quad (1-\epsilon^{\text{S}})(\phi_{\text{S}\rightarrow\text{NP VP}}) \times (1-\epsilon^{\text{VP}})(\phi_{\text{VP}\rightarrow\text{V NP}})$$

$$(2)\quad (\epsilon^{\text{S}})(\psi_{\text{S}\dashrightarrow\text{VP NP}}) \times (\epsilon^{\text{VP}})(\psi_{\text{VP}\dashrightarrow\text{NP V}})$$

$$(3)\quad (\epsilon^{\text{S}})(\psi_{\text{S}\dashrightarrow\text{NP S}}) \times (\epsilon^{\text{S}})(\psi_{\text{S}\dashrightarrow\text{VP}}) \times (1-\epsilon^{\text{VP}})(\phi_{\text{VP}\rightarrow\text{V NP}})$$

Using values from Fig. 1, these three trees therefore have probabilities of 0.336, 0.013 and 0.001, respectively; and so the total probability of the string np v np is 0.350.[2]

Although we are restricting attention to PCFGs here, exactly the same approach could be used to formulate "mixture" versions of any kind of probabilistic grammar where the probability of a complex structure is the product of the probabilities of certain local choices (e.g. HMMs or PFSAs). The sampling methods we employ below for inference are compatible with any model where these local choices are expressed as multinomial distributions.

In the learning scenarios modeled below, the learner will have some set of hypotheses to choose from, each of which is represented by a Mixture PCFG such as that in Fig. 1. One of the competitor hypotheses might be represented by a similar Mixture PCFG that has the basic clause structure of an SOV language (rather than SVO) reflected in its ϕ-rules, and has some of the same ψ-rules as Fig. 1. Each of these two hypotheses will therefore generate strings that diverge from the strict SVO or SOV pattern licensed by its particular ϕ-rules. Deciding which of these two Mixture PCFGs provides a better explanation of some observed strings is therefore analogous to the decision between Bag H and Bag T in Section 2, with the ϕ-rules corresponding to the two-headed and two-tailed coins, and the ψ-rules corresponding to the head-tail coins.[3] Just as the decision between Bag H and Bag T could be made by considering (i.e. marginalizing over) all possible values of the unknown weight θ, we can make the decision between competing Mixture PCFGs in a way that considers all possible ϕ, ψ and ϵ values. The logic outlined in Section 2, whereby explanations in terms of core mechanisms that align with skews in the data are preferred, carries over to the case where the core mechanisms are either SVO or SOV word order.

4 Case study: Learning basic word order

We show that the approach of deciding among competing Mixture PCFGs provides a novel solution to the problem of word order acquisition in early development. Children acquire the basic word order of their language from data that contains a large amount of noise. For example, English learners identify that their language is canonically SVO in infancy, before they can identify the processes that produce non-canonical word orders in sentences like *wh*-questions (Hirsh-Pasek and Golinkoff, 1996; Lidz et al., 2017; Perkins and Lidz, 2020, 2021). Many accounts assume that learners have the ability to "filter" non-basic sentences of this sort, ignoring them when drawing

[2]The overall mechanics of a Mixture PCFG can be recast as a single classical PCFG. Specifically: add nonterminals S_ϕ and S_ψ alongside S, and include the rules $\text{S} \rightarrow \text{S}_\phi$ and and $\text{S} \rightarrow \text{S}_\psi$ with probabilities $(1 - \epsilon^{\text{S}})$ and ϵ^{S}, respectively; the subsequent expansions of S_ϕ and S_ψ are determined by the ϕ-rules for S and the ψ-rules for S, respectively. Our implementation in fact works with exactly this classical PCFG.

[3]Bag H is analogous to a Mixture PCFG with $\phi_{S\rightarrow\text{h}} = 1$, $\psi_{S\dashrightarrow\text{h}} = \theta$ and $\psi_{S\dashrightarrow\text{t}} = 1 - \theta$, and ϵ^{S} representing the proportion of head-tail coins in the bag.

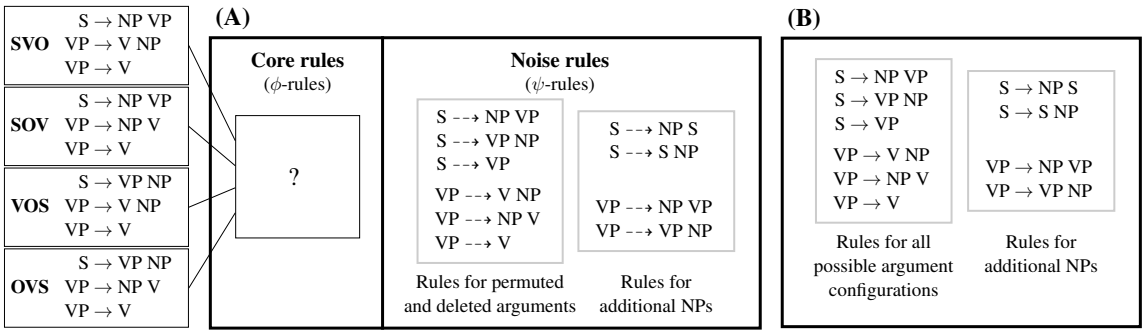

Figure 3: (A) Hypothesis space for our noise-tolerant learner; (B) Fully-flexible learner for comparison.

early syntactic inferences (e.g. Pinker, 1984). But if learners do not yet know what counts as basic, how do they identify which sentence types count as *non*-basic, in order to filter them out (Gleitman, 1990; Perkins et al., 2022)? Our model provides a way to implement the essence of this filtering idea, while avoiding potential issues of circularity.

Our learner's hypothesis space consists of four sets of ϕ-rules and one shared set of ψ-rules, giving rise to the four Mixture PCFGs in Fig. 3A. The ϕ-rules generate the core predicate-argument structure of basic transitive and intransitive clauses, deterministically putting subjects before or after verb phrases and objects before or after verbs. This yields a 4-way choice of canonical word order: SVO, SOV, VOS, OVS.[4] Subjects are are obligatory and objects are optional, reflecting the learner's belief that canonical clauses need subjects. All four grammars share the same set of noise rules, which allow for all permutations and deletions of NP arguments, and for additions of NPs into non-argument positions. The flexibility in the noise rules produces many more possibilities for expanding a given nonterminal than are provided by the core rules, mirroring the asymmetry between restrictive two-headed (and two-tailed) coins and flexible head-tail coins.

Crucially, while the learner's noise rules contain hypotheses about which non-canonical processes might operate in its language, the learner does not know ahead of time the ψ and ϵ probabilities associated with these rules: it does not know which kinds of non-canonical clauses it will encounter, or how frequently. We show that our learner is able to identify the correct Mixture PCFG— the correct combination of core and noise rules— using only the distributions of noun phrases and verbs that a 15-month-old infant might be able to represent. This inference does not require information about underlying clause structure. However, a similar mechanism could be generalized to make use of structural cues from meaning or prosody (Pinker, 1984; Christophe et al., 2008).

Using strings of imperfectly-identified noun phrases and verbs, the learner evaluates the following three questions, corresponding to the ϕ, ψ, and ϵ parameters of its input filter, respectively: (1) What do the data from the core rules look like? (2) What do the data from the noise rules look like? (3) What is the right division into signal vs. noise? For each grammar in its hypothesis space, the learner considers the possible answers to these questions in order to determine how well that grammar explains the data it observes. Comparing across the four grammars, the learner selects the grammar that provides the best explanation.

4.1 Generative model

The model's data consists of a collection $\vec{w}$ of strings, each comprising a single v with any number of satellite np's (i.e., of the form np^* v np^*). The model assumes that these are generated by one of the Mixture PCFGs in its hypothesis space (Fig. 3A), each of which has equal prior probability; the learner is not biased *a priori* in favor of any particular word orders.

Given any particular Mixture PCFG, we can construct an equivalent standard PCFG that defines the same distribution over strings (via some additional nonterminals and unary rules; see Footnote 2 for details). Let $\vec{\theta}^{A_G}$ be the vector of weights of the allowable expansions of a given nonterminal A in this resulting standard PCFG G; the prior over $\vec{\theta}^{A_G}$

[4]We limit our focus to these four word orders because they are the options generated by a 2x2 choice of subject and object position. Natural languages allow more complex argument structure profiles, including canonical orders in which the verb and object are separated (VSO and OSV), or variability from argument-drop or scrambling. How these properties are learned is an important question that we leave for future work.

	English	French
Corpus	Brown: Eve	Lyon
# Children	1	5
Ages	1;6-2;31	1;0-3;0
# Words	81,687	885,334
# Utterances	14,232	182,511

Table 1: Corpora of child-directed English and French

English	French
0.36 np v	0.48 np v
0.20 v	0.21 np v np
0.20 np v np	0.13 v
0.17 v np	0.05 np np v
0.04 np v np np	0.03 np v np np
0.03 v np np	0.03 v np

Table 2: Proportions of most frequent string types

is a Dirichlet distribution with parameters $\vec{\alpha}^{A_G}$. We begin with the assumption that all components $\alpha_i^{A_G}$ are equal to 1, resulting in a uniform prior distribution, i.e. the model considers all possible expansions for A with equal probability.

4.2 Inference

From the observed strings, the model infers the posterior distribution over all grammars in its hypothesis space, $P(G \mid \vec{w})$. Calculating this posterior analytically would require marginalizing over both $\vec{\theta^G}$ and $\vec{t}$— i.e., integrating over the rule weights and summing over all possible trees for a string, for all strings in the data. This calculation is intractable. So, instead of marginalizing over all of the information in $\vec{t}$, we marginalize over only some of it, and sample the remaining partial analyses. We call these partial analyses "coarse structures" ($\vec{s}$), described below. We begin by randomly initializing a set of possible coarse structures for the observed strings. Then, we use Gibbs sampling to jointly infer the posterior $P(G, \vec{s} \mid \vec{w})$, alternating between sampling a new grammar according to $P(G \mid \vec{s}, \vec{w})$, and sampling new coarse structures according to $P(\vec{s} \mid G, \vec{w})$. This process will converge to the joint posterior distribution over G and $\vec{s}$.

The coarse structures $\vec{s}$ take the same shape as the trees generated by the learner's grammars, but abstract away from the distinction between core and noise rewrites in those trees. This corresponds to abstracting away from the distinction between solid and dashed lines in Fig. 2. Unlike a full tree, which commits to particular core vs. noise distinctions and therefore is compatible with only some grammars, any coarse structure is consistent with all of the grammars in the learner's hypothesis space: it might be generated by core rules in certain grammars, or by some combination of noise and core rules, or by *only* noise rules, which are shared across all grammars. Therefore, for every grammar G, $P(G \mid \vec{s}, \vec{w})$ is always non-zero, allowing us to draw samples from this posterior in a feasible way. We sample $\vec{s}$ from the posterior $P(\vec{s} \mid G, \vec{w})$ with a Hastings proposal, using a variant of an algorithm introduced by Johnson et al. (2007) and marginalizing over $\vec{\theta^G}$. See Appendix B for details.

5 Simulations

We tested our model on English and French. These languages are both canonically SVO, but differ in how strictly they adhere to this canonical pattern: English has rigid word order, whereas French allows a greater degree of argument dislocation. We show that our model sucessfully identifies SVO as the target grammar for its noisy data, and does so even in an expanded hypothesis space that allows a choice among more flexible discrete hypotheses. Moreover, our model out-performs a learner whose grammar allows all word-order rules with some probability (Fig. 3B), with a numerical bias to prefer rule weights that are close to 0 or 1. This shows that for this case study, our model fares better than the more common type of explicit regularization bias in prior literature.

5.1 Data

We used the Eve and Lyon CHILDES corpora (Brown, 1973; Demuth and Tremblay, 2008), which contain speech directed to English- and French-learning 1- and 2-year-olds (see Table 1). We searched these corpora for strings of one v and any number of satellite np's. We used a noisy heuristic to approximate the knowledge of infants at 15 months and younger, who can use functional cues— determiners, pronouns, and auxiliaries— to differentiate nouns and verbs (Babineau et al., 2020; Shi and Melançon, 2010; Hicks et al., 2007). We categorized any full pronoun as an np; any word following a determiner as the head of an np; and any word following an auxiliary as a v. *Wh*-words and object clitics were not categorized as np's, because they may not be recognized as such by infants learning basic word order (Perkins and Lidz, 2021; Brusini et al., 2017). Object clitics that are homophonous with determiners were treated erroneously as determiners, to simulate the uncertainty that infants might have about their category.

To create the datasets for our learner, we sampled 50 strings in their relevant proportions in each language (see Table 2). Over 30% of the strings in each language are incompatible with the core rules of the target SVO grammar. As a whole, these data cannot be generated by the core rules of any single grammar in the learner's hypothesis space, without considering the option of noise.

5.2 Results: Our model

Fig. 4 displays our model's inferred posterior probability distribution over the four Mixture PCFGs in its hypothesis space, averaged over 10 runs of the model in each language. In both English and French, the SVO grammar was assigned a higher posterior probability than any other grammar in the learner's hypothesis space (all $ps < 0.001$, Binomial tests). This shows that the learner's filtering mechanism allowed it to overcome the large amount of noise in its data. The learner successfully discovered that the best explanation for its data involved identifying some portions that were signal for core SVO word order, and some portions that came from noise processes.

5.3 Comparison: Fully-flexible model

In order to assess how much our model's success depended on a choice of discrete canonical word-order grammars, we constructed a comparison learner whose hypothesis space collapses the distinction between canonical and non-canonical structures. This "fully-flexible" hypothesis space consists of a single standard PCFG comprising all of the word-order rules across our learner's four grammars (Fig. 3B). For this model, learning canonical word order would mean identifying that some of its rules have probabilities near zero.

We tested two variants of this model. The first assumes that all rules in its hypothesis space are equally probable *a priori*, as in our original model. The second is numerically biased to regularize its rule weights, following the regularization approach in prior literature (Reali and Griffiths, 2009; Culbertson et al., 2013; Ferdinand et al., 2019). This regularization bias takes the form of a skewed prior over the rule weights $\vec{\theta}$ in the learner's grammar. For each nonterminal A, we set all component parameters α_i^A of the model's Dirichlet prior to a small value, 0.001. This biases the learner to put probability mass on only one expansion of a given nonterminal, and push the probabilities of other expansions towards zero.

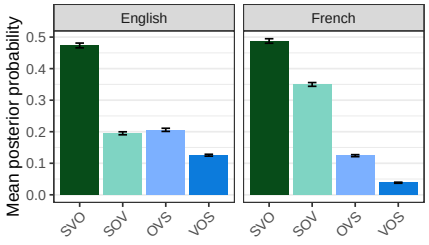

Figure 4: Posterior distribution over grammars

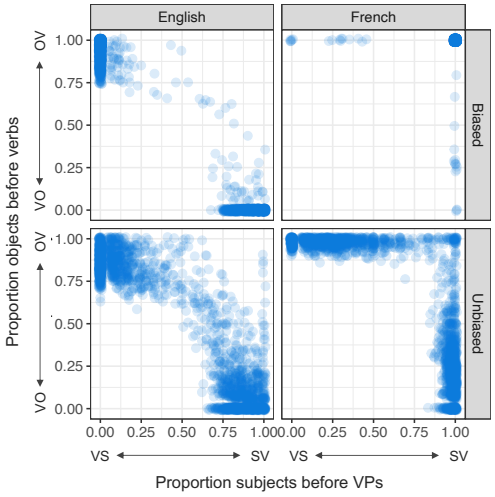

Figure 5: Posterior distribution over subject and object position in sampled treesets ($\vec{t}$), fully-flexible learner

The learner's inference process consists of one of the steps in our original Gibbs sampler. We sample trees for the learner's data from the posterior given its sole grammar, $P(\vec{t} \mid G, \vec{w})$, just as we sampled $P(\vec{s} \mid G, \vec{w})$ in our original model.

We assessed whether the fully-flexible learner had identified a canonical word order by calculating the proportion of the learner's sampled trees that contained subject NPs before verb phrases and object NPs before verbs. These proportions are plotted in Fig. 5, where each point corresponds to a sampled set of trees, aggregated across ten runs of the model in each language. These plotted distributions provide an estimate of the learner's inferred posterior probabilities of subject-initial and object-initial structures. The four possibilities for canonical word order correspond approximately to the four corners in each panel: clockwise from top left, these are OVS, SOV, SVO, and VOS.

If the learner had successfully identified that English and French are canonically SVO, the majority of tree samples would lie close to the lower right corners of these graphs. Instead, the unbi-

ased learner (bottom) inferred a distribution over tree structures that mirrored its noisy data. These ranged from the OVS to the SVO regions in English, and across the OVS, SOV, and SVO regions in French. The biased learner (top) inferred distributions closer to the corners corresponding to canonical word orders. However, the English learner gave equal posterior probability to both OVS and SVO structures; its mean proportions of subject-initial and object-final trees were not significantly different from 0.5 (mean subject-initial: 0.51, mean object-final: 0.54, $ps > 0.67$). The French learner converged to SOV structures instead of SVO (mean subject-initial: 0.99, mean subject-final: 0.01, $ps < 0.001$). The learner's regularization bias helped it identify one or two canonical word orders for its noisy data. But unlike our model, it did not correctly converge on SVO as the most probable word order in either language.

Why would our approach fare better than the more common approach to regularization in past work? Our model's success comes in large part from its expectation that canonical clauses require subjects; subject-drop can occur only in non-canonical clause types. This allows our learner to use the large number of np v strings as evidence for a subject-initial grammar. Given the choice between using its restricted core rules to analyze the sole np as a canonical subject, versus using its noise rules to analyze the np in a different position, a preference emerges for the canonical-subject analysis— just as we prefer to analyze a sequence of heads as coming from a two-headed rather than a head-tail coin. The fully-flexible learner does not distinguish between canonical structures in which subjects are required, and non-canonical structures in which they are not, so no preference emerges to to analyze a sole np in a specific clausal position.

For this learning problem, it appears helpful to have a hypothesis space with a distinction between core rules that provide deterministic options for canonical word order, and noise rules that produce non-canonical structures. This mixture of deterministic and non-deterministic options is what allows the target basic clause structure to emerge as the best explanation of the learner's noisy data.

5.4 Comparison: A data-coverage heuristic

Our learner successfully shows a preference for some hypotheses over others in a scenario where none are compatible with all of the data. But

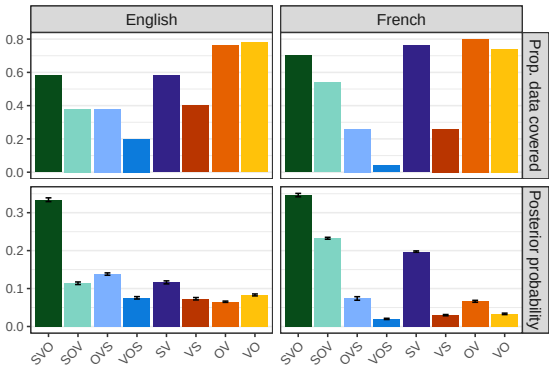

Figure 6: Eight-way hypothesis space: proportion data coverage vs. model's posterior distribution

one might ask whether the same result could be achieved via a much simpler approach: the core rules of the SVO grammar can generate 56% of the English data, which is a greater proportion than can be generated by the core rules of any of the alternatives (each less than 40%), and so this alone might lead a learner to identify SVO as the preferred option. Our model's inference mechanism does more than simply recapitulate this "data coverage" heuristic. To see this, it is useful to consider a scenario where the learner has a wider range of discrete hypotheses to choose among, including some that are more restrictive than others.

We constructed a comparison learner that considers an eight-way choice among Mixture PCFGs. These include all four deterministic options from our original model: grammars whose core rules fix subject and object position. In addition, the hypothesis space includes four more flexible grammars whose core rules fix only one of those argument positions, and allow the other to vary. For instance, the grammar we call "SV" fixes the subject preverbally, but allows the object to appear either before or after the verb: its rules are the union of the rules in the SVO and SOV grammars. Similarly, the "VS" grammar fixes the subject post-verbally but allows object position to vary, and the "OV" and "VO" grammars fix object position, but allow subject position to vary. All eight grammars share the same set of noise rules as in our original learner.

Given a choice among these eight grammars, the data-coverage heuristic will always favor one of the four more flexible ones, since they generate unions of the stringsets generated by the original four. In each of the top panels in Fig. 6, where a comparison only among the leftmost four grammars would have SVO as the winner (roughly mirroring Fig. 4),

we see that the more flexible grammars in general fare better by the data-coverage metric. But our learner, on both languages, assigned SVO higher posterior probability than any other grammar in the hypothesis space (Fig. 6, bottom; all $ps < 0.001$, averaged across 10 runs of the learner).

Why does our learner still succeed at identifying that English and French are SVO, even when there are other hypotheses that cover more of the data? Intuitively, our learner considers a tradeoff between fit to the data and restrictiveness of its hypotheses. Given the choice between the restrictive SVO hypothesis that provides a decent fit to the data, and the more flexible hypotheses that provide slightly better fits, a preference emerges for the more restrictive option— again paralleling our intuitive preference to attribute as many coin flips as possible to a two-headed rather than a head-tail coin. In our original model, this preference for restrictive hypotheses applied *within* each grammar, governing the learner's choice of attributing data to the restrictive core rules vs. the flexible noise rules. Here, we show that this same mechanism informs the learner's choice *across* grammars.

These findings demonstrate the flexibility and robustness of this learning mechanism. Our learner identifies strict SVO word order as its preferred hypothesis not only in comparison with other equally-strict alternatives, but also when other less restrictive options are available; the fact that it settled on deterministic SVO order in Fig. 4 was not simply a by-product of the fact that we provided only deterministic options. An implicit tradeoff between a grammar's restrictiveness and its fit to the data, and the expectation that this fit will be noisy, together enable the learner to identify the target deterministic word order among more flexible hypotheses.

6 Discussion

We introduce a general mechanism for noise-tolerant learning of deterministic grammars. Our learner assumes that its data are generated by a complex system: the particular grammatical processes that the learner is currently trying to acquire, and other unknown processes that conspire to introduce variability into the data. We model the inference process as a special case of probabilistic grammar learning, in which the learner evaluates a choice among different *Mixture PCFGs*: composite grammars in which each node might be introduced either by a restricted set of "core" rules, or by a less restricted set of "noise" rules.

We apply this approach to the problem of acquiring basic word order from immature sentence representations. Using distributions of imperfectly-identified noun phrases and verbs, our model successfully infers that English and French are SVO, without further cues to underlying sentence structure. It does so by separating signal for canonical word order from noise due to non-canonical structures, thereby implementing a proposal that young learners "filter" non-canonical clauses from their data (Pinker, 1984; Perkins et al., 2022). Because the learner's grammatical hypotheses allow only certain restricted core rules, a preference emerges to use these core rules to explain the skews in its data when possible, rather than analyzing most of the data as noise. This provides the impetus for successful filtering, even though our learner does not know ahead of time the rate or properties of non-canonical clauses in the language.

While we focus here on Mixture PCFGs, this same approach can be applied to "mixture" versions of other sorts of grammars that generate complex structures as a function of local choices about smaller subparts. This approach may therefore generalize to many other problems in grammar learning: e.g., learning phonological constraints that can be expressed in mixture finite-state systems, or learning syntactic dependencies that can be expressed in mixture multiple context-free grammars.

More broadly, this approach provides a novel mechanism for regularization in grammar learning. Here, a learner's tendency to regularize variable data is not driven by an explicit bias to prefer extreme points in a fully-gradient space, but instead emerges from the learner's expectation that its data are a noisy realization of a restrictive underlying system. This invites the possibility that other observed cases of regularization may be accounted for without adopting a fully-flexible hypothesis space. Instead, successful learning in certain domains may be underwritten by deterministic options in the learner's hypothesis space, combined with a general mechanism for filtering signal from noise.

Acknowledgments

We thank Xinyue Cui, Naomi Feldman, Jeff Lidz, Shalinee Maitra, the audiences at BUCLD 2022 and the UCLA Psycholinguistics/Computational Linguistics Seminar, and two anonymous reviewers for helpful feedback and assistance.

References

Mireille Babineau, Rushen Shi, and Anne Christophe. 2020. 14-month-olds exploit verbs' syntactic contexts to build expectations about novel words. *Infancy*, 25(5):719–733. Publisher: Wiley Online Library.

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

Perrine Brusini, Ghislaine Dehaene-Lambertz, Marieke Van Heugten, Alex De Carvalho, François Goffinet, Anne-Caroline Fiévet, and Anne Christophe. 2017. Ambiguous function words do not prevent 18-month-olds from building accurate syntactic category expectations: An ERP study. *Neuropsychologia*, 98:4–12. Publisher: Elsevier.

Anne Christophe, Séverine Millotte, Savita Bernal, and Jeffrey Lidz. 2008. Bootstrapping Lexical and Syntactic Acquisition. *Language and Speech*, 51(1-2):61–75.

Jennifer Culbertson, Paul Smolensky, and Colin Wilson. 2013. Cognitive biases, linguistic universals, and constraint-based grammar learning. *Topics in Cognitive Science*, 5(3):392–424.

Katherine Demuth and Annie Tremblay. 2008. Prosodically-conditioned variability in children's production of French determiners. *Journal of child language*, 35(1):99–127. Publisher: Cambridge University Press.

Vanessa Ferdinand, Simon Kirby, and Kenny Smith. 2019. The cognitive roots of regularization in language. *Cognition*, 184:53–68.

Lila Gleitman. 1990. The structural sources of verb meanings. *Language Acquisition*, 1:3–55.

Jessica Hicks, Jessica Maye, and Jeffrey Lidz. 2007. The role of function words in infants' syntactic categorization of novel words. In *Proceedings of the Linguistic Society of America Annual Meeting*, Anaheim, CA.

Kathy Hirsh-Pasek and Roberta Michnick Golinkoff. 1996. The intermodal preferential looking paradigm: A window onto emerging language comprehension. In Dana McDaniel, Cecile McKee, and Helen S. Cairns, editors, *Methods for assessing children's syntax*, pages 105–124. The MIT Press, Cambridge, MA.

Carla L. Hudson Kam and Elissa L. Newport. 2005. Regularizing Unpredictable Variation: The Roles of Adult and Child Learners in Language Formation and Change. *Language Learning and Development*, 1(2):151–195.

Carla L. Hudson Kam and Elissa L. Newport. 2009. Getting it right by getting it wrong: When learners change languages. *Cognitive Psychology*, 59(1):30–66.

Mark Johnson, 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, New York. Association for Computational Linguistics.

Jeffrey Lidz, Aaron Steven White, and Rebecca Baier. 2017. The role of incremental parsing in syntactically conditioned word learning. *Cognitive Psychology*, 97:62–78.

Laurel Perkins, Naomi H. Feldman, and Jeffrey Lidz. 2022. The Power of Ignoring: Filtering Input for Argument Structure Acquisition. *Cognitive Science*, 46:e13080.

Laurel Perkins and Jeffrey Lidz. 2020. Filler-gap dependency comprehension at 15 months: The role of vocabulary. *Language Acquisition*, 27(1):98–115.

Laurel Perkins and Jeffrey Lidz. 2021. 18-month-old infants represent non-local syntactic dependencies. *Proceedings of the National Academy of Sciences*, 118(41):e2026469118.

Steven Pinker. 1984. *Language learnability and language development*. Harvard University Press, Cambridge, MA.

Florencia Reali and Thomas L. Griffiths. 2009. The evolution of frequency distributions: Relating regularization to inductive biases through iterated learning. *Cognition*, 111:317–328.

Jordan Schneider, Laurel Perkins, and Naomi H. Feldman. 2020. A noisy channel model for systematizing unpredictable input variation. In *Proceedings of the 44th Annual Boston University Conference on Language Development*, pages 533–547.

Rushen Shi and Andréane Melançon. 2010. Syntactic Categorization in French-Learning Infants. *Infancy*, 15(5):517–533.

A Details of the coins example from Section 2

Recall the scenario with just Bag H: this bag contains an unknown number of Type A coins, which always come up heads, and an unknown number of Type B coins, which all have some single unknown probability θ of coming up heads. Ten times, a coin is chosen from the bag and flipped; this produces eight heads and two tails. How many of these ten flips might we guess came from Type A coins, and how many from Type B coins?

We consider three hypotheses:

- H1: 0 Type A flips, 10 Type B flips

- H2: 6 Type A flips, 4 Type B flips
- H3: 8 Type A flips, 2 Type B flips

The three hypotheses' likelihoods, conditioned upon the unknown probability θ, are as follows.

$$(4)\quad \Pr(\text{data} \mid \text{H1}, \theta) = \binom{10}{8}\, \theta^8\, (1-\theta)^2$$

$$(5)\quad \begin{aligned}\Pr(\text{data} \mid \text{H2}, \theta) &= \binom{6}{6}\, 1^6 \cdot \binom{4}{2}\, \theta^2\, (1-\theta)^2 \\ &= \binom{4}{2}\, \theta^2\, (1-\theta)^2\end{aligned}$$

$$(6)\quad \begin{aligned}\Pr(\text{data} \mid \text{H3}, \theta) &= \binom{8}{8}\, 1^8 \cdot \binom{2}{0}\, \theta^0\, (1-\theta)^2 \\ &= (1-\theta)^2\end{aligned}$$

As noted in the main text, these expressions highlight the fact that H1 is the most costly hypothesis, since it relies most heavily on the contingent outcomes from Type B coins, and H3 is the least costly.

We can make this more precise by marginalizing over the unknown value of θ in (4) and (5). The useful general result here is that

$$(7)\quad \int_0^1 \binom{n}{k} \theta^k (1-\theta)^{n-k} d\theta = \frac{1}{n+1}$$

for any n and k; notice that the right-hand side only depends on n. Marginalizing over θ in (4), (5) and (6), under the assumption of a uniform prior on θ, yields integrals of this form.[5] For H1, $n = 10$ so $\Pr(\text{data} \mid \text{H1}) = \frac{1}{11}$. What this highlights is that the likelihood under such a hypothesis depends *only* on the number of times that hypothesis needs to invoke the uncertain Type B coin flip: *any* outcome of the ten-flip experiment invoked by H1 has probability $\frac{1}{11}$, and *any* outcome of the two-flip experiment invoked by H3 has probability $\frac{1}{3}$.

$$\begin{aligned}\Pr(\text{data} \mid \text{H1}) &= \tfrac{1}{11} \\ \Pr(\text{data} \mid \text{H2}) &= \tfrac{1}{5} \\ \Pr(\text{data} \mid \text{H3}) &= \tfrac{1}{3}\end{aligned}$$

Now consider the choice between Bag H and Bag T, as candidate explanations for a sequence of ten flips that yielded eight heads and two tails. We have seen that, using Bag H, the possible hypotheses range from those that provide "good" explanations of the data (such as H3 at one extreme) by exploiting the presence of the two-headed coins, to those that constitute "costly" explanations (such as H1 at the other extreme) because they rely heavily on flips of the head-tail coins; see Fig. 7. With Bag T, the explanations at the costly extreme are still available (e.g. the hypothesis that all ten flips came from head-tail coins; $n = 10$), but there is no way for the two-tailed coins to contribute to particularly good explanations of the observed high proportion of heads. To minimize the reliance on the contingent outcomes of head-tail coins, the best one can do is to suppose (call this H4) that the two observed tails both came from two-tailed coins, which still leaves eight uncertain flips. The likelihood under this hypothesis (compare with (5) for H2) is

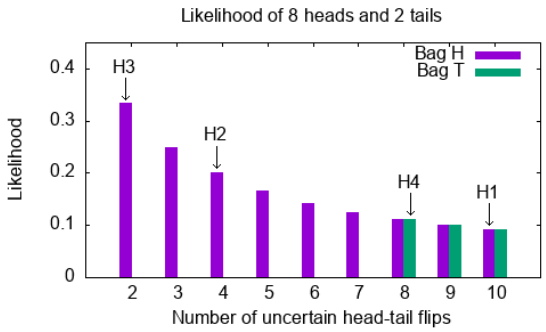

Figure 7

$$(8)\quad \begin{aligned}\Pr(\text{data} \mid \text{H4}, \theta) &= \binom{2}{2}\, 1^2 \cdot \binom{8}{0}\, \theta^8\, (1-\theta)^0 \\ &= \theta^8\end{aligned}$$

and $\Pr(\text{data} \mid \text{H4}) = \frac{1}{9}$.

Returning now to the overarching choice between the two bags: the likelihood assigned to the data by a particular bag is the sum of the heights of the associated bars in Fig. 7. This is clearly larger for Bag H, and so assuming a flat prior over the two bags, the posterior probability of Bag H will be higher than that of Bag T.

B Details of Gibbs sampling

In the first step of sampling, we use Bayes' Rule to calculate the posterior probability of each grammar given the observed strings $\vec{w}$ and a collection of hypothesized coarse structures $\vec{s}$ for those strings:

$$(9)\quad P(G|\vec{s},\vec{w}) = \frac{P(\vec{s},\vec{w}|G)P(G)}{\sum_{G'} P(\vec{s},\vec{w}|G')P(G')}$$

Bayes' Rule tells us that the posterior probability of any grammar is proportional to the product of

[5] Specifically, the uniform prior can be represented as a Beta(1,1) distribution over θ, so $\Pr(\text{data} \mid \text{H1}) = \int_0^1 \Pr(\text{data} \mid \theta, \text{H1})\text{Beta}(\theta \mid 1,1)d\theta = \int_0^1 \Pr(\text{data} \mid \theta, \text{H1})d\theta$, since $\text{Beta}(\theta \mid 1,1) = 1$ for all θ.

the likelihood (the probability of $\vec{s}$ and $\vec{w}$ under that grammar) and the prior probability of that grammar. We assume that all four grammars have equal prior probability.

Because we are only considering coarse structures that could have yielded the strings in the data, the joint likelihood of the coarse structures and strings, $P(\vec{s}, \vec{w}|G)$, is equivalent to the likelihood of the coarse structures alone, $P(\vec{s}|G)$. Calculating this likelihood requires summing over the unknown ways that each portion of these coarse structures might be analyzed as either a core (ϕ, solid line) or noise (ψ, dashed line) rewrite. The specific core vs. noise choices are interchangeable for each particular nonterminal given a grammar, so we make this calculation tractable by considering how *many* core vs. noise rewrites might have occurred for each nonterminal.

We divide the n^A total observations of a particular nonterminal A into $n_1^A \ldots n_m^A$ observations of the 1^{st} through the m^{th} possible rewrites (collapsing across ϕ-rewrites and ψ-rewrites of A). The full likelihood of the set of coarse structures, $P(\vec{s}|G)$, is the product over all nonterminals A of $P(n_1^A \ldots n_m^A \mid G)$. We divide each of the observed rewrites of a nonterminal into some number of core (solid line) rewrites (ϕ) and some number of noisy (dashed line) rewrites (ψ). The n_1^A occurrences of the first type of rewrite for A are divided into $n_1^{A^\phi}$ core occurrences and $n_1^{A^\psi}$ noisy occurrences. More generally, the n_m^A occurrences of the m^{th} rewrite type are divided into $n_m^{A^\phi}$ core occurrences and $n_m^{A^\psi}$ noisy occurrences. We can calculate the likelihood by marginalizing over $n_1^{A^\phi} \ldots n_m^{A^\psi}$:

$$P(\vec{s}|G) = \prod_A P(n_1^A \ldots n_m^A \mid G) = \tag{10}$$

$$\prod_A \Bigg[\sum_{n_1^{A^\phi}=0}^{n_1^A} \cdots \sum_{n_m^{A^\phi}=0}^{n_m^A} \Big[P(n_1^{A^\phi} \ldots n_m^{A^\phi} | n^{A^\phi}, G) \times P(n_1^{A^\psi} \ldots n_m^{A^\psi} | n^{A^\psi}, G) \times P(n^{A^\phi} | n^A, G) \Big] \Bigg]$$

The first term in the sum is the probability of observing $n_1^{A^\phi} \ldots n_m^{A^\phi}$ core occurrences of each rewrite type, out of n^{A^ϕ} total core occurrences of A. This follows a multinomial distribution with parameter $\vec{\phi}^{A_G}$. Because $\vec{\phi}^{A_G}$ is unknown, we integrate over all possible values of $\vec{\phi}^{A_G}$ to obtain

$$\frac{B(\vec{\alpha_\phi}^{A_G} + (n_1^{A^\phi} \ldots n_m^{A^\phi}))}{B(\vec{\alpha_\phi}^{A_G})} \tag{11}$$

for this first term, where $\vec{\alpha_\phi}^{A_G}$ represents the parameters of the Dirichlet prior over $\vec{\phi}^{A_G}$, and $B(\cdot)$ is the multivariate Beta function.

The second term in the sum in (10) is analogous: this is the probability, given n^{A^ψ} total noisy occurrences of A, of observing $n_1^{A^\psi} \ldots n_m^{A^\psi}$ noisy occurrences of each rewrite type, which follows a multinomial distribution with parameter $\vec{\psi}^{A_G}$. The third term is the probability of observing n^{A^ϕ} total core occurrences out of n^A overall occurrences of A. This follows a binomial distribution with parameter $(1 - \epsilon^{A_G})$. We again integrate over all possible values of $\vec{\psi}^{A_G}$ and ϵ^{A_G}, obtaining results analogous to Eq. (11).

This allows us to calculate the likelihood $P(\vec{s} \mid G)$ for each G in our hypothesis space, and (since we assume a flat prior of grammars) sample a new G with probability proportional to this likelihood.

After re-sampling a new grammar G, we then use a component-wise Hastings proposal to sample a new set of coarse structures $\vec{s}$ for the observed strings, given G. Following Johnson et al. (2007), we consider the probability of a particular coarse structure s_i for corresponding string w_i, given G and the current hypotheses about coarse structures $\vec{s}_{-i}$ for all the other strings. We can define a function f that is proportional to the posterior distribution over s_i, $f(s_i) \propto P(s_i|w_i, \vec{s}_{-i}, G)$, as

$$f(s_i) = P(w_i|s_i)P(s_i|\vec{s}_{-i}, G) \tag{12}$$

The probability of a string being the yield of a given coarse structure, $P(w_i|s_i)$, is always 1 or 0. The probability of a coarse structure given all other coarse structures and G, $P(s_i|\vec{s}_{-i}, G)$, is

$$P(s_i|\vec{s}_{-i}, G) = \frac{P(\vec{s}|G)}{P(\vec{s}_{-i}|G)} \tag{13}$$

Both $P(\vec{s}|G)$ and $P(\vec{s}_{-i}|G)$ are calculated according to to Eq. (10).

We can use this function f to sample $\vec{s}$ given G and $\vec{w}$ as follows. Within each iteration of the Gibbs sampler, we re-sample $\vec{s}$ using a procedure modified from Johnson et al. (2007). First, we choose a string w_i and its current corresponding s_i at random. Second, we take the other coarse structures $\vec{s}_{-i}$, to be the output of a simple PCFG

which generates coarse structures directly, rather than the full trees generated by a Mixture PCFG. We estimate of the weights of this PCFG, $\vec{\theta^s}$, from the relative frequencies of each observed rewrite, using add-one smoothing to account for accidental gaps. Third, we generate a new proposed coarse structure $s_i{}'$ for w_i by sampling from this grammar's distribution using $\vec{\theta^s}$. Finally, we decide to accept this proposal with probability

$$A(s_i') = min\left(1, \frac{f(s_i{}')P(s_i|w_i, \vec{\theta^s})}{f(s_i)P(s_i{}'|w_i, \vec{\theta^s})}\right) \quad (14)$$

We ran multiple chains from different starting places to test convergence. For the simulations reported in Sec. 5.2, we ran chains of 20,000 iterations of Gibbs sampling each, and analyzed every 10th iteration from the last half of each chain. We report averages across 10 chains as estimates of the posterior over G and $\vec{s}$. To simulate the "fully-flexible" learner described in Sec. 5.3, we estimate the posterior distribution over $\vec{t}$ by using a component-wise Hastings sampler analogous to that for estimating $P(\vec{s}|G, \vec{w})$ in our original model. We ran 10 chains of 20,000 Hastings iterations each, and analyzed every 10th iteration from the last half of each chain.