Modeling Regularization in Language Acquisition as Noise-Tolerant Grammar Selection

Laurel Perkins, Tim Hunter

Department of Linguistics, University of California Los Angeles

Author Note

Address for correspondence:

Laurel Perkins, Tim Hunter

3125 Campbell Hall

Los Angeles, CA 90025

perkinsl@ucla.edu, timhunter@ucla.edu

Abstract

Language acquisition involves drawing systematic generalizations from messy data. On one hypothesis, this is facilitated by a domain-general bias for children to "regularize" inconsistent variability, sharpening the statistical distributions in their input towards more systematic extremes. We introduce a general computational framework for modeling a different explanation: on this view, children expect that their data are a noisy realization of a restrictive underlying grammatical system. We implement a learner that evaluates a choice among composite context-free grammars, in which a restricted set of "core" rules, comprising the particular grammatical processes that the learner is currently trying to acquire, operate alongside a less restricted set of "noise" rules, representing other independent processes that have yet to be learned, and conspire to introduce variability into the data. Our *Noisy CFG Learner* partitions its data into portions that serve as evidence for one of the possible core grammars in its hypothesis space, and portions generated by these noise processes. It does so without knowing in advance how much noise occurs or what its properties are. We compare our learner to a common implementation of the general regularization bias approach, and show that both can account for children's behavior in a representative artificial language learning experiment. However, we find that our approach performs better on two naturalistic case studies in early syntax acquisition: learning the rules governing canonical word-order and case-marking, given natural language data with "noise" from non-canonical sentence types. We show that our learner succeeds because its architecture allows a natural way to express linguistically-motivated expectations about the character of those rules. This suggests that, in certain domains, successful learning from messy data may be enabled by a hypothesis space comprising restrictive grammatical options.

Keywords: language acquisition, syntax, grammar, regularization, computational modelling, Bayesian reasoning

1 Introduction

Language acquisition involves drawing systematic generalizations from data that appear on the surface to be messy and variable. For instance, infants acquire the system of sound categories in their language from noisy, overlapping distributions of sound tokens in the speech that they hear (Bion, Miyazawa, Kikuchi, & Mazuka, 2013; Cristia, 2018; Hitczenko & Feldman, 2022; Maye, Werker, & Gerken, 2002; Swingley, 2019). They learn the phonetic and phonological regularities that mark word boundaries in their language, despite a high degree of variability in how words are pronounced (Beech & Swingley, 2023; Cristia, Dupoux, Ratner, & Soderstrom, 2019; Jusczyk & Aslin, 1995; Mattys, Jusczyk, Luce, & Morgan, 1999; Mattys & Jusczyk, 2001). In the domain of syntax, infants learn their language's basic word order despite immature abilities to identify subjects and objects in the sentences that they hear, and despite variability from constructions such as *wh*-questions and passives, where these clause arguments occur in non-canonical orders (Dautriche et al., 2014; Hirsh-Pasek & Golinkoff, 1996; Gertner, Fisher, & Eisengart, 2006; Lidz, White, & Baier, 2017; Perkins & Lidz, 2020, 2021; Pinker, 1984). What kind of mechanisms allow for learning to abstract away from such messiness in the learner's representation of the data?

On one view, this behavior arises at least in part from a general bias, observed not only in language but also in other cognitive domains, for young learners to "regularize" unpredictably variable input. 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. An influential literature has proposed that this tendency underlies children's behavior both in learning probabilistic regularities in non-linguistic domains, and in the context of acquiring language from non-native speakers, where it could be a mechanism responsible for language change (Austin, Schuler, Furlong, & Newport, 2022; Hudson Kam & Newport, 2005, 2009; Newport, 1999; Reali & Griffiths, 2009; Singleton & Newport, 2004; Smith & Wonnacott, 2010).

There are two important assumptions implicit in this approach, which we will call the

general regularization bias account. The first is that learners' regularization behavior across various linguistic and non-linguistic domains can be attributed at their core to the same set of non-linguistic cognitive factors operative in early development (Austin et al., 2022; Culbertson & Kirby, 2016; Ferdinand, Kirby, & Smith, 2019; Keogh, Kirby, & Culbertson, 2024; Hudson Kam & Newport, 2005, 2009; Newport, 1990). The second is that across all of these domains, learners have 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 (Austin et al., 2022; Hudson Kam & Newport, 2005, 2009). This can be seen as the result of a regularization bias operating within a learner's fully-flexible hypothesis space, pushing learners to prefer probabilities closer to zero or one and producing near-categorical learning outcomes. This idea could be applied to the learning of basic word order in infancy, such as learning that English is canonically subject-verb-object (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 towards one consistent order. Importantly, because this behavior is taken to be the product of a shared set of cognitive factors operating across different learning domains, no specific extreme within the learner's hypothesis space is preferred *a priori* for domain-specific reasons. What might differ by domain or learning context is simply the degree to which this overall skewing or numerical sharpening occurs (Ferdinand et al., 2019; Culbertson & Kirby, 2016; Reali & Griffiths, 2009).

Here we explore an alternative account, on which learners' regularization behavior primarily reflects an expectation about the specific character of the system that generated their data in a particular domain. We propose that in certain circumstances, learners face a choice among discrete hypotheses, each of which is restrictive or deterministic in a way that is incompatible with the full variability of the observed data. Learners assume that their

data result from an opaque interaction between (i) one of the restrictive hypotheses that they are currently considering, and (ii) various other processes that might introduce "noise" into the data. This proposal is consistent with certain accounts of language acquisition, which posit that learners bring restrictive domain-specific expectations about the types of grammatical mechanisms that may have generated their linguistic input. Learning is not an attempt to encode the full distributions in the data veridically, but rather to use those distributions as evidence for the parameters of this generative system, under the assumption that these parameters will interact in opaque ways (Chomsky, 1965, 1975; Fodor, 1998; Lidz & Gagliardi, 2015; Lightfoot, 1991; Valian, 1990; Yang, 2002).

This *restrictive hypotheses* proposal can explain children's linguistic regularization behavior in the following ways. For a child learning an artificial determiner system, the data might reflect a combination of signal for restrictive rules governing determiner distributions, and noise coming from unknown grammatical or extra-grammatical processes. For a child learning the syntax of basic clauses, the data reflect a combination of signal for the restrictive rules governing 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 restrictive hypothesis within their noisy data, and adopt this hypothesis as the best explanation despite its surface mismatch with the data.

While the compatibility of this proposal with previous experimental findings has been acknowledged in the prior literature (Austin et al., 2022; Ferdinand et al., 2019; Hudson Kam & Newport, 2009), it has not yet been explicitly investigated computationally. A general computational approach exists for modeling learning under the first, general regularization bias account (e.g., Culbertson, Smolensky, & Wilson, 2013; Perfors, 2012; Reali & Griffiths, 2009; Smith et al., 2017). But no such general approach exists for the second, which has been studied only narrowly through a handful of specific case studies (Perkins, Feldman, & Lidz, 2022; Schneider, Perkins, & Feldman, 2020). Here, we provide this approach.

We introduce a general computational framework for choosing among discrete,

restrictive hypotheses in the face of noisy data. A learner of the sort we describe below expects that its data are generated by a complex system: a restrictive core component that the learner is attempting to acquire, operating alongside a "noise" component whose properties are currently unknown. The learner aims to separate evidence for a restrictive core grammar from the distorting effects of non-canonical noise processes, without knowing ahead of time how much noise is present in its data, or what the properties of that noise are. We compare this learning architecture to a common implementation of the general regularization bias approach and show that (i) both can indeed account for children's regularization behavior in a representative artificial language experiment (Austin et al., 2022), but (ii) our approach fares better in modeling two case studies of naturalistic phenomena in early syntax acquisition: learning word order and case-marking from messy and variable representations of data. We argue that a key to our model's success in these learning problems, and the reason that it performs better than the previous regularization bias approach, is that it can naturally encode substantive, domain-specific expectations about the nature of grammatical rules— for instance, the expectation that canonical clauses require subjects. This provides support for views in which successful learning from noisy data depends on a hypothesis space comprising restrictive grammatical options.

2 Modeling language learning from noisy data

2.1 The phenomenon of regularization

At early stages of grammatical development, children's representations of their linguistic input are immature, incomplete, and sometimes inaccurate. Acquiring any piece of grammatical knowledge therefore requires mechanisms for abstracting away from messiness in the data, as a child perceives it (Fodor, 1998; Lidz & Gagliardi, 2015; Perkins et al., 2022; Valian, 1990).

A potential clue to the nature of these learning mechanisms emerges from studies of learning from non-native language speakers. In such cases, a child's data may contain noise

introduced not only by their own incomplete grammatical knowledge, but also by the incomplete grammatical knowledge of their parents (J. S. Johnson, Shenkman, Newport, & Medin, 1996; Singleton & Newport, 2004; Newport, 1999; Wolfram, 1985). An example is reported in Singleton and Newport (2004), who studied a Deaf child acquiring American Sign Language solely from parents who were late learners of the language. His parents' morphological productions were inconsistently accurate, and yet his own were strikingly consistent, approaching the accuracy of children learning the language from native signers. Singleton and Newport posit that this illustrates a general trend for young learners to regularize their input, drawing generalizations that are more systematic or categorical than their variable data would seem to support. In the contexts of learning from non-native speakers, this phenomenon could be responsible for creolization and language change (Hudson Kam & Newport, 2005, 2009; Newport, 1999; see also Bickerton, 1981, 1984; Senghas & Coppola, 2001). And it may broadly apply to many other cases of learning language from data distorted by noise, whether real or perceived.

This phenomenon has been extensively investigated in experimental settings, in both child and adult learners (e.g., Austin et al., 2022; Culbertson et al., 2013; Ferdinand et al., 2019; Hudson Kam & Newport, 2005, 2009; Perfors, 2012; Reali & Griffiths, 2009; Smith & Wonnacott, 2010). For instance, a series of artificial-language learning experiments exposed children and adults to input in which novel determiners occurred with novel nouns inconsistently (Austin et al., 2022; Hudson Kam & Newport, 2005, 2009). When asked to produce sentences in this language, adults tended to maintain this variability, producing determiners at roughly the rates at which they had heard them in the training data. By contrast, young children tended to behave as if they had learned more systematic or skewed rules, producing one determiner at a much higher rate.

The contrast between adult and child behavior in these experiments resembles a developmental trend in an older literature on non-linguistic "probability learning." In a representative task, Gardner (1957) exposed adults to two flashing lights: in one condition,

Light A flashed on a random 70% of trials, and Light B flashed on a random 30% of trials. When asked to predict which light would flash next, the participants closely matched these proportions: in 70% of trials they predicted Light A, and in 30% of trials they predicted Light B. This "probability-matching" behavior has been found in a wide variety of tasks with adults, older children, and some non-human animals (Behrend & Bitterman, 1961; Bullock & Bitterman, 1962; Estes, 1964, 1976; Gardner, 1957; Myers, 1976; Stevenson & Weir, 1959). Increasing the complexity of the task can sometimes lead adults to stop matching the distributions in their data as closely and start regularizing, producing more skewed responses (e.g., Gardner, 1957; Weir, 1964). But although adults can behave variably in these tasks, an important finding that emerges is that children regularize remarkably consistently, particularly at young ages (Bever, 1982; Craig & Myers, 1963; Derks & Paclisanu, 1967; Stevenson & Weir, 1959; Weir, 1964).

How does the phenomenon of regularization in young learners bear on the mechanisms underlying first language acquisition? On what we are calling the *restrictive hypotheses* account, this phenomenon is informative about the domain-specific expectations that children bring with them to the learning task. In particular, regularization behavior may tell us something about the types of regularities that children believe they are are likely to encounter in a particular learning domain. In the domain of language, children may expect that the grammatical system generating their data contains particular sorts of restrictive rules (Bickerton, 1981, 1984; Chomsky, 1965, 1975; Lidz & Gagliardi, 2015; Lightfoot, 1991; Pinker, 1984; Senghas & Coppola, 2001). For instance, they may expect that determiner distributions are systematic. In non-linguistic domains, children may bring other domain-specific hypotheses about rule systems responsible for their data (e.g., Schulz & Sommerville, 2006). Regularization emerges when children can abstract away from unhelpful variability in their data— "noise" coming from parts of the language (or other domain) that haven't yet been acquired— in such a way as to identify signal for one of the restrictive hypotheses that they are considering.

This approach stands in contrast to an account in which learners have a flexible hypothesis space that can accommodate any degree of variability. On the *general regularization bias* account, children consider faithfully encoding the statistical distributions in their data, but are pushed to prefer extreme points in their gradient hypothesis space, arriving at generalizations that "sharpen" the statistics of their input and sometimes appear near-categorical (Austin et al., 2022; Culbertson & Kirby, 2016; Ferdinand et al., 2019; Hudson Kam & Newport, 2005, 2009; Perfors, 2012; Reali & Griffiths, 2009; Smith & Wonnacott, 2010). This account offers two sorts of explanations for the results of the artificial language learning experiments described above. Children may have considered the veridical distribution of determiner occurrences in their training data, but did not settle on this distribution due to a domain-general prior belief that probabilities close to zero or one are generally more likely to occur than intermediate values. Alternatively, children may have attempted to veridically encode the distribution of determiner occurrences, but failed due to constraints on developing cognitive systems that are shared across learning domains, such as information processing, cognitive control, or working memory (Austin et al., 2022; Hudson Kam & Newport, 2005, 2009; Newport, 1990, 1999). These cognitive constraints may conspire to skew learning towards more categorical outcomes. (For more on these issues, see Culbertson & Kirby, 2016; Ferdinand et al., 2019; Keogh et al., 2024; and Perfors, 2012.)

The general regularization bias proposal has been extensively investigated computationally (e.g., Culbertson et al., 2013; Reali & Griffiths, 2009; Perfors, 2012; Smith et al., 2017). However, despite its compatibility with prior experimental findings, no general computational architecture exists for modelling the restrictive hypotheses proposal. In the remainder of this section, we will first illustrate a computational approach that has been commonly adopted to implement a general regularization bias, taking as our case study a representative artificial language learning experiment reported in Austin et al. (2022). Specifically, we will show how a previously-proposed Bayesian learning model that includes a numerical bias in its prior can account for these results (Perfors, 2012; Reali & Griffiths,

2009). We will then introduce an alternative computational approach that implements the idea of learning in a noise-tolerant way with restrictive hypotheses.

2.2 A representative artificial language experiment

We will consider children's behavior in the "inconsistent language" condition from Austin et al.'s (2022) Experiment 1. Participants were trained on an artificial language with VSO word order and novel vocabulary items. The noun-phrases (NPs) in subject and object position each comprised a noun followed by a determiner. There were two determiners in the language, 'ka' and 'bo';[1] one was designated as the primary determiner, and appeared in 67% of determiner positions, and the other determiner appeared in the other 33% of determiner positions. Apart from this proportional requirement, the choice between 'ka' and 'bo' in any given determiner position was unpredictable. Sentences in this artificial language were used to label scenes with puppets interacting in short video clips. In the training sessions, participants were taught the names of these puppets and were asked to repeat the sentences that they heard. The scene to be described dictated whether the sentence was transitive or intransitive, and dictated the choices of nouns and verbs, but left open the choice of determiner. A single session of "sentence exposure" consisted of 24 sentences, each labeling a scene, of which 18 were transitive (two NPs each) and 6 were intransitive (one NP each). Each one of these sessions therefore exposed participants to 42 determiner-noun pairs in total. Of these, 28 had the primary determiner; for concreteness, we'll take 'ka' to be the primary determiner, so participants saw 28 occurrences of 'ka' and 14 of 'bo'.

After three training days, participants were then asked to produce their own sentences to describe scenes with new combinations of the puppets and actions. A single session of "production test" consisted of 12 scenes to be labeled, of which 8 were transitive (two NPs each) and 4 were intransitive (one NP each). Three of these production test sessions were conducted over the course of the experiment (on Days 3–5). Given sufficient training, a

[1] These determiners were actually 'ka' and 'po', but we're calling the second one 'bo' to avoid notational confusions with $p(\cdot)$.

participant could learn the word order of this language and the labels for the puppets and actions, and therefore learn to appropriately label scenes like the ones that they had been trained on. The interesting question is how they chose between 'ka' and 'bo' in determiner positions in these NPs. In line with the experimental results surveyed above, adult participants matched the proportions of 'ka' and 'bo' in their training data: they produced 'ka' in 67% of determiner positions, and 'bo' in 33% of determiner positions. The key finding for our purposes is that five- and six-year-olds strengthened the observed dominance of 'ka.' In the aggregate, they produced 'ka' about 86% of the time and 'bo' 14% of the time. Moreover, 6 of the 15 children at this age showed categorical behavior, using 'ka' all of the time and 'bo' not at all.

2.3 Regularization through a numerical bias

We can formally model learning in this experiment as the task of estimating an unknown parameter that governs the probability of 'ka' vs. 'bo' appearing in an NP in the language. Suppose that learners assume that a given determiner position will have 'ka' with some single unknown probability θ, and 'bo' with probability $1 - \theta$. For any particular probability θ, the *likelihood* of observing k instances of 'ka' out of m total determiners is the binomial probability $\binom{m}{k}\theta^k(1-\theta)^{m-k}$. This is the same as the probability of tossing a coin with a weight θ of coming up heads, and observing k total heads out of m total tosses. Given a particular estimate of θ, the likelihood of the learner's training data before their first production test on Day 3 (84 instances of 'ka' and 42 instances of 'bo')[2] is

$$P(\begin{smallmatrix}84\text{ 'ka'}\\42\text{ 'bo'}\end{smallmatrix} \mid \theta) = \binom{126}{84}\theta^{84}(1-\theta)^{42} \tag{1}$$

[2] We model learning prior to the first production test, as Austin et al. (2022) find no differences in children's behavior from the first through the third day of production testing. However, this approach could be generalized to model learning across the remaining days of the experiment.

The value of θ that maximizes this likelihood is 0.67, which is the proportion of 'ka' in the training data. In order to account for children's higher rate of 'ka' production, we need a learning model that does something other than simply maximizing likelihood.

Rational Bayesian reasoning can be applied to this problem by combining the likelihood of the data with a learner's *prior* beliefs about θ, before having seen any data. Bayes' Rule tells us that the *posterior* probability of a particular hypothesis about θ given the observed data, $p(\theta \mid {}^{84\text{ 'ka'}}_{42\text{ 'bo'}})$, depends on the likelihood of the data under θ, in (1), and the learner's prior degree of belief in θ, expressed in the prior probability distribution $p(\theta)$. The denominator is frequently calculated by *marginalizing* over all possible values of θ.

$$p(\theta \mid {}^{84\text{ 'ka'}}_{42\text{ 'bo'}}) = \frac{P({}^{84\text{ 'ka'}}_{42\text{ 'bo'}} \mid \theta)\ p(\theta)}{P({}^{84\text{ 'ka'}}_{42\text{ 'bo'}})} = \frac{P({}^{84\text{ 'ka'}}_{42\text{ 'bo'}} \mid \theta)\ p(\theta)}{\int_0^1 P({}^{84\text{ 'ka'}}_{42\text{ 'bo'}} \mid \theta)\ p(\theta)\ d\theta} \tag{2}$$

More generally, for data consisting of k observations of 'ka' and b observations of 'bo', the posterior probability distribution over θ is

$$p(\theta \mid k, b) = \frac{P(k, b \mid \theta)\ p(\theta)}{\int_0^1 P(k, b \mid \theta)\ p(\theta)\ d\theta} = \frac{\binom{k+b}{k}\theta^k(1-\theta)^b\ p(\theta)}{\int_0^1 \binom{k+b}{k}\theta^k(1-\theta)^b\ p(\theta)\ d\theta} \tag{3}$$

Following an approach introduced in Reali and Griffiths (2009) and pursued in Perfors (2012), we can account for children's regularization behavior by assuming that young learners' prior beliefs about θ are skewed. In particular, we can suppose that a learner's prior distribution $p(\theta)$ takes the form of a symmetric Beta distribution

$$p(\theta) = \text{Beta}(\alpha, \alpha) = \frac{\theta^{\alpha-1}(1-\theta)^{\alpha-1}}{B(\alpha, \alpha)} \tag{4}$$

where $B(\cdot, \cdot)$ is the Beta function.[3] The symmetric Beta distribution has a single α shape parameter, which governs the type and degree of skew in this distribution (Figure 1). When $\alpha = 1$, the distribution is uniform, so a learner with this prior will have no preference *a*

[3] $B(a, b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$. For positive integers n, $\Gamma(n) = (n-1)!$.

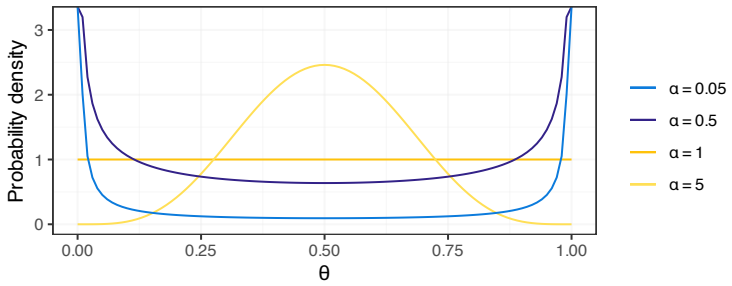


Figure 1. Illustration of symmetric Beta prior distributions, with different shape parameters

priori for any particular value of θ. When $\alpha > 1$, the distribution is unimodal, so a learner with such a prior will have some degree of preference for θ values close to 0.5. When $0 < \alpha < 1$, the distribution is u-shaped; a learner with this form of prior is biased towards regularization, preferring θ values close to zero or one. The strength of this bias is controlled by the numerical value of α. As α becomes smaller, the learner's prior preference for these extremes becomes stronger. But importantly, the prior is symmetric, so the learner has no prior belief about which of these two extremes is more likely.

A child with a u-shaped Beta prior will combine a preference for the endpoints of the distribution with the preference to fit the 67% observed proportion of 'ka' in the Austin et al. training data. For example, after observing 6 instances of 'ka' and 3 instances of 'bo,' such a learner will infer that θ is greater than 0.67: if $\alpha = 0.05$, the value of θ with highest posterior probability is 0.71. More concretely, if a learner's data consist of k observations of 'ka' and b observations of 'bo,' and the learner's prior takes the form of a symmetric Beta with parameter α, the value of θ with highest posterior probability under (3) is

$$\hat{\theta}_{\text{MAP}} = \frac{k + \alpha - 1}{k + b + 2\alpha - 2} \tag{5}$$

We see that because the posterior incorporates the learner's prior beliefs about θ, the value of θ that maximizes this posterior does not correspond to the proportion of 'ka' in the learner's data, but rather corresponds to what would be the proportion of 'ka' in a collection of $(k + \alpha - 1)$ 'ka' observations and $(b + \alpha - 1)$ 'bo' observations.

However, (5) also shows us that as k and b increase, this prior bias will play less of a role. In this sort of model, the prior can be overcome with sufficient data, resulting in posterior estimates that are very close to the maximum-likelihood estimate. This leads us to a wrinkle for the specifics of the Austin et al. (2022) experiment. Given the large amount of training data that children observe, such a prior will not by itself lead to visible regularization in the learner's posterior estimate of θ. So, accounting for children's regularization behavior with this approach requires a further assumption, which is that young learners' general cognitive limitations significantly limit the amount of data that they learn from (Keogh et al., 2024; Perfors, 2012; see also Newport, 1990, 1999). Only in combination with a mechanism that limits the size of the learner's data will this form of prior numerical regularization bias exert its influence. For the purposes of illustration, we use Perfors' 2012 simple model of "memory limitations," whereby a certain proportion of a learner's data is forgotten. For a learner with a "forgetting rate" of r, each data point that the learner observes in the input is dropped with probability r. (For more sophisticated ways to model memory limitations, see the further simulations in Perfors 2012, which find qualitatively similar results with several more complex models.) Note that this forgetting process does not asymmetrically target any particular variant in the learner's data; it merely reduces the size of the learner's intake (the remembered data) while leaving the proportion of 'ka' vs. 'bo' approximately unchanged.

In Figure 2, we plot the maximum *a posteriori* estimates of θ for models with varying degrees of prior regularization bias, and varying rates at which data is forgotten. Each bar represents an average across 10 randomly-generated intake datasets at a particular forgetting rate, from actual input data consisting of 84 'ka' and 42 'bo'. When the learner's regularization bias is weak or a large amount of data is remembered, the posterior estimates of θ are around 0.67, matching the actual proportion of 'ka' in the data (represented by the dashed line). But with an increasing bias towards regularization and less data remembered, the learner's posterior estimates of θ are more extreme.

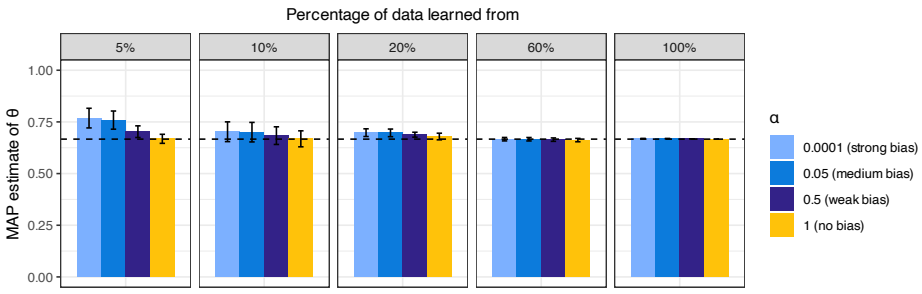


Figure 2. Maximum *a posteriori* estimates of θ across varying regularization biases, and varying amounts of data randomly sampled from 84 observations of 'ka' and 42 observations of 'bo'. The dashed line indicates the proportion of 'ka' in the training data.

This illustration shows us that a learner with a skewed Beta prior, in combination with cognitive limitations that reduce the amount of data to learn from, would infer that the true posterior probability of 'ka' in the language is greater than the rate of 'ka' in the training data. There are several ways that one might apply this result to model the patterns of children's responses in Austin et al. (2022). One might suppose that each child identifies the maximum *a posteriori* estimate of θ given her remembered training data, and uses that estimate to generate her own productions of 'ka' and 'bo' in the testing sessions. Or, a child might sample from the entire inferred posterior distribution over θ, choosing a value of θ on each trial in proportion to its posterior probability under (3). (For the details of these two approaches, see Reali and Griffiths 2009.) In either case, a child whose inferred posterior distribution is centered on a value sufficiently higher than 0.67 will show some degree of regularization at test, producing 'ka' at a higher rate than attested in the training data.

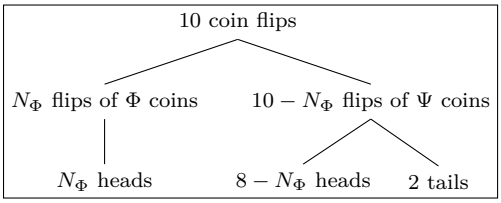
In summary, children's regularization behavior can be modelled by applying Bayesian inference with a symmetrical Beta prior distribution, whose degree of skew is given by the numerical α parameter. When the amount of data that children learn from is sufficiently small, possibly as a result of limited memory or other cognitive resources, this skewed prior enforces a preference for extreme points in a gradient space. This type of prior has been applied to many cases of regularization in experimental contexts (Culbertson et al., 2013; Perfors, 2012; Reali & Griffiths, 2009; Smith et al., 2017). There are two primary

assumptions inherent in this approach: (i) learners operate with a flexible hypothesis space that can accommodate any degree of variability, but (ii) they are biased *a priori* to expect that particular outcomes will occur in given contexts at rates close to zero or one. Varying the parameter of this prior can alter the strength of a learner's regularization bias, which may be needed to account for different degrees of regularization in different individuals or domains (e.g., Culbertson & Kirby, 2016; Ferdinand et al., 2019). But importantly for this account, the shape of the prior need not vary across domains: each of these cases of numerical regularization assumes a symmetrical skew in the learner's prior distribution, with no preference for one extreme over another. A learner comes to regularize simply by balancing this symmetrical prior with a desire to fit observed skews in the data.

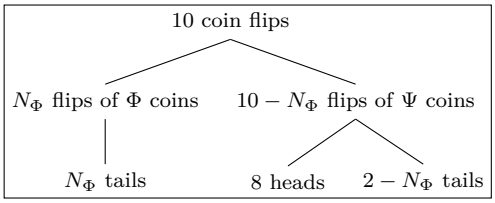
2.4 Regularization as selection among noisy hypotheses

We will now introduce the distinct approach to explaining regularization that is the main focus of this paper, which we offer as an alternative to the numerical approach described above. We will start with a simple illustration involving coin flips that will introduce the important intuitions, and then show how the Austin et al. (2022) experiment can be seen as an instance of the very same coin-flipping scenario. In Section 3, we will scale up the idea in order to apply it to the learning of grammars: finite systems that generate unbounded collections of sentences. The grammar-based system that we introduce there serves as the basis for the detailed case studies in Sections 4 and 5.

2.4.1 Illustration with coin flips. Suppose we have two bags of coins: one, which we'll call Bag H, contains coins where both sides are heads, and the other, Bag T, contains coins where both sides are tails. We will consider a scenario where we are trying to decide which bag's coins were responsible for generating some observed sequence of coin flips. The catch is that each bag also contains some unknown proportion of "noise coins", which we'll call Ψ coins, which all have some single unknown probability ψ of coming up heads (and $1 - \psi$ of tails). To decide which bag provides a better explanation of the observed



(a) Using Bag H



(b) Using Bag T

Figure 3. Partitioning 8 heads and 2 tails into signal and noise

collection of heads and tails, we will calculate the likelihood of the data under each bag, which will involve guesses about how many of the observed coin flips were flips of a Ψ coin — in other words, how many of the observed coin flips were "signal" bearing on the decision between Bag H and Bag T, and how many were noise. The difference in how this partitioning into signal and noise plays out across the two bags will be the basis of two bags' differing likelihoods, which will in turn lead the learner to choose one bag over the other.

We'll start by restricting attention to Bag H, where the "core" coins — we'll call these Φ coins — always produce heads. Suppose that ten times, a coin is drawn from this bag (with replacement) and flipped, producing eight heads and two tails. How many of these flips should we guess came from Φ coins, and how many from Ψ coins? There is a wide range of options, including the possibility that all ten flips came from the unpredictable Ψ coins. But given the observed skew towards heads, there is a clear intuition that the double-headed Φ coins were probably responsible for a significant portion of the observations. Why is this?

Figure 3a illustrates the possible ways of breaking down the ten observed coin flips, where N_Φ is the number of flips of Φ coins. Recall that if a coin has probability θ of coming up heads, then the probability of k heads given m flips is $\binom{m}{k}\theta^k(1-\theta)^{m-k}$. So if we suppose that all ten flips came from the unpredictable Ψ coins, i.e. $N_\Phi = 0$, the likelihood of the observed eight heads and two tails is $\binom{10}{8}\psi^8(1-\psi)^2$. Contrast this with the more intuitively plausible possibility that $N_\Phi = 6$. This way, it is guaranteed that the six Φ flips will come up heads ($\binom{6}{6}1^6 0^0 = 1$), so generating the observed data just amounts to having the four Ψ flips produce two heads and two tails, the likelihood of which is $\binom{4}{2}\psi^2(1-\psi)^2$. This is clearly less

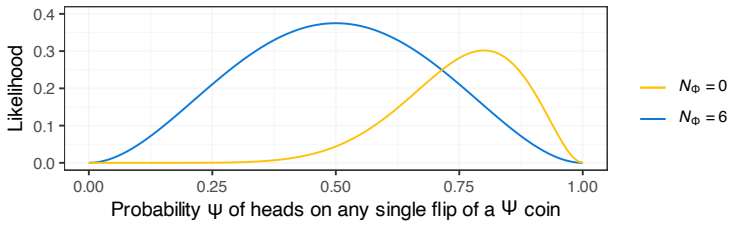


Figure 4. Likelihood of 8 heads and 2 tails under two hypotheses about N_Φ for Bag H

"costly" than the requirement that ten Ψ flips produce eight heads and two tails: by positing six Φ flips, six of the heads that we need to generate come for free, whereas positing that all of the flips were Ψ flips gives us no such head start.

These two likelihoods are graphed as a function of ψ in Figure 4. This shows that the $N_\Phi = 6$ hypothesis has a higher likelihood for "most" values of the unknown ψ (all those less than about 0.71), in line with our intuitive preference for hypotheses that invoke Φ flips. But we can make this more precise by marginalizing over ψ, as in (6) and (7).

$$(6) \qquad P(\,{}^{8\text{ heads}}_{2\text{ tails}} \mid N_\Phi = 0, \text{Bag H}) = \int_0^1 P(\,{}^{8\text{ heads}}_{2\text{ tails}} \mid N_\Phi = 0, \text{Bag H}, \psi)\; p(\psi)\; d\psi$$
$$= \int_0^1 \binom{10}{8}\; \psi^8\; (1-\psi)^2\; p(\psi)\; d\psi$$

$$(7) \qquad P(\,{}^{8\text{ heads}}_{2\text{ tails}} \mid N_\Phi = 6, \text{Bag H}) = \int_0^1 P(\,{}^{8\text{ heads}}_{2\text{ tails}} \mid N_\Phi = 6, \text{Bag H}, \psi)\; p(\psi)\; d\psi$$
$$= \int_0^1 \binom{4}{2}\; \psi^2\; (1-\psi)^2\; p(\psi)\; d\psi$$

These expressions make reference to $p(\psi)$, which encodes our prior beliefs about how likely the various possible values of ψ are. If we had reason to believe that some values were *a priori* more likely than others, then we might want to assign more "weight" to the corresponding portions of the x-axis in Figure 4. In the previous numerical regularization approach, such a belief could be expressed by assuming that $p(\psi)$ takes the form of a skewed Beta distribution (with $\alpha < 1$). But on our approach, any skewed learning outcomes will arise from a choice among discrete restrictive systems, rather than a numerically skewed prior. So, here we assume that $p(\psi)$ follows a uniform Beta(1,1) distribution, meaning that

all possible values of ψ are equally likely *a priori*. This means that the integrals in (6) and (7) simply amount to the areas under the two curves in Figure 4.

This assumption of a uniform Beta prior makes available particularly simple closed-form solutions for the integrals in (6) and (7).[4] Specifically, for any m and k:

$$\int_0^1 \binom{m}{k} \theta^k \, (1-\theta)^{m-k} \, p(\theta) \, d\theta = \frac{1}{m+1} \qquad \text{(when } p(\cdot) \text{ is the flat/uniform prior over } [0,1]\text{)} \tag{8}$$

This tells us that if a θ-weighted coin is tossed m times, then the probability of heads appearing k times is $\frac{1}{m+1}$. It may be surprising at first that this does not depend on k, but since we make no assumptions at all about θ, any value of k from the set of options $\{0, 1, \ldots, m\}$ is equally likely; and since there are $m+1$ options, each has probability $\frac{1}{m+1}$.

The integrals in (6) and (7) therefore evaluate to $\frac{1}{11}$ and $\frac{1}{5}$; these are the likelihoods of the data given $N_\Phi = 0$ and $N_\Phi = 6$, respectively, after marginalizing over ψ. More generally, if we hypothesize n_Φ flips of Φ coins, then we must invoke $10 - n_\Phi$ flips of Ψ coins, and so the likelihood is $\frac{1}{(10-n_\Phi)+1} = \frac{1}{11-n_\Phi}$. Even better than positing six Φ flips, then, is positing eight Φ flips, and leaving only two Ψ flips, for a likelihood of $\frac{1}{3}$.[5] Notice, however, that $N_\Phi = 8$ is as far as we can go: with more than eight Φ flips, the likelihood would be zero, because there are only eight heads in the data. The observed number of heads puts a cap on the degree to which large values of n_Φ can be used to achieve high likelihoods.

$$P(\,{}^{8\text{ heads}}_{2\text{ tails}} \mid N_\Phi = n_\Phi, \text{Bag H}) = \begin{cases} \dfrac{1}{11 - n_\Phi} & \text{if } n_\Phi \leq 8 \\ 0 & \text{otherwise} \end{cases} \tag{9}$$

The full range of ways that Bag H can explain the observed data is represented by the blue bars in Figure 5. The crucial point is that the better explanations make use of the

[4] There is a closed-form solution for any prior that takes the form of a Beta distribution. For simplicity we are leaving aside the more general case which is not relevant here. See the Appendix for more detail.

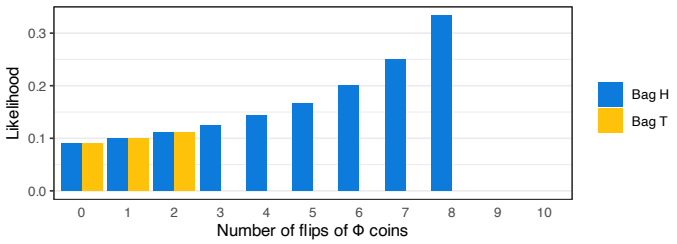


Figure 5. Likelihood of 8 heads and 2 tails

two-headed Φ coins to account for the large proportion of heads in the data.

Now let us consider how the same observed data might have arisen from Bag T, in which Φ coins always produce tails rather than heads. Here it is the two observed tails whose source is uncertain (see Figure 3b); the eight heads necessarily all came from Ψ coins. As with Bag H, the hypothesis that there were zero Φ flips is the most costly, and alternatives that make use of larger numbers of Φ flips provide higher likelihoods. But since Φ flips in Bag T produce certain *tails*, the number of such flips is capped at two. The best available explanation still leaves eight uncertain Ψ flips and therefore only achieves a likelihood of $\frac{1}{9}$; see (10). There is no way for the two-tailed Φ coins in Bag T to contribute to particularly good explanations of the observed high proportion of heads.

$$P(\begin{smallmatrix}8\text{ heads}\\2\text{ tails}\end{smallmatrix} \mid N_\Phi = n_\Phi, \text{Bag T}) = \begin{cases} \dfrac{1}{11 - n_\Phi} & \text{if } n_\Phi \leq 2 \\ 0 & \text{otherwise} \end{cases} \tag{10}$$

Figure 5 shows how this compares with the wider range of possibilities made available by Bag H. Since the explanations made available by Bag H range from very good ones to more costly ones, and Bag T makes available a subset at the costly end of that range, it is intuitive that Bag H is a better explanation overall — even though the presence of head-tail coins ensures that any observed combination of heads and tails is *compatible* with both bags.

[5] If this is surprising, it may be because of an intuition that we expect the Ψ flips to yield a roughly equal number of heads and tails, contrary to our assumption of a uniform prior on ψ. In combination with a prior encoding an expectation that ψ will have a value close to 0.5, the $N_\Phi = 6$ hypothesis can win out over $N_\Phi = 8$; such a prior would heavily weight the high part of the $N_\Phi = 6$ curve in Figure 4.

To make this more precise we can marginalize over N_Φ. This involves summing over the full range possible choices of $n_\Phi \in \{0, 1, \ldots, 10\}$, but (9) and (10) tell us that only certain subsets of that range will contribute non-zero values to the sum.

$$P(\,{}^{8\text{ heads}}_{2\text{ tails}} \mid \text{Bag H}) = \sum_{n_\Phi=0}^{10} \left[P(\,{}^{8\text{ heads}}_{2\text{ tails}} \mid N_\Phi = n_\Phi, \text{Bag H}) \times P(N_\Phi = n_\Phi \mid \text{Bag H}) \right] \tag{11}$$

$$= \sum_{n_\Phi=0}^{8} \left[\frac{1}{11 - n_\Phi} \times P(N_\Phi = n_\Phi \mid \text{Bag H}) \right]$$

$$P(\,{}^{8\text{ heads}}_{2\text{ tails}} \mid \text{Bag T}) = \sum_{n_\Phi=0}^{2} \left[\frac{1}{11 - n_\Phi} \times P(N_\Phi = n_\Phi \mid \text{Bag T}) \right] \tag{12}$$

The remaining factors $P(N_\Phi = n_\Phi \mid \cdots)$ in (11) and (12) are the probability that n_Φ of the ten draws yield Φ coins. We have assumed that we know nothing about the ratio of Φ coins to Ψ coins, which means that the probability of each value of n_Φ is uniform across the eleven values on the x-axis of Figure 5, following the logic from (8) again.[6] Each bag's likelihood is therefore proportional to the sum of its corresponding bars in Figure 5.

$$P(\,{}^{8\text{ heads}}_{2\text{ tails}} \mid \text{Bag H}) = \sum_{n_\Phi=0}^{8} \left[\frac{1}{11 - n_\Phi} \times \frac{1}{11} \right] = 0.138 \tag{13}$$

$$P(\,{}^{8\text{ heads}}_{2\text{ tails}} \mid \text{Bag T}) = \sum_{n_\Phi=0}^{2} \left[\frac{1}{11 - n_\Phi} \times \frac{1}{11} \right] = 0.027 \tag{14}$$

Notice that the product of $\frac{1}{11}$ and $\frac{1}{11-n_\Phi}$ inside the summations here corresponds to combining (i) a choice of how to split 10 coin flips into Φ flips and Ψ flips, shown at the root of the tree diagrams in Figure 3, and (ii) a choice of how to split $10 - n_\Phi$ flips of Ψ coins into heads and tails, shown on the right-hand side of the tree diagrams. This can straightforwardly be adapted to the more general scenario with h heads and t tails.

[6] Given different assumptions about the ratio of Φ coins to Ψ coins in the bags, we might want some portions of the x-axis of Figure 5 to be weighted more heavily than others. But the subset-superset relationship shown on the graph makes it clear that as long as our assumptions about this ratio are *the same* for the two bags, there is no way to distribute this weight that will make Bag T's total larger than Bag H's.

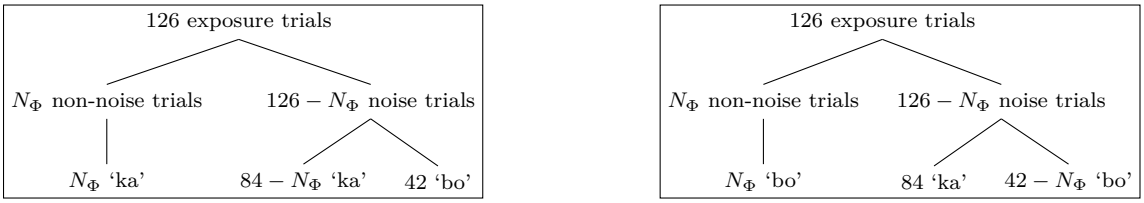
$$P(\,{}^{h\text{ heads}}_{t\text{ tails}}\mid \text{Bag H}) = \sum_{n_\Phi=0}^{h}\left[\frac{1}{(h+t+1)-n_\Phi}\times\frac{1}{h+t+1}\right] \tag{15}$$

$$P(\,{}^{h\text{ heads}}_{t\text{ tails}}\mid \text{Bag T}) = \sum_{n_\Phi=0}^{t}\left[\frac{1}{(h+t+1)-n_\Phi}\times\frac{1}{h+t+1}\right] \tag{16}$$

Given the likelihoods in (13) and (14) and a prior probability for each bag, we can calculate each bag's posterior probability. This applies Bayes' Rule, introduced in (3) earlier, but here we have two discrete hypotheses rather than a continuous range of hypothesized values for the parameter θ. Assuming an uninformative prior of 0.5 for each bag, we find that Bag H is about five times more likely to have generated the data than Bag T.

$$\begin{aligned} P(\text{Bag H}\mid {}^{8\text{ heads}}_{2\text{ tails}}) &= \frac{P(\,{}^{8\text{ heads}}_{2\text{ tails}}\mid \text{Bag H})P(\text{Bag H})}{P(\,{}^{8\text{ heads}}_{2\text{ tails}}\mid \text{Bag H})P(\text{Bag H}) + P(\,{}^{8\text{ heads}}_{2\text{ tails}}\mid \text{Bag T})P(\text{Bag T})} \\ &= \frac{0.138\times 0.5}{0.138\times 0.5 + 0.027\times 0.5} \\ &= 0.834 \\ P(\text{Bag T}\mid {}^{8\text{ heads}}_{2\text{ tails}}) &= 0.166 \end{aligned} \tag{17}$$

This choice between Bag H and Bag T will correspond to the choice between competing restrictive hypotheses in the learners we describe below. It is essentially a choice between the two-headed coins and the two-tailed coins, where either choice is embedded in a system where head-tail coins also produce some noise: divergences from what would be generated by the core mechanisms (the Φ coins) 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 (the Ψ coins) 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 (how much of the data came from Ψ coins) or what the contribution of noise looks like (the probability ψ of noise contributing a head).



(a) Under hypothesis G_{ka} (b) Under hypothesis G_{bo}

Figure 6. Partitioning 84 'ka' and 42 'bo' into signal and noise

2.4.2 Applying this to the Austin et al results. Suppose the participant is construing the task in Austin et al.'s (2022) experiment as one of choosing between two highly restrictive grammars, one of which uses 'ka' in all determiner positions, and the other of which similarly uses only 'bo'. Obviously neither of these grammars is directly compatible with the observed data in the exposure sessions. But we will consider each to be embedded in a noisy system, in the same way that the always-heads coins were embedded in Bag H and the always-tails coins were embedded in Bag T. One system, which we'll call G_{ka}, contains the always-'ka' grammar along with some noise-generating machinery that flips a coin to decide between 'ka' and 'bo', while the other, G_{bo}, contains the always-'bo' grammar along with that same noise-generating machinery. The observed data with 84 occurrences of 'ka' and 42 of 'bo' is therefore compatible with both G_{ka} and G_{bo}, but G_{ka} will provide a better explanation of the data, via exactly the same logic that we saw with the coins.

The ways in which the 126 total determiner observations could break down under each of these two hypotheses are illustrated in Figure 6, where again N_Φ is the number of observations for which the core mechanisms were responsible. Then following (15) and (16), the likelihood of the data under one of these hypotheses is a sum over possible values of this variable, where each contribution to the sum represents a particular choice among 127 ways to split the total exposure trials and among $(127 - n_\Phi)$ ways to split the noise trials. The two hypotheses differ only in the range of contributing values in the summation.

$$(18) \qquad P(\,{}^{84\ \text{`ka'}}_{42\ \text{`bo'}} \mid G_{\text{ka}}) = \sum_{n_\Phi=0}^{84} \left[\frac{1}{127 - n_\Phi} \times \frac{1}{127}\right] = 8.65 \times 10^{-3}$$

$$(19) \qquad P(\,{}^{84\ \text{`ka'}}_{42\ \text{`bo'}} \mid G_{\text{bo}}) = \sum_{n_\Phi=0}^{42} \left[\frac{1}{127 - n_\Phi} \times \frac{1}{127}\right] = 3.24 \times 10^{-3}$$

Since the data imposes a cap of 42 on the number of observations that can be attributed to the core mechanisms of G_{bo}, this hypothesis's likelihood is significantly lower than that G_{ka} hypothesis, where up to 84 of the observations can be attributed to the core mechanisms.

Using the likelihoods in (18) and (19), and assuming equal prior probabilities $P(G_{\text{ka}}) = P(G_{\text{bo}}) = 0.5$, we can calculate the posterior probabilities $P(G_{\text{ka}} \mid {}^{84\ \text{`ka'}}_{42\ \text{`bo'}}) = 0.728$ and $P(G_{\text{bo}} \mid {}^{84\ \text{`ka'}}_{42\ \text{`bo'}}) = 0.272$ just as we did in (17) for the bags of coins. We plot this posterior distribution in the rightmost panel of Figure 7. To set up a direct comparison with the simulation in Section 2.3, we also include the inferred posterior distributions for the previously-created intake datasets of different sizes, from actual input consisting of 84 'ka' and 42 'bo'. For the majority of these datasets, we see that the learner arrives at posterior distribution where G_{ka} is more than 2 times as likely as G_{bo}, on the basis of data where 'ka' was only 2 times as frequent as 'bo'. With 100% of the data, the learner infers that G_{ka} is 2.7 times as likely.

Thus, like we found in the numerical regularization approach, the learner arrives at a grammatical hypothesis that strengthens the dominance of 'ka' over 'bo' observed in the training data. However, here we see a difference in how learning interacts with the size of data. In the numerical regularization approach, a sharper skew in favor of 'ka' is a property of learning with very small amounts of data, and diminishes as more data is observed, as the likelihood overcomes the learner's prior regularization bias. In the current approach, a sharper skew in favor of G_{ka} only emerges as more data is observed, as the learner gains confidence in how to split the data into noise vs. non-noise. In the case studies that follow, we will test the learner's inference across datasets of varying sizes, but will remain agnostic

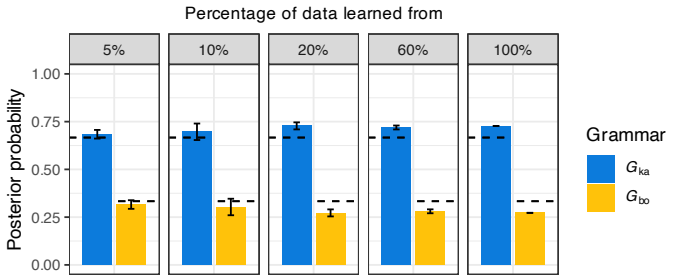


Figure 7. Posterior distribution over G_{ka} and G_{bo} across varying amounts of data randomly sampled from 84 observations of 'ka' and 42 observations of 'bo'. The dashed lines indicate the proportions of 'ka' and 'bo' in the training data.

about the precise amount of data being learned from; our method will be to examine the abstract skews in the posterior distributions that the learner infers.

There are a number of ways that we might take this approach to be the core of an explanation of the observed phenomenon of regularization in an artificial language experiment. We might imagine that each child arrives at a posterior distribution over grammars on the basis of the training data, and chooses to adopt either G_{ka} or G_{bo} in accord with this distribution; if 100% of the data is learned from, then 72.8% of children would produce only 'ka' at testing, and 27.2% would produce only 'bo'. Alternatively, to more closely model the reported results of the Austin et al. (2022) experiment, one can imagine that children choose (sample) either G_{ka} or G_{bo} on each production trial, with these same probabilities; this will again lead to production of 'ka' at a higher rate than observed in the training data. What we would like to emphasize is the following broader point: this approach, where the learner chooses among restrictive core mechanisms embedded in a system that also produces some noise, provides a candidate explanation for the phenomenon of regularization, and an alternative to the better-known numerical approach outlined in Section 2.3.

Perkins et al. (2022) applied this approach to model a naturalistic phenomenon in language acquisition that resembles regularization: how learners identify the core transitivity properties of verbs in their language— which verbs require objects— 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 deterministically require or deterministically disallow objects— for roughly the same reason that Bag H above provides a good explanation for data that do not consist entirely of heads, and that G_{ka} provides a good explanation for data that do not consist entirely of 'ka'.

In the naturalistic case studies that we consider below, we will be asking how learners identify a grammar of basic syntax that generates subjects and objects according to some canonical order (SVO, SOV, etc.), and/or with some canonical case-marking, yielding surface strings of verbs and noun phrases. Just like in Perkins et al. (2022), 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 the core grammatical rules that generated them. Learning will therefore require abstracting away from the noise in the data in order to identify the core rules of the target restrictive grammar.

3 Noisy CFG learners

In this section, we show how the idea of restrictive hypotheses operating alongside noise mechanisms can be incorporated into the learning of a system of grammatical rules. To illustrate, we'll first recast our account of the Austin et al. (2022) results in terms of a choice between two extremely simple grammars. Then we will show how to generalize the idea to more complex grammars made up of collections of interacting rules. To set up our case studies of basic syntax acquisition, we will formalize this as a case of Context-Free Grammar (CFG) learning, but this same approach can generalize to learning of other grammatical formalisms.

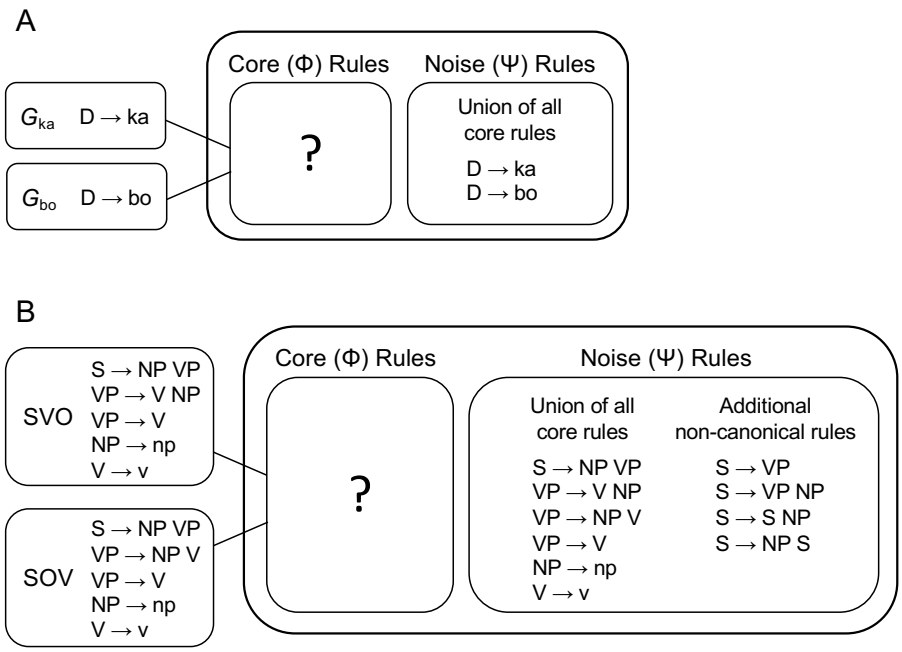


Figure 8. The hypothesis spaces of two example Noisy CFG learners

Probability	Rule
$1-\epsilon_{\mathrm{D}}$	$\mathrm{D} \rightarrow \mathrm{D}_{\Phi}$
ϵ_{D}	$\mathrm{D} \rightarrow \mathrm{D}_{\Psi}$
$\phi_{\mathrm{D}\rightarrow\mathrm{ka}}$	$\mathrm{D}_{\Phi} \rightarrow \mathrm{ka}$
$\psi_{\mathrm{D}\rightarrow\mathrm{ka}}$	$\mathrm{D}_{\Psi} \rightarrow \mathrm{ka}$
$\psi_{\mathrm{D}\rightarrow\mathrm{bo}}$	$\mathrm{D}_{\Psi} \rightarrow \mathrm{bo}$

(a) For G_{ka}

Probability	Rule
$1-\epsilon_{\mathrm{D}}$	$\mathrm{D} \rightarrow \mathrm{D}_{\Phi}$
ϵ_{D}	$\mathrm{D} \rightarrow \mathrm{D}_{\Psi}$
$\phi_{\mathrm{D}\rightarrow\mathrm{bo}}$	$\mathrm{D}_{\Phi} \rightarrow \mathrm{bo}$
$\psi_{\mathrm{D}\rightarrow\mathrm{ka}}$	$\mathrm{D}_{\Psi} \rightarrow \mathrm{ka}$
$\psi_{\mathrm{D}\rightarrow\mathrm{bo}}$	$\mathrm{D}_{\Psi} \rightarrow \mathrm{bo}$

(b) For G_{bo}

Figure 9. Compiled-out PCFGs for the two grammars in Fig. 8a

3.1 From surface forms to context-free grammar rules

To begin, we can represent the relationship between the G_{ka} and G_{bo} hypotheses from the previous section as in Figure 8a, where the two possible realizations of a determiner, 'ka' and 'bo', are expressed as two possible rules in the notation of Context-Free Grammars for rewriting the nonterminal symbol D.[7] Both of these rules are available as "Ψ rules" under either hypothesis; where the two hypotheses differ is which of the two rules they choose to

[7] A Context-Free Grammar specifies a set of rules for rewriting nonterminal symbols (in uppercase) as sequences of other nonterminal or terminal symbols (in lowercase). In the examples we consider here, the terminal symbols are words, and the nonterminal symbols are grammatical and phrasal categories.

include in their Φ component.

To implement the idea that each realization of D is mediated by a choice between whether it should be realized noisily or not, we can think of G_{ka} and G_{bo} as each taking the form of a single standard Probabilistic Context-Free Grammar (PCFG; Booth & Thompson, 1973; Wetherell, 1980) that "compiles out" the signal/noise distinction into two additional intermediate nonterminals, D_Φ and D_Ψ. These two PCFGs are shown in Figure 9.

Each rule in a PCFG is associated with a particular probability. The rule probabilities define a local distribution over possible rewrites for each nonterminal, so in each of these two grammars the probabilities of the two rules rewriting D_Ψ must sum to one (i.e., $\psi_{\text{D}\to\text{ka}} + \psi_{\text{D}\to\text{bo}} = 1$). Similarly, in these very simple examples that each only include one possible rewrite of D_Φ, the parameter $\phi_{\text{D}\to\text{ka}}$ in G_{ka} and the parameter $\phi_{\text{D}\to\text{bo}}$ in G_{bo} will necessarily be 1. But the generic notation in Figure 9 for these ϕ and ψ parameters will be useful when we extend to more involved examples. The parameter ϵ_{D} controls the choice between a noisy or non-noisy realization of D, and the ψ parameters specify the probabilities of 'ka' and 'bo' if a noisy realization is chosen. In terms of the earlier coins example, we can think of G_{ka} and G_{bo} as two "bags", where for each bag there is a parameter ϵ_{D} that expresses the ratio of Φ coins to Ψ coins, and there are ψ parameters that express the distribution over outcomes of the Ψ coins.

Each of the two compiled-out PCFGs in Figure 9 allows exactly three derivations. Those allowed by G_{ka} are shown in (20), along with their probabilities. Note that these rules implement precisely the branching structure in Figure 6a. An observed 'ka' might have been generated either via this grammar's Φ component or its Ψ component, but an observed 'bo' necessarily came from the Ψ component.

(20)

Derivation	Probability
$\text{D} \to \text{D}_\Phi \to \text{ka}$	$(1 - \epsilon_{\text{D}}) \times \phi_{\text{D}\to\text{ka}}$
$\text{D} \to \text{D}_\Psi \to \text{ka}$	$\epsilon_{\text{D}} \times \psi_{\text{D}\to\text{ka}}$
$\text{D} \to \text{D}_\Psi \to \text{bo}$	$\epsilon_{\text{D}} \times \psi_{\text{D}\to\text{bo}}$

The total probability assigned to the realization of D as 'ka' is the sum of the two corresponding derivations' probabilities.

The example we worked through in the previous section can be thought of as asking which of the two PCFGs in Figure 9 provides a better explanation for a corpus of 126 observed "sentences", each of which was just one word long; 84 were 'ka', and 42 were 'bo'. The system we propose below extends this idea to allow for noise not only in choices of surface forms (e.g. D → 'ka' or D → 'bo'), but also "higher level" choice points such as the order in which a sentence arranges its subject noun phrase and its verb phrase (e.g. S → NP VP or S → VP NP). Unlike the choice to realize D as 'ka', a choice to realize S as NP VP does not introduce observable surface forms, but rather other nonterminals, which themselves will be realized either noisily or non-noisily via other rewrite rules. The observed data will be unboundedly long strings, derived via sequential expanding rewrites of nonterminals, with the possibility of noise at each rewriting step. Our learner will ask which of a range of hypothesized grammars like the two in Figure 9 provides the best explanation for a corpus of such strings.

3.2 Noisy CFG learners: Model and key intuitions

Specifying the hypothesis space of a Noisy CFG learner involves specifying (i) a collection $R_1, R_2, R_3, \ldots$ of sets of context-free rules, and (ii) a further set R of context-free rules that has all the others as subsets (i.e. each $R_i \subseteq R$).[8] The learner will see data that has been generated by a composite system that has the full set R as its Ψ component, and has one of the subsets R_i as its Φ component; the learner's task is to choose which of the subsets is playing this role. In the introductory example above, $R_1 = \{\text{D} \rightarrow \text{ka}\}$, $R_2 = \{\text{D} \rightarrow \text{bo}\}$, and $R = \{\text{D} \rightarrow \text{ka}, \text{D} \rightarrow \text{bo}\}$; but in general there may be additional noise rules in R that are not part of any candidate Φ component. In the first of our two case studies below, the learner will have four different candidate sets of Φ rules, corresponding to the word orders

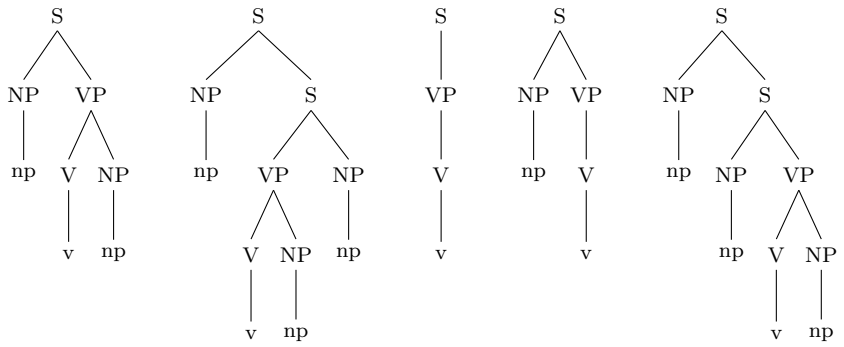
[8] More minimally, the system is completely determined by the collection of sets $R_1, R_2, \ldots$ and the set $R - \bigcup_i R_i$ of remaining rules, if any.

SVO, SOV, VOS and OVS, and the set of Ψ rules will include the rules for all four word orders plus some additional rules allowing for possibilities that are non-canonical under all four hypotheses, such as clauses that lack a subject. The outcome of the simulations we run using this learner will be a posterior distribution over these four sets of Φ rules, i.e. over these four different ways a canonical, or non-noise, (transitive) sentence can be realized.

Asking how well a particular hypothesized Φ component accounts for the observed data amounts to asking how likely those data are given the combination of that particular set of Φ rules and the shared set of Ψ rules — marginalizing over the individual ϕ parameters associated with the Φ rules, the individual ψ parameters associated with the Ψ rules, and also the ϵ parameters that control how likely each nonterminal is to be realized (non-)noisily.

In the simulations that we report below, the observed data take the form of a corpus of strings. The Φ and Ψ rules control the generation of tree structures that underlie those observed strings in the manner of CFGs; the probability of a string is a sum over the probabilities of its possible tree structures. To illustrate how different choices of Φ rules lead to better or worse explanations of the data, we will consider here only the task of calculating a likelihood for a collection of trees for a given hypothesized Φ component. In the case studies in Sections 4 and 5, we will use the tree-based calculations here as the basis for calculating string likelihoods (see the Appendix for more detail).

Consider the task of working out the likelihood of the collection of trees in Figure 10, for each of the choices of Φ components within the hypothesis space shown in Figure 8b. This is a slimmed-down version of the full hypothesis space that we use in the case study in Section 4, with just two candidate canonical word orders (SVO and SOV) and some non-canonical rules that only appear in the noise component, such as $\text{S} \rightarrow \text{VP}$ and $\text{S} \rightarrow \text{NP S}$. Via the standard independence assumptions of context-free grammars, observing this collection of trees amounts to observing the counts of rewrites shown in the table in Figure 10. Furthermore, we can treat the rewrites of S nodes independently from the rewrites of VP nodes: the likelihood of the trees is exactly the product of the probability of



Parent	Children	Freq
S	NP S	2
S	S NP	0
S	NP VP	3
S	VP NP	1
S	VP	1
VP	V NP	2
VP	NP V	1
VP	V	2

Figure 10

seven S rewrites breaking down as shown and the probability of five VP rewrites breaking down as shown. (The rewrites of the NP and V nodes have probability one, since there is no optionality, and therefore can be ignored.)

We will see shortly that these likelihoods can be calculated directly from the summary counts shown in Figure 10, but underlyingly each choice of Φ component actually corresponds to a compiled-out PCFG of the sort in Figure 9. These PCFGs generate trees that include additional layers of structure encoding the choice to treat each rewrite as either signal or noise. Two examples are shown in Figure 11; we call these *articulated trees*. The probability of a standard tree like the ones in Figure 10 is the sum of the probabilities of its articulated trees, just as the probability of 'ka' in (20) is the sum of the probabilities of two derivations, one proceeding via D_Φ and one proceeding via D_Ψ. What is more interesting about Figure 11 is that this choice between signal and noise mechanisms can be made independently at each node, even within the same tree. In both of these trees, the S $\rightarrow$ NP S rewrite at the top is analyzed as noise (i.e. via S_Ψ) — as it must be, since this rewrite does not occur in either of the two hypotheses' Φ components — but the S node that is introduced by that rewrite is itself realized *non-noisily* (via S_Φ). The two articulated trees differ in their treatment of the VP $\rightarrow$ V NP rewrite: the one on the left attributes this to the Φ component, and is therefore only compatible with the SVO hypothesis, whereas the one on the right attributes it to the Ψ component and is therefore compatible with either SVO or SOV.

Given the compiled-out PCFG for, say, the Φ component representing SVO word-order,

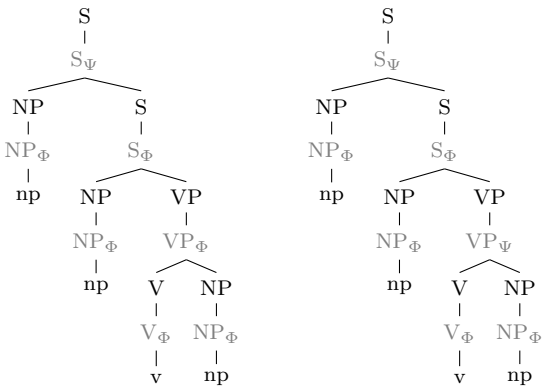


Figure 11. Two possible articulated trees for the rightmost tree shown in Fig. 10

and given specific values for all of the corresponding rule probabilities — one for $S \rightarrow S_\Psi$, which we could call ϵ_S, and one for $VP_\Phi \rightarrow V\ NP$, which we could call $\phi_{VP \rightarrow V\ NP}$, etc. — we could calculate a probability for each of the articulated trees in Figure 11. These are two of the many articulated trees whose probabilities we would sum to calculate the probability of the one corresponding tree in Figure 10. But what we really want is a likelihood, given the choice of the SVO Φ component, with the values of the ϵ, ϕ and ψ parameters (i.e., the PCFG's rule probabilities) marginalized out. We assume a uniform prior over all the possible settings of those parameters, which means that the desired likelihood can be calculated by making use of only rewrite counts like those shown alongside the trees in Figure 10, via essentially the same logic that we introduced in the simpler coins and 'ka'/'bo' examples.

Turning to the VP nodes first, we see that the five rewrites of VP resulted in two occurrences of the right-hand side V NP, one of NP V and two of just V. For this to have been generated by the SVO hypothesis, the NP V rewrite would necessarily have been noise, and the other four could be either noise or non-noise. The details of how this partitioning could work are shown in Figure 12, where M_Φ and N_Φ are random variables expressing the counts of the two non-noise possibilities.

Let us consider the likelihood of this observed collection of VP rewrites arising via a specific combination of noise and non-noise, i.e. via specific values $m_\Phi \in \{0, 1, 2\}$ and $n_\Phi \in \{0, 1, 2\}$. There are six possible ways that the five VP rewrites could have been split

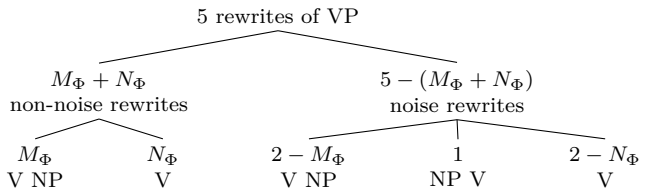


Figure 12. Partitioning the VP rewrites in Fig. 10 into signal and noise under the SVO hypothesis

between noise and non-noise, so the particular split that has $m_\Phi + n_\Phi$ non-noise rewrites has a probability of $\frac{1}{6}$. Similarly, the probability of the $m_\Phi + n_\Phi$ non-noise rewrites breaking into two groups as they did (on the left of Figure 12) is $\frac{1}{m_\Phi+n_\Phi+1}$. For the breakdown of the noise rewrites (on the right of Figure 12), we must consider the ways things can be split into *three* bins, rather than two as we've had in all other examples to this point. Just as in the two-bin case, however, all the different ways of doing the splitting have equal probability. The number of ways to split n objects into k bins is $\binom{n+k-1}{k-1} = \frac{(n+k-1)!}{n!(k-1)!}$. For $n = 5-(m_\Phi+n_\Phi)$ and $k = 3$, this is $\frac{(7-(m_\Phi+n_\Phi))!}{(5-(m_\Phi+n_\Phi))!2!}$. The overall likelihood is therefore

$$(21)\qquad P\left(\begin{smallmatrix}2\ \text{VP}\to\text{V NP}\\ 1\ \text{VP}\to\text{NP V}\\ 2\ \ \text{VP}\to\text{V}\end{smallmatrix},\ \begin{smallmatrix}M_\Phi=m_\Phi\\ N_\Phi=n_\Phi\end{smallmatrix}\ \middle|\ G_{\text{SVO}}\right) = \frac{1}{6}\times\frac{1}{m_\Phi+n_\Phi+1}\times\frac{(5-(m_\Phi+n_\Phi))!2!}{(7-(m_\Phi+n_\Phi))!}$$

and from here we can marginalize over the unknown values of m_Φ and n_Φ to find a likelihood conditioned only on the choice of G_{SVO}.

$$(22)\qquad P\left(\begin{smallmatrix}2\ \text{VP}\to\text{V NP}\\ 1\ \text{VP}\to\text{NP V}\\ 2\ \ \text{VP}\to\text{V}\end{smallmatrix}\ \middle|\ G_{\text{SVO}}\right) = \sum_{m_\Phi=0}^{2}\sum_{n_\Phi=0}^{2}\left[\frac{1}{6}\times\frac{1}{m_\Phi+n_\Phi+1}\times\frac{(5-(m_\Phi+n_\Phi))!2!}{(7-(m_\Phi+n_\Phi))!}\right]$$
$$= 6.07\times10^{-2}$$

The three fractions inside the summation in (22) correspond to the three branch-points in Figure 12, just as the fractions inside the summations in (18) and (19) correspond to the branch-points in Figure 6. The situation in (22) differs from the earlier examples in having non-determinism inside the Φ component — unlike the G_{ka} and G_{bo} hypotheses where there is only one non-noise outcome — but what matters is that the range of non-noise options is

more restricted than the range of noise options in the Ψ component. As long as there is this asymmetry, explanations that invoke more non-noise mechanisms will be preferred, because the non-noise options share probability mass with fewer competitors. The earlier examples illustrated the special case where there was *no* competition among non-noise options.

The situation for G_{SOV} is similar, except that m_Φ, rather than expressing how many of the two V NP rewrites to treat as non-noise, will now express "how many of" the one NP V rewrite to treat as non-noise. As in (18) and (19), this distinction has no effect on the fractions inside the summation, but constrains the range of values to sum over for m_Φ: rather than the cap of 2 in (22), for the SOV hypothesis it is capped at 1.

$$(23) \qquad P\left(\begin{smallmatrix}2\ \text{VP}\to\text{V NP}\\ 1\ \text{VP}\to\text{NP V}\\ 2\ \ \text{VP}\to\text{V}\end{smallmatrix}\,\middle|\, G_{\text{SOV}}\right) = \sum_{m_\Phi=0}^{1}\sum_{n_\Phi=0}^{2}\left[\frac{1}{6}\times\frac{1}{m_\Phi+n_\Phi+1}\times\frac{(5-(m_\Phi+n_\Phi))!2!}{(7-(m_\Phi+n_\Phi))!}\right]$$
$$= 3.71\times10^{-2}$$

An analogous calculation can be done for the frequencies of the various rewrites of S nodes in Figure 10. The likelihood of the trees in Figure 10 under any hypothesized Φ component is the product of the probabilities of the VP rewrites and the S rewrites. Since G_{SVO} and G_{SOV} include the same S rewrite rules in their Φ components, the S rewrites contribute the same probability for each hypothesis.

$$(24) \qquad P(\text{trees in Figure 10} \mid G_{\text{SVO}}) = P\left(\begin{smallmatrix}2\ \text{VP}\to\text{V NP}\\ 1\ \text{VP}\to\text{NP V}\\ 2\ \ \text{VP}\to\text{V}\end{smallmatrix}\,\middle|\, G_{\text{SVO}}\right) \times P\left(\begin{smallmatrix}2\ \ \text{S}\to\text{NP S}\\ 0\ \ \text{S}\to\text{S NP}\\ 3\ \text{S}\to\text{NP VP}\\ 1\ \text{S}\to\text{VP NP}\\ 1\ \ \text{S}\to\text{VP}\end{smallmatrix}\,\middle|\, G_{\text{SVO}}\right)$$
$$= (6.07\times10^{-2})\times(1.45\times10^{-3})$$
$$= 8.82\times10^{-5}$$
$$P(\text{trees in Figure 10} \mid G_{\text{SOV}}) = (3.71\times10^{-2})\times(1.45\times10^{-3})$$
$$= 5.39\times10^{-5}$$

The result is therefore that the posterior probability of G_{SVO} given the trees in Figure 10 is $\frac{8.82}{8.82+5.39} = 0.621$, compared to 0.379 for G_{SOV}.

3.3 A tractable inference mechanism for learning from strings

We have seen that a Noisy CFG learner corresponds to a collection of compiled-out PCFGs. From a collection $\vec{w}$ of observed strings, the goal is to infer the posterior distribution over this collection of grammars, $P(G \mid \vec{w})$.

We will use $\vec{\theta}^G$ to represent the vector of rule probabilities in a compiled-out PCFG G; this represents the full collection of ϕ, ψ, and ϵ probabilities. Let $\vec{\theta}^{A_G}$ be the weights of the allowable expansions of a given nonterminal A. We can represent the prior over $\vec{\theta}^{A_G}$ as a Dirichlet distribution with parameters $\vec{\alpha}^{A_G}$; the Dirichlet distribution is a generalization of the Beta distribution to cases with more than two possible outcomes. We assume here that all components $\alpha_i^{A_G}$ are equal to 1. Just as in the case of a Beta distribution with $\alpha = 1$, this results in a uniform prior distribution: the model has no preference for or against assigning probability mass to any particular expansions of A.

In the small example that we worked through in the section above, where we assumed that the strings came with observable tree structures (a set of trees $\vec{t}$, i.e., Figure 10), it was possible to calculate the posterior $P(G \mid \vec{t}, \vec{w})$ analytically by marginalizing over the rule weights ($\vec{\theta^G}$) and the partitionings of rewrites into signal and noise (the choice of articulated trees for each tree). But calculating the posterior $P(G \mid \vec{w})$ given only the strings would require further marginalizing over all the possible ways to choose a tree for each string in the data ($\vec{t}$). This calculation is intractable. So we instead infer the posterior $P(G, \vec{t} \mid \vec{w})$ via Gibbs sampling (Geman & Geman, 1984). After randomly initializing a set of possible trees for the observed strings, we alternate between sampling a new grammar according to $P(G \mid \vec{t}, \vec{w})$, and sampling new trees according to $P(\vec{t} \mid G, \vec{w})$. This process will converge to the joint posterior distribution over G and $\vec{t}$.

The step of sampling a grammar from $P(G \mid \vec{t}, \vec{w})$ involves exactly the calculations illustrated in Section 3.2. Note that any treeset $\vec{t}$ is compatible 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. Thus, for every grammar G, $P(\vec{t}, \vec{w} \mid G)$ is always non-zero, and so $P(G \mid \vec{t}, \vec{w})$ is also always non-zero, allowing us to draw samples from this posterior in a feasible way.

We sample trees from the posterior $P(\vec{t} \mid G, \vec{w})$ with a Hastings proposal (Hastings, 1970), using a variant of an algorithm introduced by M. Johnson, Griffiths, and Goldwater (2007) and marginalizing over $\vec{\theta^G}$. See the Appendix for details.

3.4 Summary

Section 2.4 introduced the idea of a learner that makes a choice among discrete restrictive hypotheses, each of which is embedded in a system that also produces some noisy, non-canonical output. In this section we have shown how to extend this approach to systems of interacting rules that can model the fundamentals of natural language syntax, to define what we call a *Noisy CFG Learner*. Given some data that reflects a mixture of core and noise mechanisms, the learner evaluates the following three questions, corresponding to the three branch-points in a choice for how to partition its data into noise and non-noise components: (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 of the core grammars in its hypothesis space, the learner considers the possible answers to these three questions in order to determine how well that grammar can explain the observed data. In the case studies below, we show how two particular aspects of syntax acquisition can each be set up as a problem of the form in Figure 8, with a choice among certain sets of core rules against the backdrop of a certain set of noise rules.

4 Case study 1: Learning basic word order

In the following case studies, we show that the approach of deciding among competing Noisy CFGs can be applied to model naturalistic phenomena in language acquisition that resemble regularization in previous artificial language experiments—cases of "regularization in the wild." Very early in development, children acquire the basic word order of their language from data that, from their perspective, is variable and messy. 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 and relative clauses (Hirsh-Pasek & Golinkoff, 1996; Lidz et al., 2017; Perkins & Lidz, 2020, 2021). In sentences like *What did you bring?* and *I like the toys that you brought*, a fronted phrase acts as the object of the verb in a non-canonical position, rather than post-verbally. These "non-basic" clause types are common: *wh*-questions comprise approximately 15% of the input to 1-year-olds (Cameron-Faulkner, Lieven, & Tomasello, 2003; Stromswold, 1995). Learners nonetheless manage to abstract away from the messiness in their representations of their data in order to draw an accurate inference about the structure of basic clauses.

Some accounts assume that learners have the ability to "filter" non-basic sentences of this sort, ignoring them when drawing early syntactic inferences (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.

In our first case study, we tested our model on child-directed English, French, and Japanese. Both English and French are canonically SVO, but allow different types of argument dislocation. Japanese is canonically SOV, but also has a large amount of argument dislocation, and moreover allows both subject and object drop. We show that our model sucessfully identifies the target grammars for its noisy data in all three languages. Moreover, our model out-performs a learner that is capable of encoding the full messiness of its data— its hypothesis space allows all word-order rules with some probability— but is numerically biased to prefer extreme points in that hypothesis space, favoring rule weights that are close to 0 or 1. Thus, we show that our approach fares better in this learning problem than the numerical regularization approach taken in prior literature.

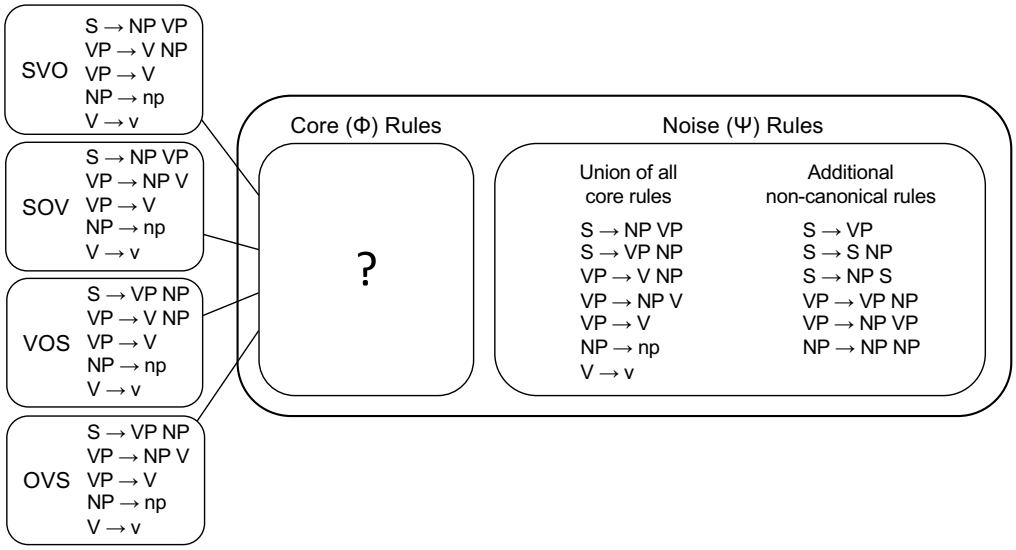


Figure 13. Hypothesis space of Noisy CFG learner for word order

4.1 Noisy CFG learner for word order

Our learner's hypothesis space is an expanded version of the sample hypothesis space for word order in Figure 8b. It consists of four sets of Φ-rules and one shared set of Ψ-rules, giving rise to the four Noisy CFGs in Figure 13. Here, we explain more fully the structure of this hypothesis space as it pertains to the current case study.

The Φ-rules generate the core predicate-argument structure of basic transitive and intransitive clauses. These rules have two important properties. The first is a degree of determinism: each hypothesized set of core rules deterministically puts 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.[9] The second property is a substantive expectation about the nature of the rules governing basic clause syntax. In each of these hypothesized grammars, subjects are are obligatory and objects are optional, reflecting the learner's belief that canonical clauses need subjects. Furthermore, noun phrases must occupy either

[9] 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.

canonical subject or canonical object position, reflecting the learner's belief that adjunction is in some sense non-canonical. In the simulations in this section, we show that these properties play a key role in the learner's ability to draw the correct word order generalizations.

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. These are analogous to some of the Ψ-rules that we saw in Figure 8b in Section 3. 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 or two-tailed coins, and flexible head-tail coins.

Crucially, while the learner's rules contain hypotheses about which canonical and non-canonical processes might be operative, the learner does not know ahead of time the ϕ, ψ and ϵ probabilities associated with these rules. It does not know how frequently it will encounter canonical transitive vs. intransitive clauses, and 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 Noisy CFG— the correct combination of core and noise rules— by marginalizing over all possible choices of these parameters, following the logic we saw above.

Our learner infers the posterior probabilities of the Noisy CFGs in its hypothesis space using distributions of imperfectly-identified noun phrases and verbs that a young infant might be able to represent. The learner performs this inference based on these string distributions alone, by using the inference mechanism described in Section 3. This does not require supplementary information about underlying clause structure; we note, however, that a similar mechanism could be generalized to make use of structural cues from other domains, such as meaning or prosody (Christophe, Millotte, Bernal, & Lidz, 2008; Morgan & Demuth, 1996; Pinker, 1984).

	English	**French**	**Japanese**
Corpus	Brown	Lyon	Miipro
# Children	3	5	4
Ages	1;6 – 5;1	1;0 – 3;0	1;2 – 5;0
# Words	380,423	515,827	328,502
# Utterances	85,787	139,800	115,368

Table 1
Corpora of child-directed English, French, and Japanese

4.2 Data

We used the CHILDES Brown, Lyon, and Miipro corpora (Brown, 1973; Demuth & Tremblay, 2008; Oshima-Takane, MacWhinney, Sirai, Miyata, & Naka, 1995), which contain speech directed to English, French, and Japanese learning children (see Table 1).

We followed a procedure for identifying verbs and noun phrases on the basis of distributional cues that an infant around the age of 15 months might be able to use. Based on empirical evidence, we assume that infants at this age can recognize a small number of functional elements of different sorts (e.g., Höhle, Weissenborn, Kiefer, Schulz, & Schmitz, 2004; Hicks, Maye, & Lidz, 2007; Shi & Melançon, 2010; Kim & Sundara, 2021; Mintz, 2013; Babineau, Shi, & Christophe, 2020). For English, for example, we assume that an infant can recognize *you* as a pronoun, *the* as a determiner, *-ed* as a verbal suffix and *will* as an auxiliary; these are four of the 75 elements that met our criteria for being recognizable, which included being among the 100 most frequent tokens in the corpus.

Once these recognizable elements have been tagged in the corpus, we use some simple heuristics to guess at the positions of verbs and noun phrases on the basis of only those tags. A noun phrase (np) is taken to appear wherever there is an element recognized as a pronoun or a name, and wherever there is an unrecognized element in a "np-cue position"; a verb (v) is taken to appear wherever there is an unrecognized element in a "v-cue position". The definition of the cue positions differs by language. For English, for example, np-cue positions include those following a determiner and those preceding the suffix *-s*, and v-cue positions include those following an auxiliary and those preceding the suffix *-ed*. For Japanese, np-cue

English	French	Japanese
0.34 np v	0.45 np v	0.63 v
0.27 np v np	0.20 np v np	0.23 np v
0.10 v	0.12 v	0.05 v np
0.08 v np	0.08 np np v	0.05 np np v
0.07 np v np np	0.05 v np	0.02 np v np
0.04 np np v	0.03 np np v np	
0.03 np np v np	0.03 np v np np	
0.02 v np np	0.01 np np np v	
0.01 np v np np np	0.01 v np np	
0.01 np np np v		
0.01 np np np v np		

Table 2
Proportions of most frequent string types in each language

positions are those preceding a case-marker, and v-cue positions include those preceding the negation marker *-nai*.[10] *Wh*-words and object clitics were not identified as np's, because they may not be recognized as such by infants learning basic word order (Perkins & Lidz, 2021; Brusini et al., 2017). Object clitics in French that are homophonous with determiners were treated erroneously as determiners, to simulate the uncertainty that infants might have about their category. Case-markers in Japanese were used merely as cues for identifying np's; their grammatical function (e.g., nominative or accusative) was not encoded, as we assume that a learner at this age does not yet know which case-markers mark subjects and which mark objects (a problem we consider in Section 5).

This allows us to map each corpus utterance to a string consisting of any number (possibly zero) of occurrences of np and v. From these we retain only those that contain exactly one v, since these are the ones relevant to the learner's question of how the elements

[10] A few more details about this procedure. Words in the corpus are split into multiple tokens either when indicated by the tilde character in CHILDES' tags, or at a small number of hand-identified affixes in each language. In English these are '-s', '-ed' and '-ing'; in French, '-é(e)(s)', '-er' and '-ons'; and in Japanese, '-nai', '-te', '-de', and '-ba'. The ambiguous English suffix '-s' (plural or third-person present) was not disambiguated: each time it occurs, it was counted as the plural suffix, because this is more frequent according to CHILDES' tagging. In English and French, the cue positions are the same: in addition to those in the main text, following a negation that follows an auxiliary counts as a v-cue position. In Japanese, the other v-cue positions are those preceding certain sentence-final particles (e.g. 'ne', 'ka', 'yo') or the suffixes '-te', '-de', and '-ba'. Of the 75 elements that we assumed are recognizable in English, 29 play a role in the definition of cue positions. In French, there are 76 recognizable elements, of which 27 are used as cues; in Japanese, there are 57 recognizable elements, of which 24 are used as cues.

of a single clause are arranged. The proportions of the relevant string types that we arrive at for each language are shown in Table 2.

We created datasets ranging in size from 10–50 strings, sampled according to the distribution in Table 2 for each language. Over 30% of the strings in each language are incompatible with the core rules of the target grammar (SVO for English and French, SOV for Japanese). 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.

4.3 Results

4.3.1 Noisy CFG learner. Figure 14 displays our model's inferred posterior probability distribution over the four Noisy CFGs in its hypothesis space, averaged over 10 runs of the model in each dataset. Visual inspection shows that the model's posterior distributions take the same qualitative shapes across datasets of varying sizes within each language. Just as in the artificial language 'ka'/'bo' simulations in Section 2.4, we see that asymmetries in these posterior distributions are weaker with smaller amounts of data, and become sharper as the amount of data increases. But while the amount of data presented to a learner can be controlled in an experimental setting, here it is an empirical question how many sentences a child is exposed to before identifying the target word order of the language; likely this number is much higher than 10 or even 50. For simplicity, we analyze the largest dataset presented to our model.

We find that 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 Welch's $t > 17.24$, all $p < 0.001$, Bonferroni correction for multiple comparisons). In Japanese, the SOV grammar had highest posterior probability compared to all other grammars (all $t > 7.14$, all $p < 0.001$). Thus, the learner's filtering mechanism allowed it to successfully overcome the large amount of noise in its data. The learner came to partition its data into noise and non-noise portions, in such a way that the non-noise portion provided a signal for

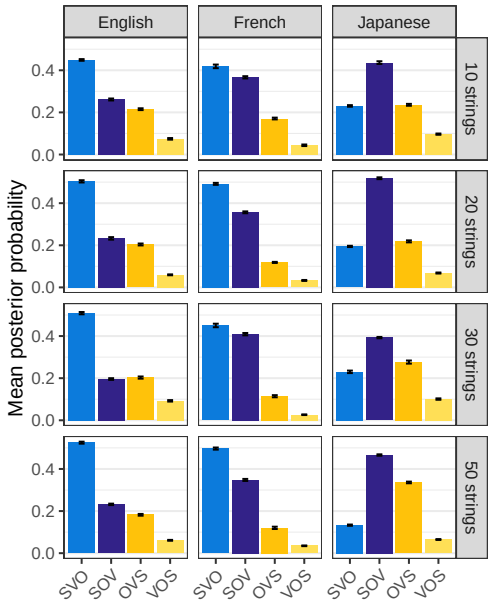


Figure 14. Posterior distribution over word-order grammars, Noisy CFG learner

the correct target word order.

Examining the behavior of the Japanese learner more closely, we find something interesting. With sufficient data, the learner in Japanese assigned the OVS grammar higher posterior probability than the SVO grammar, unlike in English and French. But this is puzzling: in Japanese, just like in English and French, the core rules of the OVS grammar can generate *fewer* strings in the learner's data than can the core rules of the SVO grammar. Only 7% of the strings in the Japanese dataset are compatible with the core OVS rules, compared to 25% compatible with the core SVO rules. Why, then, did the Japanese learner judge OVS more probable than SVO?

This difference reveals some subtleties in how the learner makes use of the distributions that it observes. Specifically, because Japanese allows argument-drop, the Japanese learner observed a relatively large number of bare `v` strings. This is the most common string type in the Japanese data, but comprises only about a tenth of the data in English and French. In order to analyze a string with only a `v` and no `np` satellites, the learner must make use of a noise rule that allows S to be rewritten as VP directly (S $\rightarrow$ VP), without an NP subject.

For Japanese, the learner came to identify that that this noise rule for omitting subjects was more useful than any of the core rules for introducing subjects, with a consequence for how it analyzed the np v strings in its data. Rather than taking these as evidence for a subject-initial grammar, the Japanese learner preferentially analyzed np v strings as subjectless clauses with an object-initial verb phrase, increasing the probability of OVS over SVO. In English and French, by contrast, the learner did not come to the same conclusion, because its data did not lead it to infer that the noise rule for omitting subjects had high probability. Thus, the different shapes of the posterior distributions in Fig. 14 arise in part because the learner correctly identified that subject-drop has high probability in Japanese, but not in English or French.

4.3.2 Comparison: A data-coverage heuristic. The results above illustrate that the learner's conclusions can diverge from what one would expect from a simple "data coverage" heuristic, where the best-fitting grammar is simply the one whose core rules can account for the most data. A further demonstration comes from considering comparisons between hypotheses that differ in restrictiveness in their core rules. To examine this situation, we added a fifth "free-order" grammar to the learner's hypothesis space, in which neither subjects nor objects have a fixed position. This grammar's core ruleset is the union of the core rules in the learner's four original restrictive grammars, and its noise rules are the same as those in the original four grammars.

Given a choice among the original restrictive grammars and this free-order grammar, the data-coverage heuristic will always favor the free-order grammar, since it generates the union of the stringsets generated by the original four. In Figure 15, we plot the predictions of the data-coverage heuristic and the model's inferred posterior over this five-way hypothesis space, for the 50-sentence dataset in each language. In each of the top panels, where a comparison only among the leftmost four grammars would have SVO or SOV as the winner, we see that the more flexible grammar fares better by the data-coverage metric. But our learner still assigned SVO higher posterior probability than any other grammar in the

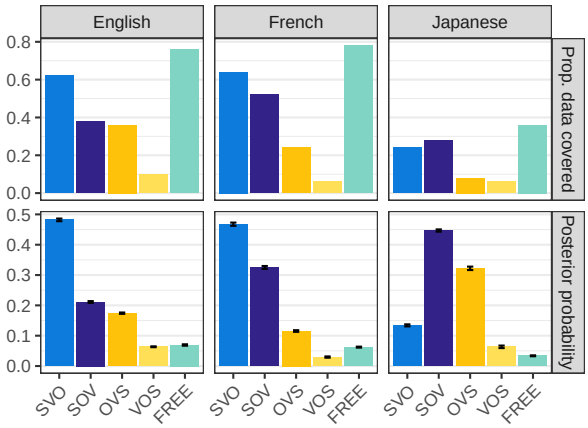


Figure 15. Five-way hypothesis space: Proportion data coverage vs. model's posterior distribution (50-sentence datasets)

hypothesis space in English and French, and SOV highest posterior probability in Japanese (Fig. 15, bottom, averaged across 10 runs of the learner; all $t > 4.80$, all $p < 0.001$).

Why does our learner still succeed at identifying the target word order in each language, even in the presence of another hypothesis that covers more of the data? Intuitively, our learner considers a tradeoff between fit to the data and restrictiveness of its hypotheses, an instance of what is often called "Bayesian Occam's Razor" (Tenenbaum & Griffiths, 2001; see also Maitra & Perkins, 2023). Given the choice between a restrictive hypothesis that provides a decent fit to the data, and a more flexible hypothesis that provides a slightly better fit, a preference emerges for the more restrictive option— just as we intuitively prefer to attribute as many coin flips as possible to a two-headed rather than a head-tail coin.

These findings demonstrate the flexibility and robustness of this learning mechanism. Our learner identifies the target strict 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 a restrictive word order in Figure 14 was not simply a by-product of the fact that we provided only restrictive 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 restrictive word order among

more flexible hypotheses.

4.3.3 Comparison: Numerical regularization. We now turn to the question of whether our model's success depends on a choice of discrete canonical word-order grammars in the learner's hypothesis space. To answer this question, 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. For this fully-flexible model, learning canonical word order would mean identifying that some of its rules have probabilities near zero. Within this fully-flexible architecture, we impose a numerical regularization bias in the same manner as for the artificial language 'ka'/'bo' simulation in Section 2.3, thereby assessing how our model compares to the general regularization bias approach taken in previous literature (Reali & Griffiths, 2009; Culbertson et al., 2013; Perfors, 2012).

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. This regularization bias takes the form of a skewed prior over the rule weights $\vec{\theta}$ in the learner's grammar (M. Johnson et al., 2007), analogous to the skewed prior used in Section 2.3. For each nonterminal A, we set all component parameters α_i^A of the model's Dirichlet prior to a small value, 0.0001. Because the Dirichlet distribution is a generalization of the Beta distribution, when $\alpha < 1$, this again has the effect of symmetrically skewing the learner's prior towards extreme values. Here, this biases the learner to put probability mass on only one expansion of a given nonterminal, and to push the probabilities of other expansions towards zero. Because the distribution is symmetric, the learner has no prior belief about which particular expansion is more likely. For simplicity, we test only one very extreme value of α in order to give the learner the best chance of regularizing successfully.

As this learner does not make a choice among discrete grammars, the way that it

would arrive at the target canonical word order for the language is to put most probability mass on the appropriate rules for subject and object position in its posterior distribution over $\vec{\theta}$. As an indication of whether the posterior distribution over $\vec{\theta}$ displays these properties, we examine the learner's posterior distribution over trees for its data. Our sampling process consists of just 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 $P(\vec{t} \mid G, \vec{w})$ in our original model.

We assessed whether the fully-flexible learner had identified a canonical word order by analyzing the trees in which the learner had identified a subject NP (daughter of S, sister of VP) and/or an object NP (daughter of VP, sister of V). For each sampled treeset in which at least one tree had a subject and at least one tree had an object, we calculated the proportion of subjects that appeared before the verb phrase, and objects that appeared before the verb. These proportions are plotted in Figures 16-17, where each point corresponds to a sampled treeset, aggregated across ten runs of the model in each language. For the sake of space, we plot the unbiased and biased model results together only for the 50-sentence dataset (Fig. 16). Results from the biased model for the smaller datasets are plotted in Fig. 17.

These plotted distributions provide an indication 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. For Japanese, we would expect to see the majority of tree samples close to the top right corner of the graph, corresponding to canonical SOV order.

To provide a direct comparison with the analysis of our Noisy CFG learner, we analyze the learner's samples for the 50-sentence dataset (Fig. 16). For the unbiased learner, the number of sampled treesets that we are able to analyze is high across the three languages: 99% of the English samples, 90% of the French samples, and 94% of the Japanese samples

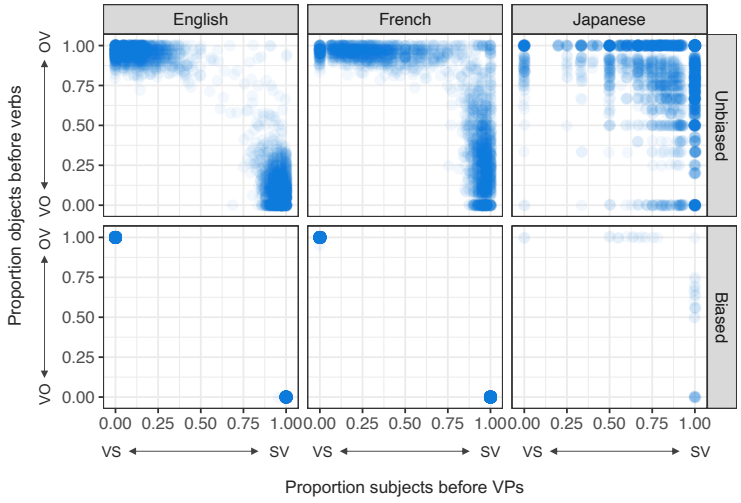


Figure 16. Posterior distribution over subject and object position in sampled treesets ($\vec{t}$), fully-flexible learner, 50-sentence dataset

contained at least one tree with a subject and at least one tree with an object. But across these analyzable samples, in each language the learner inferred distributions over tree structures that mirrored its noisy data. These ranged from the OVS (top-left) to the SVO (bottom-right) region in English, and across the OVS, SOV, and SVO regions in French and Japanese (Fig. 16, top). Thus, the unbiased learner failed to identify a restrictive word order for any of the three languages.

For the biased learner, the number of sampled treesets that we can analyze differs by language. Looking first at English and French, these numbers were high: 100% of the English samples and 99% of the French samples contained at least one tree with a subject and at least one tree with an object. In both languages, the biased learner inferred distributions over subject and object position that lie close to the corners corresponding to canonical word orders (Fig. 16, bottom). However, the learner assigned equal posterior probability to both OVS and SVO structures; the mean proportions of subject-initial and object-final trees were not significantly different from 0.5 in either language (English: both mean subject-initial and mean object-final = 0.51; French: both mean subject-initial and mean object-final = 0.54; all $t < 0.71$, all $p > 0.49$). The learner's numerical regularization bias led it towards the deterministic corners rather than the flexible middle, but it did not

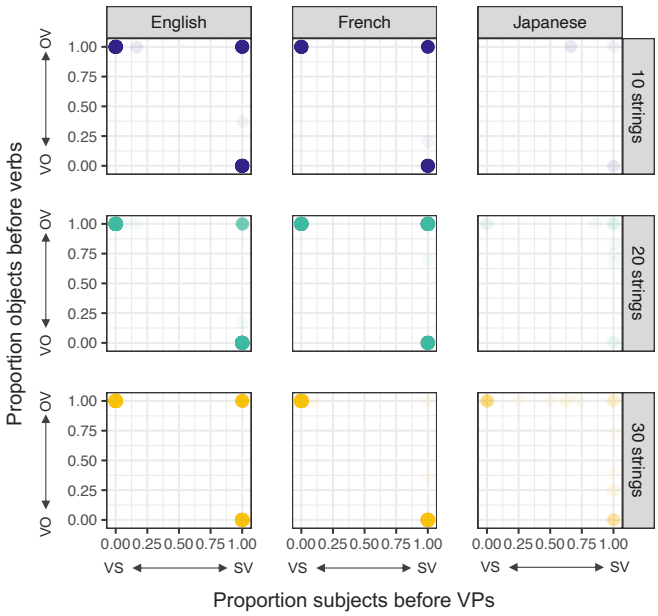


Figure 17. Posterior distribution over subject and object position in sampled treesets ($\vec{t}$), fully-flexible biased learner, smaller datasets

correctly identify SVO as a better corner than OVS. This same qualitative pattern is found with the smaller datasets as well (Fig. 17). With less data, the learner put additional probability mass on the SOV corner, but did not successfully discriminate SVO from OVS.

In Japanese, the number of sampled treesets that we can analyze for the biased learner is surprisingly low, resulting in very few data points plotted for that learner for any size of dataset in Figs. 16-17. For the 50-sentence dataset, only 1.6% of the learner's samples contain at least one tree with a subject and at least one tree with an object. Instead, the Japanese biased learner overwhelmingly converged to analyses in which np's occupied non-subject and non-object positions, introduced by recursive adjunction rules such as S $\rightarrow$ NP S. For the 1.6% of the sampled treesets that do contain a subject and an object, we can perform the same analysis as for English and French, and find that more of these samples are in the SOV quadrant (mean subject-initial = 0.83, $t(9) = 13.00$, $p < 0.001$; mean object-initial = 0.68, $t(9) = 2.53, p < 0.05$). But this analysis of the relative positions of subjects and objects tells us less about the learner's overall conclusions than it did for English and French, as the Japanese learner strongly preferred structures in which no

subjects or objects were present. This difference arises again from the Japanese learner's need to explain the prominence of bare v strings in its data. The Japanese learner concluded that the most probable way to introduce a VP is to use a rule with no NP subject, driving the probabilities of the subject-introducing rules towards zero. Similarly, it inferred that the most probable way to introduce a V is to use a rule with no NP object, driving the probabilities of the object-introducing rules towards zero. The flexibility in the learner's hypothesis space, in combination with the argument-drop in its data, led the biased Japanese learner away from analyses in which np's are clause arguments.

4.3.4 Role of substantive grammatical expectations. Unlike our Noisy CFG model, the fully-flexible learner did not identify the target canonical word order for its noisy data. Why does our approach perform better in this learning problem than the numerical regularization bias approach? There are two important properties of our model's hypothesis space that distinguish it from the fully-flexible learner, and may have allowed it to make better use of its observed data. First, our model's hypothesis space encodes an expectation that the grammatical systems generating its data comprise a mixture of restrictive core rules and processes that introduce distortions of those rules. Second, it also encodes substantive expectations about the content of those rules: specifically, it expects canonical clauses to have subjects and not to have adjuncts.

These two substantive expectations may have helped our learner in different ways across the languages tested. In English and French, we hypothesize that the crucial expectation is that canonical clauses require subjects. This may have allowed our learner to use one of the most common string types in its data— np v— as evidence for a subject-initial grammar. Given the choice between using its restrictive core rules to analyze the sole np as a canonical subject, versus using its noise rules to analyze the np in a different position (leaving the clause subjectless), a preference will emerge for the canonical-subject analysis, again paralleling our preference to analyze a sequence of heads as coming from a two-headed rather than a head-tail coin. In our comparison against the "data-coverage" heuristic above,

we saw that this preference for restrictive hypotheses can inform the learner's choice *across* grammars, when some are more restrictive than others. Here, we expect that this same mechanism applies *within* each grammar, governing the learner's choice of attributing data to the restrictive core rules vs. the flexible noise rules. 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 analyze a sole np in a specific clausal position.

In Japanese, we hypothesize that the crucial assumption on the part of our learner was that adjuncts are non-canonical. As discussed above for the fully-flexible learner, the prevalence of bare v strings in Japanese provides evidence that unary-branching structures (S dominating only VP, VP dominating only V) are common; and strings that do contain np's can make use of those common substructures by analyzing the np's as adjuncts. But for our learner, we expect that this temptation to analyze np's as adjuncts will be balanced against a motivation to use restrictive core rules rather than more flexible noise rules. This is likely what prevented our learner from analyzing all np's as adjuncts, and therefore allowed it to draw firmer conclusions about the positions of subjects and objects than the fully-flexible learner did.

Thus, the learner's two sorts of substantive expectations work in formally similar ways. Our learner's hypothesis space encodes an asymmetry between restrictive core rules and less restrictive noise rules, leading the learner to prefer core rule analyses. By allowing only particular sorts of core rules, we are able to express an expectation that canonical clauses have particular shapes. We suggest that this contributed in important ways to the learner's ability to draw sophisticated inferences from its noisy data.

We test this possibility with a further model comparison, focusing specifically on the learner's expectation that canonical clauses require subjects. We ask: does our learner's success in English and French depend on its substantive expectations about the core rules for subjects? Or would our learner succeed just as long as its hypothesis space encodes a

distinction between restrictive core rules and flexible noise rules, regardless of the nature of those core rules? Because the learner's expectations about subjects and adjuncts have formally similar effects on its learning mechanism, answering this question for one of these cases will speak to the broader question of whether the nature of the core rules matters in our learner's inference.

We constructed a comparison Noisy CFG model whose hypothesis space lacks the requirement that canonical clauses have subjects. Each of the grammars in this hypothesis space now includes S $\rightarrow$ VP within its core ruleset, allowing the core rules to produce subjectless analyses. The hypothesis space is otherwise identical to that of our original model. By the reasoning above, we predict that this change should not affect our learner's success in Japanese, but it should lead to worse performance in English and French.

Figure 18 displays the resulting posterior probability distribution over grammars that this comparison learner inferred for the 50-sentence datasets. These results are consistent with our predictions. Changing the learner's expectation about subjects in canonical clauses did not affect its success in Japanese: like our original model, this Japanese learner successfully assigned SOV highest posterior probability (all $t > 75.96$, all $p < 0.001$). But unlike our original model, this learner was unable to identify that English and French are SVO. In English, it inferred that SVO and OVS were tied as most probable (Welch's $t(16.93) = 0.49, p = 0.63$). In French, it assigned highest posterior probability to SOV (all $t > 6.98$, all $p < 0.001$), and did not differentiate between SVO and OVS as the next-most-probable options (Welch's $t(13.31) = 1.03, p = 0.32$). When the hypothesis space no longer encoded a requirement for subjects in canonical clauses, the English and French learners were not able to discriminate between subject-initial and object-initial grammars. Thus, it is not merely the presence of restrictive grammatical rules in our learner's hypothesis space that matters for its success in this learning problem; the content of those rules also plays a large role.

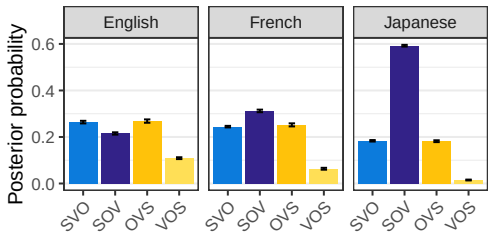


Figure 18. Posterior distribution over word-order grammars, Noisy CFG learner without subjects required canonically (50-sentence datasets)

4.4 Summary

Through this case study, we show that the approach of deciding among competing Noisy CFGs can be applied to a naturalistic language learning phenomenon which resembles the phenomenon of regularization observed in previous artificial language experiments. We find that our model can successfully acquire basic word order from the immature sentence representations available to an infant at early stages of syntax acquisition. Using distributions of imperfectly-identified noun phrases and verbs, our model successfully infers that English and French are SVO and Japanese is SOV, without further cues to underlying sentence structure. It does so by separating signal for canonical word order from noise coming from 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 restrictive 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.

Moreover, we find that our model out-performs a learner that instantiates the general regularization bias approach in prior literature (Reali & Griffiths, 2009; Culbertson et al., 2013; Perfors, 2012). We show that two properties of our learner are crucial for its success. First, the learner's hypothesis space encodes a distinction between restrictive core rules that

produce canonical clause structures, and more flexible noise rules that introduce distortions into its data. Second, it encodes substantive expectations about the nature of those core rules: that canonical clauses have subjects, and that noun phrases canonically occupy argument positions within a clause. Without these properties, the learner is unable to identify signal for the target word order within its noisy data.

These findings suggest that, for this learning problem, it is important for learners to have specific expectations about the types of grammatical rules that may have generated their data, embedded within a broader system that introduces noise. By encoding a distinction between restrictive grammatical rules and noise processes, our learning architecture provides a natural way to express linguistically-motivated expectations about the nature of basic clause syntax. And by embedding these expectations within a noise-tolerant system, the architecture makes it possible for a learners to recover the target basic clause structure for their language from data that appear inconsistent on the surface.

5 Case study 2: Learning case-marking

In our second case study, we demonstrate how our approach generalizes to another learning problem within early morphosyntax acquisition. We consider the problem of acquiring a case-marking system in which subjects and objects are each marked with a particular affix, as in Japanese. Early in language development, children acquiring case-marking languages come to identify how specific affixes function as case-markers, despite sometimes variable and inconsistent evidence in their input (e.g., Fisher, Jin, & Scott, 2019; Suzuki & Kobayashi, 2017; Suzuki, 1999; Matsuo, Kita, Shinya, Wood, & Naigles, 2012; Göksun, Küntay, & Naigles, 2008). Here, we model a stage in development in which a child may have identified certain affixes as candidate case-markers, but does not know which marks subjects (nominative) and which marks objects (accusative). We show how the system that we have introduced above can be applied to this scenario. This case study demonstrates how learning problems that seem on the surface to be very different—

acquiring morphological marking vs. clause structure— can both be expressed within the same kind of system, as a choice between Noisy CFGs.

We conducted simulations on a new dataset generated from child-directed Japanese, in which noun phrases sometimes co-occur with the *ga* (nominative) and *o* (accusative) suffixes. Because Japanese has scrambling and argument-drop, and the pronunciation of case-markers is optional, this dataset contains noisy and sparse evidence for the grammatical function of these affixes. We consider a learner that is attempting to identify which grammatical relation *ga* and *o* are each marking at the same time as identifying the canonical positions of subjects and objects. We show that our model simultaneously identifies *ga* as nominative and *o* as accusative, along with canonical SOV word order. Moreover, as in the previous section, our approach out-performs a learner that does not make a discrete choice among case-marking grammars, but rather implements a numerical regularization bias over a fully-flexible hypothesis space.

5.1 Noisy CFG learner for case-marking

Our learner's hypothesis space consists of augmented versions of the Noisy CFGs for the word order learner in the previous case study. We expanded the learner's hypothesis space in Section 4 to include Φ rules producing all four possible orders of subjects and objects, crossed with both possible mappings of the case-markers *ga* and *o* to subjects vs. objects. This gives rise to the eight Noisy CFGs in Figure 19. The Φ-rules mirror the four grammars our word-order learner considered: they generate basic transitive and intransitive clause structures, with subjects obligatory and objects optional. But here subjects and objects are represented with distinct nonterminal symbols (NPS and NPO), and are re-written as terminal symbols with distinct case-marking (np-ga and np-o). Each case marker deterministically realizes either subjects or objects, yielding a 2-way choice for case-marking systems: either subjects are rewritten as np-ga and objects as np-o, or objects are rewritten as np-ga and subjects as np-o. Adjunct NPs are now represented by a distinct

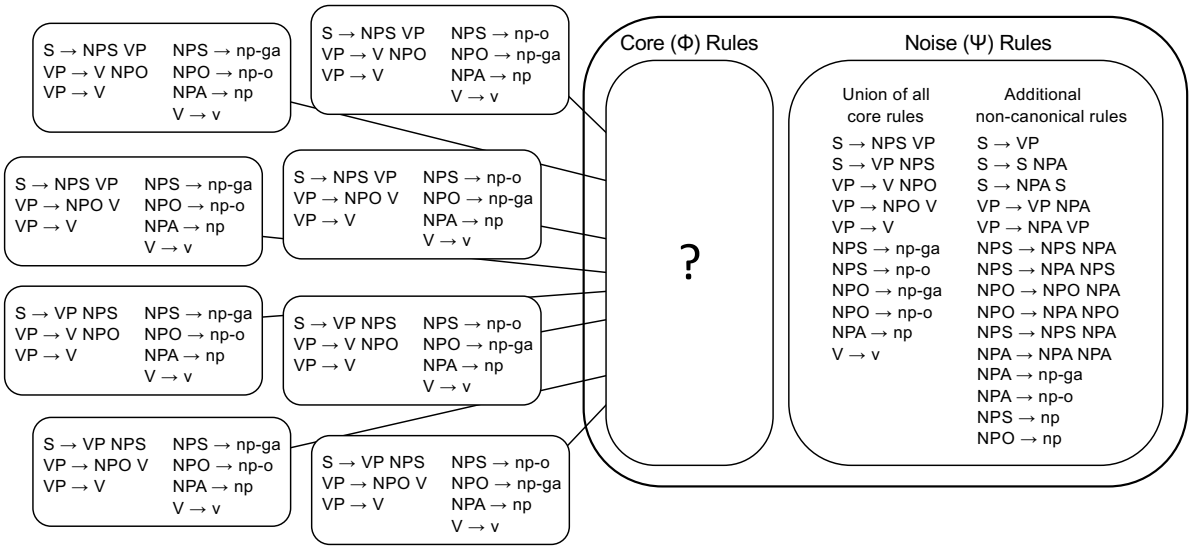


Figure 19. Hypothesis space of Noisy CFG learner for case-marking

nonterminal symbol, NPA, and these are canonically rewritten as np, i.e., noun-phrases that are neither marked with *ga* nor with *o*.

All grammars again share the same set of Ψ rules. As before, these comprise a superset of the union of the core rules across the learner's eight grammars, allowing for permutations and deletions of subjects and objects, and insertions of NPs into non-argument positions. In addition, the noise rules allow any type of NP to be rewritten by np-ga, np-o, or np. Thus, as before, we encode an asymmetry between highly restrictive core rules for case-marking and flexible noise rules that produce many more possibilities for expanding a given nonterminal. Here, the noise operates both internal to a tree, manipulating the position of NP nonterminals, and at a tree's frontier, manipulating the position of np terminal symbols.

Similar to our previous case study, the learner observes strings of imperfectly-identified verbs and noun phrases, some affixed with *ga* and *o*. Without knowing ahead of time the rate of noise or its properties, the learner evaluates how to partition its data into portions generated by the core case-marking and word-order component, and portions generated from noise. Comparing across the eight grammars, the learner selects the one that provides the best explanation for its data, thereby identifying both the target system of

Japanese	
0.45 np-ga v	0.03 np np np-ga v
0.14 np np-ga v	0.02 np-ga np v
0.13 np-o v	0.02 np v np-ga
0.07 v np-ga	0.01 v np np-ga
0.05 np np-o v	0.01 v np-o
0.04 np-ga v np	

Table 3
Proportions of most frequent string types, case-marking learner

nominative/accusative case-marking and the basic word order of the language.

5.2 Data

We again used the CHILDES Miipro corpus of child-directed Japanese (Oshima-Takane et al., 1995), and followed a procedure similar to the one described in Section 4.2 to arrive at the distribution of strings shown in Table 3. The noun phrases that were all represented simply as np in the previous case study were subdivided into three types: those that were followed by 'ga' (np-ga), those that were followed by 'o' (np-o), and all others (np).[11] We restrict the dataset to strings in which there is exactly one v, as before, and furthermore retain only strings that contain at least one case-marked noun phrase (either np-ga or np-o), since the learner's goal is to identify case-marking. This restriction means that this model does not learn from strings consisting of bare v, unlike the word-order learner.

For simplicity, we test our learner on one size of dataset, comprising 50 strings sampled in their relevant proportions, given in Table 3. 55% of the strings are incompatible with the core rules of the target grammar (SOV, with np-ga realizing subjects and np-o realizing objects), and no strings contain both case-markers simultaneously. Thus, the learner's evidence for the target grammar is noisy and sparse.

[11] This represents a simplification of the learning problem faced by a Japanese-learning child, who must identify the functions of *ga* and *o* among several other case-markers in the language. How a learner solves this more complicated problem is an important question that we leave for future work.

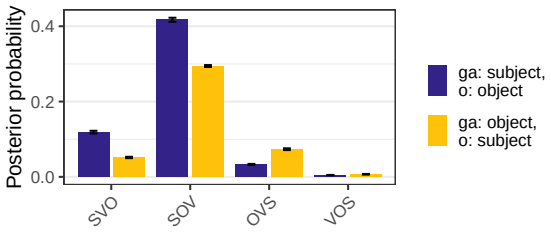


Figure 20. Posterior distribution over case-marking and word-order grammars, Noisy CFG learner

5.3 Results

5.3.1 Noisy CFG learner. Figure 20 displays the model's inferred posterior probability distribution over the eight Noisy CFGs in its hypothesis space, averaged across 10 runs of the model. The SOV grammar in which np-ga realizes subjects and np-o realizes objects was assigned significantly higher posterior probability than any other grammar (all $t > 20.51$, all $p < 0.001$). Thus, the learner successfully chose SOV as the correct word order simultaneously with identifying the correct case-marking system in Japanese.

5.3.2 Comparison: Numerical regularization. To assess how our approach compares to the more common regularization approach in this learning problem, we again constructed a comparison learner with a "fully-flexible" hypothesis space that does not encode a distinction between canonical rules for case-marking and non-canonical noise processes. As before, this hypothesis space consists of a single grammar whose rules include all rules in the core and noise components of our learner's eight Noisy CFGs. These rules allow all possible word order options and all possible mappings of case-marked np's to clause arguments and adjuncts. Identifying a case-marking system would mean identifying that np-ga is introduced with probability near 1 by only one clause argument (NPS or NPO), and that np-o is introduced with probability near 1 by the other.

We again tested two versions of this fully-flexible learner: one that assumes that all rules are equally probable *a priori*, and one with a strong numerical bias to put most probability mass on a single expansion for a given nonterminal (all α_i of the learner's Dirichlet prior $= 0.0001$). For the learner to match our model's success, we would expect to

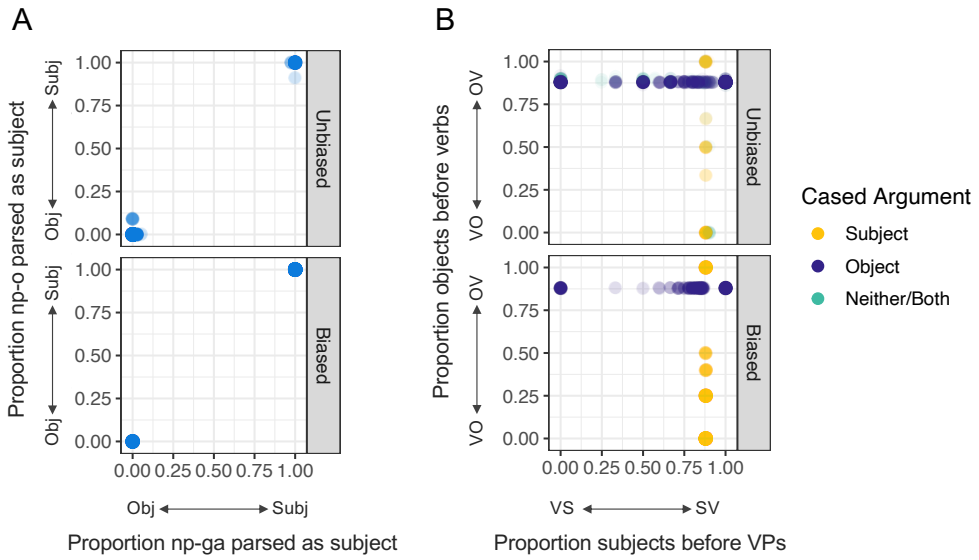


Figure 21. Posterior distribution in sampled treesets ($\vec{t}$) over (A) np-ga and np-o and (B) subject and object position, fully-flexible learner

see it converge to a posterior in which it puts most probability mass on the correct rules for introducing np-ga and np-o, and also on the correct rules for subject and object position.

To assess the learner's inference about case-marking, we examined its posterior distribution over trees for its data, $P(\vec{t} \mid G, \vec{w})$, following the same Gibbs sampling procedure described in the previous section. For each of these treesets, we calculated the proportion of np-ga that the model had parsed as NPS vs. NPO, and the proportion of np-o that the model had parsed as NPS vs. NPO. All of the treesets had at least one np-ga parsed as a subject or object, and at least one np-o parsed as a subject or object, so these proportions are well-defined. These proportions are plotted in Figure 21a.

If the learner had successfully mapped np-ga to subjects and np-o to objects, then we would expect to see the majority of tree samples in the lower right corners of the plots. Instead, we see that this fully-flexible learner failed to identify the target case-marking system of Japanese. The learner's posterior distribution over case-marked np's lies in two different corners: analyses in which np-ga and np-o both realize subjects (upper right corner), or analyses in which they both realize objects (lower left corner). Both the biased and the unbiased learner showed an approximately similar pattern, with the biased learner converging

more strongly to these deterministic corners. The unbiased learner placed significantly higher probability mass on analyses in which np-ga and np-o realize objects (mean proportion np-ga as subjects = 0.13; mean proportion np-o as subjects = 0.13; both significantly different from 0.5, $t > 9.04$, $p < 0.001$). The biased learner assigned approximately equal posterior probability to analyses in these two corners, with the mean proportion of np-ga and np-o parsed as subjects not significantly different from 0.5 (mean proportion np-ga as subjects = 0.73; mean proportion np-o as subjects = 0.73; both $t < 2.04$, both $p > 0.07$). Importantly, the learner did not treat np-ga and np-o differently in terms of the clause arguments that they realize; instead, it preferred to treat both as markers of the same argument. Thus, while the learner's numerical regularization bias led it to more strongly prefer certain deterministic corners of its hypothesis space, the pattern that emerges is not a case-marking system.[12]

We assessed the fully-flexible learner's inference about word order by performing the same analysis described in Section 5.3.3 for our word order learner. For each sampled treeset in which at least one tree had a subject and at least one tree had an object, we calculated the proportion of subjects analyzed before the verb phrase, and the proportion of objects before the verb. These proportions are plotted in Figure 21b. Compared to the fully-flexible Japanese learner in Section 5, more of the sampled treesets had NP's in argument positions and were therefore analyzable using this method: 56% of the sampled treesets for the unbiased learner, and 65% of the sampled treesets for the biased learner.

If the learner had identified that Japanese has SOV word order, we would expect to see the samples centered on the upper right corner of these plots. Instead, the learner inferred a different distribution, which is similar for both the biased and unbiased learner, with more spread for the unbiased learner. The learner converged to two different sorts of analyses: either subjects occur 88% of the time before the VP, with object position varying; or objects

[12] One might wonder if this is a consequence of the learner solving two problems in tandem: jointly acquiring case-marking and word order. We re-ran these simulations in a context where SOV word order is known, i.e., all rules for introducing clause arguments fix the subject before the VP and the object before the verb. We found the same qualitative pattern, both for the fully-flexible learner and for our Noisy CFG learner.

occur 88% of the time before the verb, with subject position varying. This does not resemble the fully-flexible learner's inference about Japanese word order in Section 5. However, when we examine how subject and object position interacts with case-marking, a clearer picture emerges. Recall that the overwhelming majority of treesets either treated both *ga* and *o* as markers of subjects, or treated both as markers of objects. For treesets in which case-marked np's were analyzed as subjects (in yellow), the learner preferred to analyze those arguments as 88% preverbal, and did not fix the position of its non-case-marked objects. For treesets in which case-marked np's were analyzed as objects (in purple), the learner preferred to analyze those arguments as 88% preverbal, and did not fix the position of its non-cased-marked subjects. Thus, the learner converged to a solution in which case-marked np's are 88% preverbal, and non-case-marked np's vary in their position. This solution captures a pattern in the learner's dataset: approximately 88% of case-marked np's appear before the verb.

Thus, the fully-flexible learner again displayed a form of regularization, coming to prefer a more deterministic rule system for its variable data. The learner preferred to analyze *ga* and *o* as markers of the same argument, and took "before the verb" to be the defining property of this marked argument. But these tendencies towards determinism did not lead it in the right direction: it did not regularize in such a way as to identify either a case-marking system or a canonical word order.

5.4 Summary

Here we show that our approach of choosing among Noisy CFGs generalizes to a second naturalistic phenomenon that looks on the surface qualitatively different from word-order acquisition: learning which of two candidate case-markers marks subjects and which marks objects in a language like Japanese, from data that provide noisy and sparse evidence. We find that our model successfully identifies the grammar of Japanese nominative/accusative case-marking, using only distributions of verbs and noun phrases affixed with *ga* and *o* in child-directed speech. Japanese allows clause arguments and

case-markers to be dropped, introducing a large amount of noise into the learner's data. Our model nonetheless succeeds at separating out noise from signal for the canonical case-marking grammar, in tandem with learning the canonical word order of Japanese. By contrast, a learner with a numerical regularization bias and no discrete choice of grammars falls short. The biased fully-flexible learner identifies more deterministic analyses for its variable data, but it does not identify a system of case-marking, in which *ga* and *o* are identified as markers of different clause arguments.

Why does our model again out-perform the general regularization bias approach for this learning problem? We can again consider the contribution of the two properties that distinguish our approach. First, our learner's hypothesis space encodes a distinction between restrictive core rules governing the canonical positions and morphology of clause arguments, and more flexible noise rules that distort those distributions. Second, it encodes substantive expectations about the nature of the core rules. Similar to our previous case study, the learner's substantive expectations arise from knowledge of the grammatical system that it is trying to identify: here, knowledge that it is learning a nominative/accusative case-marking system, and therefore that *ga* and *o* each canonically mark a different clause argument. The fully-flexible learner's hypothesis space does not distinguish between a canonical system of morphological marking and non-canonical noise processes, and therefore does not encode an assumption that the positions of *ga* and *o* are canonically exclusive. Our findings demonstrate that this assumption is needed. A learner with only a numerical regularization bias, and no substantive expectations about the nature of the grammatical system it is acquiring, does not spontaneously identify that these morphemes should be analyzed with different grammatical functions given the data that it observes.

6 General Discussion

We offer a general computational account for how children manage to draw systematic generalizations from messy data in the process of acquiring their first language. We

introduce a mechanism for noise-tolerant learning of restrictive grammatical hypotheses. The type of learner that we consider assumes that its data are generated by a complex system: the particular grammatical processes that the learner is currently trying to acquire, and other independent processes that conspire to introduce variability into the data. We model the inference process as a special case of grammar learning, in which the learner evaluates a choice among different *Noisy CFGs*: composite grammars in which a restricted set of "core" rules operates alongside a less restricted set of "noise" rules. By partitioning the data into portions that provide signal for the core component and portions generated by noise, the learner identifies the grammar whose restrictive core rules provide the best explanation for the skews in its data. It does so without knowing ahead of time the rate or properties of noise that it will encounter.

Our approach provides an alternative to a prominent proposal that learning in early development is driven by a domain-general bias to regularize variable data (Austin et al., 2022; Hudson Kam & Newport, 2009, 2005; Newport, 1999; Reali & Griffiths, 2009; Singleton & Newport, 2004; Smith & Wonnacott, 2010). We compare our learner to a common implementation of this general regularization bias approach, and show that both are able to account for results from a representative artificial language learning experiment (Austin et al., 2022; Reali & Griffiths, 2009; Perfors, 2012). However, we find that our learner out-performs the regularization bias approach in two naturalistic case studies in early syntax acquisition: learning the rules governing basic clause structure and those governing case morphology. We show that our learner succeeds because its architecture allows a natural way to express linguistically-motivated expectations about the character of those grammatical rules. A learner with a numerical regularization bias operating over a fully-flexible hypothesis space, and no expectation that it is acquiring a particular type of restrictive grammatical system, is not able to identify the correct canonical word order or case-marking system within its messy data.

These findings invite the possibility that other observed cases of regularization in

grammar learning may be accounted for without adopting a fully-flexible hypothesis space—and that some cases may be better explained as noise-tolerant selection among discrete, restrictive grammatical hypotheses. We argue that this approach provides a straightforward way to encode knowledge about the specific types of regularities that a learner is expecting to encounter, which is important for learning to succeed in the case studies that we consider. In particular, our case studies showed that it was not sufficient to endow learners with a numerical bias towards probabilities close to zero or one, which is agnostic about the content of what is being learned and therefore does not favor any particular extreme within the gradient hypothesis space. However, another logical possibility is that a learner's domain-specific knowledge (for example, a preference for clauses to have subjects) could be expressed as a numerical bias that is asymmetrical, preferring certain extremes over others as a function of the particular learning problem being addressed.[13] In other words, the current paper contrasts an approach to learning with a discrete hypothesis space and domain-specific knowledge against an approach with a gradient hypothesis space and few domain-specific expectations, but we note that these are not the only two possibilities in the space of theoretical options. We leave exploration of these issues to future work.

In contrasting how regularization behavior arises from these two approaches, we observe a difference in how learning interacts with the size of the learner's data. On the numerical regularization account, regularization is a property of learning with small amounts of data: as a learner's intake grows, its prior bias plays less of a role. Many previous accounts have therefore proposed that regularization arises in part from cognitive constraints that significantly limit the amount of data that children are able to take in for learning (Keogh et al., 2024; Perfors, 2012; Newport, 1999, 1990). On our restrictive hypotheses account, we see that limitations to the learner's intake are not necessary in the same way. Our learner succeeds at identifying the correct canonical grammar with strikingly small

[13] In particular, instead of adopting a symmetrical Beta prior with a single α shape parameter, it is possible to encode asymmetrical preferences through a more general parameterization in which the Beta distribution has two different shape parameters (α and β).

amounts of data, but as more data are observed, its regularization tendencies do not weaken; instead, they become stronger as it becomes more confident in its guesses of how to partition data as coming from core vs. noise rules. The signal that our learner identifies for the core rules may indeed come from a small portion of its data, but observing more data allows it to determine with greater certainty which portion it should attend to. These differences invite questions about whether and how the strength of children's grammatical generalizations vary relative to the amount of data that they encode.

While here we model the learning of basic syntax as a choice among Noisy CFGs, this same approach can be applied to learners of other sorts of "Noisy" grammars, including those that are non-context-free— specifically, learners of any formalism that generates complex structures as a function of local choices about smaller subparts, where the likelihoods of the data are products of multinomials. With any such formalism, the important properties of a Noisy Grammar learner's hypothesis space are (i) that flexible noise rules take the same form as the restrictive core rules (for example, for our Noisy CFG learner, they are also CFG productions), and (ii) that these noise rules are a superset of the union of all of the core rules in the learner's hypothesis space. With these two properties, the inferential logic that we illustrate here will generalize beyond CFG learning, and might be extended to many other problems in grammar acquisition: e.g., learning phonological constraints that can be expressed in Noisy Finite-State systems, or learning syntactic dependencies that can be expressed in Noisy Multiple Context-Free Grammars.

More broadly, the two approaches to regularization that we consider in this paper relate to two general views of learning, in language and in other domains. On one view, learning involves summarizing the distributions in the learner's data; this summary may be more or less veridical, if learners are generally biased towards certain distributions *a priori* (e.g., Elman et al., 1996; Aslin & Newport, 2012; Thelen & Smith, 2007). On another view, learning involves evaluating hypotheses about the generative systems that give rise to the distributions in the learner's data in a specific domain. Learning is not an attempt to

summarize those distributions, but rather to use them as indirect evidence to infer the underlying system that generated them (e.g., Chomsky, 1965, 1975; Lidz & Gagliardi, 2015; Lightfoot, 1991; Yang, 2002; Gallistel, 1990). We provide a formally explicit architecture for performing this inference in cases where the data are messy, because the generative system contains multiple components that interact in opaque ways. Solving this learning problem does not require assuming that learners bring with them a hypothesis space that is capable of encoding the full distribution of the data, in all of its messiness. Instead, we argue that this problem can be solved if learners have specific assumptions about how their restrictive hypotheses will be noisily reflected in the data that they observe. We show that this approach can more readily account for the generalizations that infants draw in two areas of syntax acquisition. This provides support for theories in which learning, at least in certain domains, is underwritten by restrictive generative systems in a learner's hypothesis space, combined with a mechanism for filtering signal from noise.

7 References

Aslin, R. N., & Newport, E. L. (2012). Statistical learning: From acquiring specific items to forming general rules. *Current directions in psychological science*, *21*(3), 170–176.

Austin, A. C., Schuler, K. D., Furlong, S., & Newport, E. L. (2022). Learning a Language from Inconsistent Input: Regularization in Child and Adult Learners. *Language Learning and Development*, *18*(3), 249–277.

Babineau, M., Shi, R., & Christophe, A. (2020). 14-month-olds exploit verbs' syntactic contexts to build expectations about novel words. *Infancy*, *25*(5), 719–733.

Beech, C., & Swingley, D. (2023). Consequences of phonological variation for algorithmic word segmentation. *Cognition*, *235*, 105401.

Behrend, E. R., & Bitterman, M. E. (1961). Probability-matching in the fish. *The American Journal of Psychology*, *74*(4), 542–551.

Bever, T. G. (1982). Regression in the service of development. In *Regressions in Mental Development* (pp. 153–188). Routledge.

Bickerton, D. (1981). *Roots of language*. Ann Arbor, MI: Karoma.

Bickerton, D. (1984). The language bioprogram hypothesis. *Behavioral and brain sciences*, *7*(2), 173–188.

Bion, R. A., Miyazawa, K., Kikuchi, H., & Mazuka, R. (2013). Learning phonemic vowel length from naturalistic recordings of Japanese infant-directed speech. *PloS one*, *8*(2), e51594.

Booth, T. L., & Thompson, R. A. (1973). Applying probability measures to abstract languages. *IEEE Transactions on Computers*, *C-22*, 442–450.

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

Brusini, P., Dehaene-Lambertz, G., Van Heugten, M., De Carvalho, A., Goffinet, F., Fiévet, A.-C., & Christophe, A. (2017). Ambiguous function words do not prevent 18-month-olds from building accurate syntactic category expectations: An ERP study.

Neuropsychologia, *98*, 4–12.

Bullock, D. H., & Bitterman, M. E. (1962). Probability-matching in the pigeon. *The American Journal of Psychology*, *75*(4), 634–639.

Cameron-Faulkner, T., Lieven, E., & Tomasello, M. (2003). A construction based analysis of child directed speech. *Cognitive science*, *27*(6), 843–873.

Chomsky, N. (1965). *Aspects of the Theory of Syntax*. Cambridge, MA: MIT Press.

Chomsky, N. (1975). *Reflections on language*. New York, NY: Pantheon.

Christophe, A., Millotte, S., Bernal, S., & Lidz, J. (2008). Bootstrapping Lexical and Syntactic Acquisition. *Language and Speech*, *51*(1-2), 61–75.

Craig, G. J., & Myers, J. L. (1963). A developmental study of sequential two-choice decision making. *Child Development*, 483–493.

Cristia, A. (2018). Can infants learn phonology in the lab? A meta-analytic answer. *Cognition*, *170*, 312–327.

Cristia, A., Dupoux, E., Ratner, N. B., & Soderstrom, M. (2019). Segmentability differences between child-directed and adult-directed speech: A systematic test with an ecologically valid corpus. *Open Mind*, *3*, 13–22.

Culbertson, J., & Kirby, S. (2016). Simplicity and Specificity in Language: Domain-General Biases Have Domain-Specific Effects. *Frontiers in Psychology*, *6*.

Culbertson, J., Smolensky, P., & Wilson, C. (2013). Cognitive biases, linguistic universals, and constraint-based grammar learning. *Topics in cognitive science*, *5*(3), 392–424.

Dautriche, I., Cristia, A., Brusini, P., Yuan, S., Fisher, C., & Christophe, A. (2014). Toddlers Default to Canonical Surface-to-Meaning Mapping When Learning Verbs. *Child Development*, *85*(3), 1168–1180.

Demuth, K., & Tremblay, A. (2008). Prosodically-conditioned variability in children's production of French determiners. *Journal of child language*, *35*(1), 99–127.

Derks, P. L., & Paclisanu, M. I. (1967). Simple strategies in binary prediction by children and adults. *Journal of Experimental Psychology*, *73*(2), 278.

Elman, J. L., Bates, E. A., Johnson, M. H., Karmiloff-Smith, A., Plunkett, K., & Parisi, D. (1996). *Rethinking innateness: A connectionist perspective on development*. MIT press.

Estes, W. K. (1964). Probability learning. In *Categories of human learning* (pp. 89–128). Elsevier.

Estes, W. K. (1976). The cognitive side of probability learning. *Psychological Review*, *83*(1), 37.

Ferdinand, V., Kirby, S., & Smith, K. (2019). The cognitive roots of regularization in language. *Cognition*, *184*, 53–68.

Fisher, C., Jin, K.-S., & Scott, R. M. (2019). The developmental origins of syntactic bootstrapping. *Topics in Cognitive Science*, *12*(1), 48–77.

Fodor, J. D. (1998). Parsing to learn. *Journal of Psycholinguistic research*, *27*(3), 339–374.

Gallistel, C. R. (1990). *The organization of learning*. The MIT Press.

Gardner, R. A. (1957). Probability-learning with two and three choices. *The American Journal of Psychology*, *70*(2), 174–185.

Geman, S., & Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on pattern analysis and machine intelligence*(6), 721–741.

Gertner, Y., Fisher, C., & Eisengart, J. (2006). Learning words and rules: abstract knowledge of word order in early sentence comprehension. *Psychological Science*, *17*(8), 684–691.

Geyer, C. J. (1991). Markov chain Monte Carlo maximum likelihood. *Computing Science and Statistics: Proceedings of the 23rd Symposium on the Interface*, 156–163. (New York American Statistical Association)

Gleitman, L. R. (1990). The structural sources of verb meanings. *Language acquisition*, *1*(1), 3–55.

Göksun, T., Küntay, A. C., & Naigles, L. R. (2008). Turkish children use morphosyntactic bootstrapping in interpreting verb meaning. *Journal of child language*, *35*(2), 291–323.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, *57*(1), 97–109.

Hicks, J., Maye, J., & Lidz, J. (2007). The role of function words in infants' syntactic categorization of novel words. Anaheim, CA.

Hirsh-Pasek, K., & Golinkoff, R. M. (1996). The intermodal preferential looking paradigm: A window onto emerging language comprehension. In D. McDaniel, C. McKee, & H. S. Cairns (Eds.), *Methods for assessing children's syntax* (pp. 105–124). Cambridge, MA: The MIT Press.

Hitczenko, K., & Feldman, N. H. (2022). Naturalistic speech supports distributional learning across contexts. *Proceedings of the National Academy of Sciences*, *119*(38), e2123230119.

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

Hudson Kam, C. L., & Newport, E. L. (2009). Getting it right by getting it wrong: When learners change languages. *Cognitive psychology*, *59*(1), 30–66.

Höhle, B., Weissenborn, J., Kiefer, D., Schulz, A., & Schmitz, M. (2004). Functional elements in infants' speech processing: The role of determiners in the syntactic categorization of lexical elements. *Infancy*, *5*(3), 341–353.

Johnson, J. S., Shenkman, K. D., Newport, E. L., & Medin, D. L. (1996). Indeterminacy in the grammar of adult language learners. *Journal of memory and language*, *35*(3), 335–352.

Johnson, M., Griffiths, T., & Goldwater, S. (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* (pp. 139–146). Association for Computational Linguistics.

Jusczyk, P. W., & Aslin, R. N. (1995). Infants' detection of the sound patterns of words in

fluent speech. *Cognitive psychology*, *29*(1), 1–23.

Keogh, A., Kirby, S., & Culbertson, J. (2024). Predictability and Variation in Language Are Differentially Affected by Learning and Production. *Cognitive Science*, *48*(4), e13435.

Kim, Y. J., & Sundara, M. (2021, July). 6–month–olds are sensitive to English morphology. *Developmental Science*, *24*(4), e13089. doi: 10.1111/desc.13089

Lidz, J., & Gagliardi, A. (2015). How nature meets nurture: Universal grammar and statistical learning. *Annu. Rev. Linguist.*, *1*(1), 333–353.

Lidz, J., White, A. S., & Baier, R. (2017). The role of incremental parsing in syntactically conditioned word learning. *Cognitive Psychology*, *97*, 62–78.

Lightfoot, D. (1991). *How to set parameters: Arguments from language change.* Cambridge, MA: MIT Press.

Maitra, S., & Perkins, L. (2023). Filtering Input for Learning Constrained Grammatical Variability: The Case of Spanish Word Order. *Proceedings of the Society for Computation in Linguistics*, *6*(1), 108–120.

Matsuo, A., Kita, S., Shinya, Y., Wood, G. C., & Naigles, L. (2012). Japanese two-year-olds use morphosyntax to learn novel verb meanings. *Journal of child language*, *39*(3), 637–663.

Mattys, S. L., & Jusczyk, P. W. (2001). Phonotactic cues for segmentation of fluent speech by infants. *Cognition*, *78*(2), 91–121.

Mattys, S. L., Jusczyk, P. W., Luce, P. A., & Morgan, J. L. (1999). Phonotactic and prosodic effects on word segmentation in infants. *Cognitive psychology*, *38*(4), 465–494.

Maye, J., Werker, J. F., & Gerken, L. (2002). Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition*, *82*(3), 101–111.

Mintz, T. H. (2013). The segmentation of sub-lexical morphemes in English-learning 15-month-olds. *Frontiers in Psychology*, *4*(24).

Morgan, J. L., & Demuth, K. (Eds.). (1996). *Signal to syntax: Bootstrapping from speech to grammar in early acquisition.* Lawrence Erlbaum.

Myers, J. L. (1976). Probability learning and sequence learning. *Handbook of Learning and Cognitive Processes, ed. WK Estes*, 171–205.

Newport, E. L. (1990). Maturational constraints on language learning. *Cognitive science*, *14*(1), 11–28.

Newport, E. L. (1999). Reduced input in the acquisition of signed languages: Contributions to the study of creolization. In M. Degraff (Ed.), *Creolization, diachrony, and language acquisition.* Cambridge, MA: MIT Press.

Oshima-Takane, Y., MacWhinney, B., Sirai, H., Miyata, S., & Naka, N. (1995). *CHILDES manual for Japanese* (Tech. Rep.). Montreal: McGill University.

Perfors, A. (2012). When do memory limitations lead to regularization? An experimental and computational investigation. *Journal of Memory and Language*, *67*(4), 486–506.

Perkins, L., Feldman, N. H., & Lidz, J. (2022). The power of ignoring: filtering input for argument structure acquisition. *Cognitive Science*, *46*(1).

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

Perkins, L., & Lidz, J. (2021). 18-month-old infants represent non-local syntactic dependencies. *Proceedings of the National Academy of Sciences*, *118*(41), e2026469118.

Pinker, S. (1984). *Language Learnability and Language Development.* Cambridge, MA: Harvard University Press.

Reali, F., & Griffiths, T. L. (2009). The evolution of frequency distributions: Relating regularization to inductive biases through iterated learning. *Cognition*, *111*(3), 317–328.

Sambridge, M. (2014). A parallel tempering algorithm for probabilistic sampling and multimodal optimization. *Geophysical Journal International*, *196*(1), 357–374.

Schneider, J., Perkins, L., & Feldman, N. H. (2020). A noisy channel model for systematizing unpredictable input variation. In *Proceedings of the 44th Annual Boston University conference on language development* (pp. 533–547).

Schulz, L. E., & Sommerville, J. (2006). God Does Not Play Dice: Causal Determinism and Preschoolers' Causal Inferences. *Child Development*, *77*(2), 427–442.

Senghas, A., & Coppola, M. (2001). Children creating language: How Nicaraguan Sign Language acquired a spatial grammar. *Psychological science*, *12*(4), 323–328.

Shi, R., & Melançon, A. (2010). Syntactic Categorization in French-Learning Infants. *Infancy*, *15*(5), 517–533.

Singleton, J. L., & Newport, E. L. (2004). When learners surpass their models: The acquisition of American Sign Language from inconsistent input. *Cognitive Psychology*, *49*(4), 370–407.

Smith, K., Perfors, A., Fehér, O., Samara, A., Swoboda, K., & Wonnacott, E. (2017). Language learning, language use and the evolution of linguistic variation. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *372*(1711), 20160051.

Smith, K., & Wonnacott, E. (2010). Eliminating unpredictable variation through iterated learning. *Cognition*, *116*(3), 444–449.

Stevenson, H. W., & Weir, M. W. (1959). Variables affecting children's performance in a probability learning task. *Journal of Experimental Psychology*, *57*(6), 403.

Stromswold, K. (1995). The acquisition of subject and object wh-questions. *Language Acquisition*, *4*(1-2), 5–48.

Suzuki, T. (1999). *Two aspects of Japanese case in acquisition* (Doctoral dissertation). University of Hawai'i at Manoa.

Suzuki, T., & Kobayashi, T. (2017). Syntactic Cues for Inferences about Causality in Language Acquisition: Evidence from an Argument-Drop Language. *Language Learning and Development*, *13*(1), 24–37.

Swingley, D. (2019). Learning Phonology from Surface Distributions, Considering Dutch and English Vowel Duration. *Language Learning and Development*, *15*(3), 199–216.

Tenenbaum, J. B., & Griffiths, T. L. (2001). Generalization, similarity, and Bayesian inference. *Behavioral and brain sciences*, *24*(04), 629–640.

Thelen, E., & Smith, L. B. (2007). Dynamic Systems Theories. In W. Damon & R. M. Lerner (Eds.), *Handbook of Child Psychology* (1st ed.). Wiley.

Valian, V. (1990). Logical and psychological constraints on the acquisition of syntax. In L. Frazier & J. G. De Villiers (Eds.), *Language Processing and Language Acquisition.* Dordrecht: Kluwer.

Weir, M. W. (1964). Developmental changes in problem-solving strategies. *Psychological review*, *71*(6), 473.

Wetherell, C. S. (1980). Probabilistic Languages: A Review and Some Open Questions. *Computing Surveys*, *12*, 361–379.

Wolfram, W. (1985). Variability in tense marking: A case for the obvious. *Language Learning*, *35*(2), 229–253.

Yang, C. (2002). *Knowledge and learning in natural language.* Oxford: Oxford University Press.

Appendix

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 trees $\vec{t}$ for those strings:

$$P(G|\vec{t},\vec{w}) = \frac{P(\vec{t},\vec{w}|G)P(G)}{\sum_{G'} P(\vec{t},\vec{w}|G')P(G')} \tag{25}$$

Bayes' Rule tells us that the posterior probability of any grammar is proportional to the product of the likelihood (the probability of $\vec{t}$ and $\vec{w}$ under that grammar) and the prior probability of that grammar. We assume that all grammars have equal prior probability.

Because we are only considering trees that could have yielded the strings in the data, the joint likelihood of the trees and strings, $P(\vec{t},\vec{w}|G)$, is equivalent to the likelihood of the trees alone, $P(\vec{t}|G)$. Calculating this likelihood requires summing over the unknown ways that each portion of these trees might be analyzed as stemming from either a core (Φ) or noise (Ψ) rewrite, i.e., the choices of "articulated trees" for a given tree. 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. This follows similar logic to the illustration in the simple example in Section 3; here we show how this applies to the more general case.

We divide the n^A total observations of a particular nonterminal A into $n^A_1 \ldots n^A_m$ 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 trees, $P(\vec{t}|G)$, is the product over all nonterminals A of $P(n^A_1 \ldots n^A_m \mid G)$. We divide each of the observed rewrites of a nonterminal into some number of core rewrites (Φ) and some number of noise rewrites (Ψ).[14]

[14] In Section 3, these choices occurred, perhaps more intuitively, in the reverse order: first we divided the total observations of a given nonterminal into Φ vs. Ψ observations, and then we partitioned the observations coming from each component (Φ vs. Ψ) among the different ways that the nonterminal could be rewritten. These two orders produce the same final result as long as the ranges for the summations are calculated

The n_1^A occurrences of the first type of rewrite for A are divided into $n_1^{A^\phi}$ core occurrences and $n_1^{A^\psi}$ noise occurrences. More generally, the n_m^A occurrences of the m^{th} rewrite type are divided into $n_m^{A^\phi}$ core occurrences and $n_m^{A^\psi}$ noise occurrences. We can calculate the likelihood by marginalizing over $n_1^{A^\phi} \dots n_m^{A^\psi}$:

$$(26)\quad P(\vec{t}|G) = \prod_A P(n_1^A \dots n_m^A \mid G) = \prod_A \Bigg[\sum_{n_1^{A^\phi}=0}^{n_1^A} \cdots \sum_{n_m^{A^\phi}=0}^{n_m^A} \Big[P(n_1^{A^\phi} \dots n_m^{A^\phi} | n^{A^\phi}, G) \times P(n_1^{A^\psi} \dots n_m^{A^\psi} | n^{A^\psi}, G) \times P(n^{A^\phi} | n^A, G) \Big] \Bigg]$$

The first term in the summation is the probability of observing $n_1^{A^\phi} \dots n_m^{A^\phi}$ 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

$$(27)\quad \frac{B(\vec{\alpha_\phi}^{A_G} + (n_1^{A^\phi} \dots n_m^{A^\phi}))}{B(\vec{\alpha_\phi}^{A_G})}$$

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. As we noted in Section 3, when the prior over rule weights is uniform (all components of $\vec{\alpha_\phi}^{A_G}$ are equal to 1), this is equivalent to 1 divided by the number of ways to partition n^{A^ϕ} core occurrences of A into m^{A^ϕ} possible rewrite types:

$$(28)\quad \frac{1}{\binom{n^{A^\phi}+m^{A^\phi}-1}{m^{A^\phi}-1}} = \frac{n^{A^\phi}!(m^{A^\phi}-1)!}{(n^{A^\phi}+m^{A^\phi}-1)!}$$

The second term in the sum in (26) is analogous: this is the probability, given n^{A^ψ}

appropriately, but the order that we demonstrate here scales better to larger grammars.

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 (27).

This allows us to calculate the likelihood $P(\vec{t} \mid G)$ for each G in our hypothesis space, and (since we assume a flat prior over 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 trees $\vec{t}$ for the observed strings, given G. Following M. Johnson et al. (2007), we consider the probability of a tree structure t_i for corresponding string w_i, given G and the current hypotheses about trees $\vec{t}_{-i}$ for all the other strings. We can define a function f that is proportional to the posterior distribution over t_i, $f(t_i) \propto P(t_i|w_i, \vec{t}_{-i}, G)$, as

$$f(t_i) = P(w_i|t_i)P(t_i|\vec{t}_{-i}, G) \tag{29}$$

The probability of a string being the yield of a given tree, $P(w_i|t_i)$, is always 1 or 0. The probability of a tree given all other trees and G, $P(t_i|\vec{t}_{-i}, G)$, is

$$P(t_i|\vec{t}_{-i}, G) = \frac{P(\vec{t}|G)}{P(\vec{t}_{-i}|G)} \tag{30}$$

Both $P(\vec{t}|G)$ and $P(\vec{t}_{-i}|G)$ can be calculated according to to (26). In practice, $P(\vec{t}_{-i}|G)$ does not need to be calculated because it will cancel out in the acceptance function (31), below.

We can use this function f to sample $\vec{t}$ given G and $\vec{w}$ as follows. Within each iteration of the Gibbs sampler, we re-sample $\vec{t}$ using a procedure modified from M. Johnson et al. (2007). First, we choose a string w_i and its current corresponding t_i at random. Second, we take the other trees $\vec{t}_{-i}$, to be the output of a simple PCFG which generates these structures directly, rather than generating them via articulated trees that distinguish between noise vs.

non-noise rewrites. We estimate the probabilities of each rewrite for a given nonterminal in this simple PCFG, $\vec{\theta^s}$, by sampling from a multivariate Gaussian whose mean is set to the relative frequencies of each observed rewrite, using add-one smoothing to account for accidental gaps. Third, we generate a new proposed tree $t_i{}'$ for w_i by sampling from this simple grammar's distribution using $\vec{\theta^s}$. Finally, we decide to accept this proposal with probability

$$A(t_i') = min\left(1, \frac{f(t_i{}')P(t_i|w_i, \vec{\theta^s})}{f(t_i)P(t_i{}'|w_i, \vec{\theta^s})}\right) \tag{31}$$

We ran multiple chains from different starting places to test convergence. For the simulations reported in Sections 4 and 5, we ran chains of 50,000 iterations of Gibbs sampling each, and analyzed every 10th iteration from the last quarter of each chain. We report averages across 10 chains as estimates of the posterior over G and $\vec{t}$.

To simulate the "fully-flexible" learners in these sections, we estimate the posterior distribution over $\vec{t}$ by using a component-wise Hastings sampler analogous to that for estimating $P(\vec{t}|G, \vec{w})$ in our original model. To improve convergence within the learner's multimodal hypothesis space, we use parallel tempering (Geyer, 1991). We run 10 chains in parallel, of which 9 sample from a "tempered" version of the target posterior: the posterior is raised to a power between 0 and 1, flattening the distribution and allowing the chain to mix quickly. At each iteration, a state swap between the sampled trees $\vec{t}$ of two random chains is proposed, and this proposal is accepted using a Metropolis-Hastings update, which preserves the joint target posterior distributions for each chain; see Sambridge (2014) for detail. Only the chain that samples from the true target posterior $P(\vec{t}|G, \vec{w})$ is analyzed at the end of the run. We ran 10 such target chains of 50,000 Hastings iterations each, and analyzed every 10th iteration from the last quarter of each chain.