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" their input, 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 distortions into the data. Our *Noisy Grammar 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 only our approach succeeds 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 ambiguous and messy. 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). What kind of mechanisms allow for learning to abstract away from such messiness in the learner's representation of the data?

In this paper, we examine a specific case of messiness that arises as learning proceeds incrementally over the course of development: the case of *opacity*, where evidence for one grammatical property can be hidden by other interacting grammatical properties that have not yet been acquired. One example that illustrates this issue is a problem in the domain of syntax learning, of identifying the rules governing basic word order. The canonical order of subjects, verbs, and objects in a language varies cross-linguistically and therefore must be acquired, but other properties of the grammar conspire to hide this canonical word order in certain sentences. For instance, children learning English and French must identify a grammatical rule system that canonically produces subject-verb-object (SVO) orders, whereas children learning Japanese must identify a rule system that canonically produces subject-object-verb (SOV). This task can be abstractly characterized as selecting among rules for basic clauses that place the subject noun-phrase before or after the verb-phrase, and rules that place the object noun-phrase before or after the verb. Empirical evidence suggests that children learn their language's basic word order very early in development, in infancy (Hirsh-Pasek & Golinkoff, 1996; Gertner, Fisher, & Eisengart, 2006; Lidz, White, & Baier, 2017; Perkins & Lidz, 2020; Franck, Millotte, Posada, & Rizzi, 2013; Zhu, Franck, Rizzi, &

Gavarró, 2022; Gavarró, Leela, Rizzi, & Franck, 2015). What is intriguing is that they do so at ages before they can identify the sentences where interacting processes have caused subjects and objects to appear in non-canonical positions (Perkins & Lidz, 2020, 2021; Gagliardi, Mease, & Lidz, 2016). These kinds of processes include those that produce *wh*-questions and relative clauses, as in (1). In these sentences, a fronted phrase acts as the verb's direct object in a non-canonical position, rather than post-verbally; they therefore depart from the basic SVO word order of English reflected in (2).

(1) a. Which dog is she chasing?

b. That's the dog we like.

(2) a. She's chasing a dog.

b. We like that dog.

If infants cannot yet distinguish between sentences exhibiting basic word order and those that depart from this basic word order, because they do not yet know the form that *wh*-questions and relative clauses take in the language, then the representations that they will be able to construct of these sentences do not coherently point towards the correct canonical word order. The sentences in (2), with one NP preverbally and one NP postverbally, could be taken as evidence for rules that canonically produce verb-medial orders; but the sentences in (1), with multiple preverbal NPs, could equally be taken as evidence for rules that produce SOV canonically. The existence of *wh*-questions and relative clauses therefore distorts the body of evidence that an infant might use for the basic positions of clause arguments, in effect rendering infants' data messy and misleading at the stage of development when word order learning takes place. Infants nonetheless manage to abstract away from this messiness to draw accurate generalizations about basic word order.

This example illustrates how opacity can arise from a grammatical system with multiple interacting components, when a learner is attempting to acquire these components incrementally. This type of learning problem has received substantial previous attention in

the syntax learning literature (e.g., Fodor, 1998; Frank & Kapur, 1996; Gibson & Wexler, 1994; Howitt, Dey, & Sakas, 2021; Hyams, 1986; Lightfoot, 1991; Manzini & Wexler, 1987; Niyogi & Berwick, 1996; Sakas & Fodor, 2012, 2001). But the solutions proposed in this literature have tended to be problem-specific, designed to handle particular opaque interaction puzzles individually. For instance, much of the work in this tradition has treated word order learning as a problem that requires its own set of learning heuristics, with other heuristics specified for different phenomena to be acquired (Gibson & Wexler, 1994; Howitt et al., 2021; Lightfoot, 1991; Sakas & Fodor, 2012, 2001). Here, we show that learners' solutions to the problem of word order acquisition, along with many other similar problems, can be characterized under a general computational framework for learning in the face of opacity.

We propose that in certain circumstances, learners face a choice among discrete hypotheses for the systems that generated their data, each of which is restrictive or deterministic in a way that is incompatible with the full messiness of the available 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. 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, as in (2), and noise introduced by non-canonical sentence types, as in (1). This same idea generalizes beyond word order acquisition to many other learning problems. For a child learning the morphosyntax of a case-marking system, the data reflect a combination of signal for restrictive rules governing which affixes mark subjects vs. objects, and noise introduced by non-canonical argument orders, argument-drop, and case-marker omission. Successful learning emerges when children are able to identify signal for a restrictive hypothesis within their opaque data, and adopt this hypothesis as the best explanation despite its surface mismatch with the data.

We introduce a general computational framework for modeling a learner's inference when the learning problem is cast in this form. Using rational probabilistic inference, the

learner aims to separate evidence for a restrictive core grammar from the distorting effects of non-canonical noise processes. Importantly, it does so without knowing ahead of time how much noise is present in its data, or what the properties of that noise are. This mechanism therefore provides a way for learners to select among a space of restrictive grammatical hypotheses from messy data, without needing *a priori* knowledge of the kinds of other phenomena that are responsible for that opacity.

In the case studies that we discuss below, we will show that the success of this learning mechanism depends on domain-specific expectations about the types of grammatical mechanisms that may have generated the learner's input. Our proposal therefore falls under a broader class of approaches which view language learning as an attempt to use the distributions in the observed data as evidence for the parameters of this generative system, under the assumption that these parameters may interact in opaque ways (Chomsky, 1965, 1975; Fodor, 1998; Lidz & Gagliardi, 2015; Lightfoot, 1991; Valian, 1990; C. Yang, 2002). We show that general probabilistic reasoning can be combined with these approaches to resolve the tension between a restrictive hypothesis space and the need for noise-tolerant learning.

This proposal stands in contrast to certain alternative approaches to probabilistic learning, which argue that general noise-tolerant probabilistic mechanisms can substantially reduce or potentially eliminate the need for restrictive domain-specific grammatical expectations in a learner's hypothesis space (e.g., Elman et al., 1996; Lewis & Elman, 2001; Perfors, Tenenbaum, & Regier, 2011; Perfors, Tenenbaum, & Wonnacott, 2010; Reali & Griffiths, 2009; Reali & Christiansen, 2005; Sagae, 2021; K. Smith & Wonnacott, 2010; Y. Yang & Piantadosi, 2022). In what follows, we compare our *restrictive hypotheses* approach to a prominent alternative proposal in this literature, which posits that a domain-general bias to "regularize" messy data can explain many instances where learners draw generalizations that are not transparently reflected in their input. On this view, learners do not evaluate among discrete, restrictive hypotheses, but are instead equipped to consider a flexible range of hypotheses, including those that closely match the messy

distributions in their input. Divergences between the generalizations that they end up drawing and the surface distributions that they observe come from a general preference for consistency. Some motivation for this hypothesis comes from children's behavior in artificial language learning experiments. For instance, when exposed to an artificial language in which determiners occur inconsistently with nouns, children show tendencies to produce particular determiners all of the time or not at all (Austin, Schuler, Furlong, & Newport, 2022; Hudson Kam & Newport, 2005, 2009). On this account, children may be equipped to consider that the language allows determiners with any probability, but have a domain-general bias to prefer probabilities closer to zero or one. An influential literature proposes that this bias 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 driver of language change (Austin et al., 2022; Culbertson & Kirby, 2016; Ferdinand, Kirby, & Smith, 2019; Hudson Kam & Newport, 2009, 2005; Perfors, 2012; Reali & Griffiths, 2009; Singleton & Newport, 2004; K. Smith & Wonnacott, 2010).

There are two important assumptions implicit in this alternative 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, such as constraints on information processing, cognitive control, or working memory (Austin et al., 2022; Culbertson & Kirby, 2016; Ferdinand et al., 2019; Keogh, Kirby, & Culbertson, 2024; Hudson Kam & Newport, 2005, 2009; Newport, 1990). The second assumption is that across all of these various learning domains, learners have a hypothesis space that can accommodate the full distribution of the data, in all of its messiness. This might mean considering the possibility that a language allows multiple determiners in free variation, or that it allows clause arguments to freely occur in multiple basic word orders. Noise-tolerance only emerges through a bias operating within this fully-flexible hypothesis space, pushing learners away from the flexible middle and towards more skewed or systematic extremes. 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).

In previous work, the general regularization bias idea has been given an explicit mathematical formulation (Culbertson, Smolensky, & Wilson, 2013; Perfors, 2012; Reali & Griffiths, 2009; K. Smith et al., 2017), which can be incorporated into the learning of complex, layered grammatical systems (M. Johnson, Griffiths, & Goldwater, 2007). One might imagine that such an approach would allow children to infer grammatical structure from messy data: children would be equipped to consider that their data might come from any combination of grammatical rules applying with any probability, but regularization leads them to posit rules that apply with a high degree of consistency, rather than merely recapitulating the messiness in their data. For this mechanism to broadly account for the generalizations that children draw over the course of natural language acquisition, it would need to be able to overcome messiness of all sorts: not only the kind of inconsistent variability that might be encountered in the context of an artificial language learning task, but also the kind of messiness arising from opaque grammatical interactions. The case studies that we examine here show that this cannot be the full solution.

While the regularization bias proposal has been previously investigated computationally, no general computational architecture exists for modeling the restrictive hypotheses proposal. Here, we provide this architecture. We present a general computational implementation for incrementally acquiring a grammar in the face of opacity, by explicitly modeling the interaction between a set of candidate restrictive hypotheses and a component that adds noisy distortions to the data. In what follows, we compare our learning architecture to the common implementation of the general regularization bias approach: an implementation of a problem-agnostic, domain-general, widely deployed bias to prefer

extremes in general, that does not favor particular extremes for domain-specific reasons. We show that (i) both approaches can account for children's behavior in a representative artificial language experiment, but (ii) only our approach succeeds in modeling two more complex case studies of opaque learning problems in natural language syntax acquisition: acquiring word order and case-marking from data that appear messy from the perspective of young language learners. We argue that a key to our model's success in these learning problems, and the reason that it performs better than the domain-general 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 Regularization and previous computational approaches

A large body of empirical work has examined cases where learners across various linguistic and non-linguistic settings show tendencies to "regularize" their data, drawing generalizations that are not transparently reflected in the messy distributions that they observe. In this section, we review these empirical findings and describe the prominent computational approach that models this regularization behavior as deriving from a domain-general bias towards consistency. This provides the primary point of comparison for our alternative computational approach, which implements the idea of learning in a noise-tolerant way with restrictive hypotheses.

2.1 The phenomenon of regularization and how to understand it

Some of the earliest proposals for regularization in language acquisition come from studies of learning from non-native speakers, where children's data may contain substantial messiness introduced by the incomplete grammatical knowledge of their parents (J. S. Johnson, Shenkman, Newport, & Medin, 1996; Singleton & Newport, 2004; Newport, 1999; Wolfram, 1985). For example, Singleton and Newport (2004) studied a Deaf child

acquiring American Sign Language solely from parents who were late learners of the language. Despite receiving inconsistently accurate input, his morphological productions approached the accuracy of children learning from native signers. Singleton and Newport posit that this illustrates a general trend for young learners to draw generalizations that are more systematic or categorical than their noisy 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). Further experimental work has found a similar phenomenon in artificial-language learning tasks (e.g., Austin et al., 2022; Culbertson et al., 2013; Ferdinand et al., 2019; Hudson Kam & Newport, 2005, 2009; Perfors, 2012; Reali & Griffiths, 2009; K. Smith & Wonnacott, 2010). For instance, a classic series of studies 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 noun phrases in this language, adults tended to produce determiners at roughly the rates at which they had heard them in the training data. By contrast, most young children tended to behave more systematically, producing one determiner at a much higher rate.

The contrast in tendencies between adults and children in these experiments resembles a developmental trend in an older literature on non-linguistic "probability learning." For instance, Gardner (1957) asked adult participants to predict which of two lights would flash next. When exposed to Light A flashing on a random 70% of trials and Light B flashing on a random 30% of trials, participants' predictions closely matched these proportions: in 70% of trials they predicted Light A, and in 30% of trials 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; M. W. Weir,

1964). But although adults can behave variably in these tasks, an important finding that emerges is that children tend to regularize at young ages (Bever, 1982; Craig & Myers, 1963; Derks & Paclisanu, 1967; Stevenson & Weir, 1959; M. W. Weir, 1964).

To explain young learners' behavior, the prior literature has primarily explored explanations within the approach that we call the "general regularization bias" proposal. In 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 more likely to occur than intermediate values. Alternatively, they may have failed to veridically encode that distribution due to constraints on developing cognitive systems shared across learning domains, such as information processing, cognitive control, or working memory; these cognitive constraints may conspire to skew learning towards more categorical outcomes (Austin et al., 2022; Hudson Kam & Newport, 2005, 2009; Newport, 1990, 1999; see also Culbertson & Kirby, 2016; Ferdinand et al., 2019; Keogh et al., 2024; Perfors, 2012).

We explore an alternative interpretation, where children's regularization behavior derives from domain-specific expectations about the particular types of regularities that they are likely to encounter in a given learning context. 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 apparent inconsistencies in their data— "noise" coming from areas of the domain that haven't yet been acquired— in such a way as to identify one of the restrictive hypotheses under consideration.

It is important to note that the issue of how learners abstract away from various sorts of "noise" in their data is not identical to the issue of how they acquire both deterministic

and nondeterministic rules. Language learners eventually acquire many types of variable processes, conditioned by grammatical, contextual, or sociolinguistic factors (Labov, 1989; Miller, 2013; Miller & Schmitt, 2012; Roberts, 1997; Roberts & Labov, 1995; Shin & Miller, 2022; J. Smith, Durham, & Richards, 2013; J. Smith & Durham, 2019; Song, Shattuck-Hufnagel, & Demuth, 2015). For instance, while English and French have strict basic SVO word order, other languages allow more flexibility: children acquiring a language like Spanish must learn that subjects can appear both pre- and post-verbally in basic clauses (e.g., Torrego, 1989). Any theoretical approach to the phenomenon of regularization must account for the fact that not all variability in a learner's input is regularized away.

Some of the prior regularization literature proposes that children's regularization tendencies may apply most strongly at initial stages of development, leading them to favor one variant— perhaps the most frequent— when multiple are available. Over time, this tendency becomes weaker, and children incrementally learn the full range of variable processes that apply within their language and the factors that govern when those processes apply (e.g., Austin et al., 2022; see also Shin & Miller, 2022). Our proposal offers an alternative account, where learners even early in development may distinguish between variability to be acquired and variability that should be regularized away. While we use the term "restrictive hypotheses" to label the core grammatical processes that a learner is attempting to acquire at a particular stage of development, this is compatible with some amount of variability in the output of these processes. For instance, a learner evaluating among different strict word orders may entertain the possibility that verb phrases vary in whether they contain a direct object, as in the case studies that we consider below. These strict word-order hypotheses may also be further extended to consider more variable word order rules, including Spanish-like options that produce both SVO and VOS (Maitra & Perkins, 2023). Importantly, our proposal allows for a learner to encode a distinction between variability that lives within one of the restrictive hypotheses under consideration, and variability in the form of distortions to the patterns expected from any of those

hypotheses, the kind of variability that we are referring to as "noise."

The regularization bias proposal has been extensively investigated computationally (Reali & Griffiths, 2009; Culbertson et al., 2013; Perfors, 2012; K. Smith et al., 2017). However, despite its compatibility with prior experimental findings, no general computational architecture exists for modeling the restrictive hypotheses proposal, which as been studied only narrowly through specific case studies (Perkins, Feldman, & Lidz, 2022; Schneider, Perkins, & Feldman, 2020). The remainder of this section first describes the previous computational approach for implementing a general regularization bias, and illustrates its application to a representative artificial language learning experiment in Austin et al. (2022). Against this background, we then introduce our novel approach in Section 3.

2.2 Past computational work: Modeling a general regularization bias

To illustrate the previous general regularization bias approach, 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 whose noun phrases (NPs) each comprised a novel noun followed by a novel 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. These NPs were used to label puppets interacting in short video clips. During training, participants were taught the names of these puppets and were asked to repeat the sentences that they heard. Over the course of three days of training, participants heard sentences containing a total of 126 determiner-noun pairs, of which 84 had the primary determiner. For concreteness, we'll take 'ka' to be the primary determiner, so participants heard 84 occurrences of 'ka' and 42 of 'bo'.

After these three training days, participants were then asked to produce their own

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

descriptions of scenes with new combinations of the puppets and actions. Three production tests were conducted over the course of the experiment; children's behavior did not differ across these three tests, so for simplicity we'll consider learning prior to their first test. Given sufficient training, a participant could 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' as determiners in the NPs that they produced. 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. By contrast, 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.

Prior computational approaches have modelled scenarios like this one as the task of estimating an unknown parameter that governs the probability of some variant appearing in the learner's input. Suppose that learners assume that a given determiner position will have 'ka' with some unknown probability θ, and 'bo' with probability $1 - \theta$. The *likelihood* of observing k instances of 'ka' and b instances of 'bo' is the binomial probability $\binom{k+b}{k}\theta^k(1-\theta)^b$. 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 $k + b$ total tosses. Given a particular estimate of θ, the likelihood of the learner's training data before the first production test is

$$P(\begin{smallmatrix}84\text{ 'ka'}\\42\text{ 'bo'}\end{smallmatrix} \mid \theta) = \binom{126}{84}\theta^{84}(1-\theta)^{42} \tag{3}$$

A learner that chooses the value of theta according to the criterion of maximizing this likelihood would be led to $\theta = 0.67$, which is the proportion of 'ka' in the training data. On previous computational approaches, the fact that children don't produce 'ka' at this rate can be accounted for by assuming that learners combine the likelihood with their *prior* biases about the values that θ can take. This can be formalized following the principles of rational Bayesian inference. For data consisting of k observations of 'ka' and b observations of 'bo',

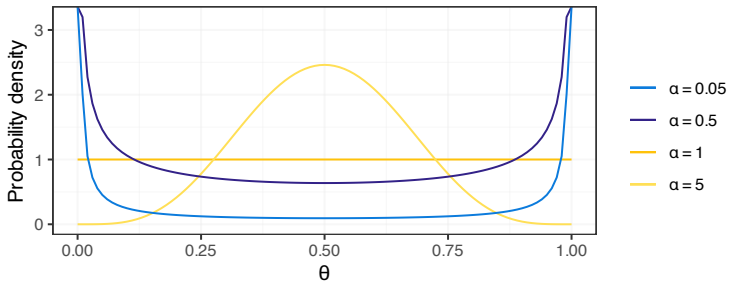


Figure 1. Illustration of symmetric Beta prior distributions, with different shape parameters

Bayes' Rule tells us that the *posterior* probability distribution over θ is proportional to the likelihood of the data under θ, combined with the learner's prior degree of belief in θ, expressed in the prior probability distribution $p(\theta)$:

$$p(\theta \mid k, b) \propto \binom{k+b}{k} \theta^k (1-\theta)^b \; p(\theta) \tag{4}$$

To model children's behavior as deriving from a domain-general regularization bias, previous work posits that young learners' prior beliefs about parameters like θ are skewed to favor extreme values. In particular, a common approach is to assume that a learner's prior distribution $p(\theta)$ takes the form of a symmetric Beta(α, α) distribution, whose α shape parameter governs the type and degree of skew (Reali & Griffiths, 2009; Culbertson et al., 2013; Perfors, 2012; K. Smith et al., 2017). When α is large, this prior distribution is skewed to favor intermediate values of θ; as α approaches zero, it encodes an *a priori* bias towards regularization, favoring values of θ close to zero or one (Figure 1). Importantly, the skew is symmetric, so a learner with this type of bias has no prior belief about which of these two extremes is more likely. In the Austin et al. experiment, a child with such a prior would combine a preference for the endpoints of the distribution with the preference to fit the 67% observed proportion of 'ka' in the 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 that maximizes the posterior probability density under (4) is 0.71.[2]

[2] More concretely, if a learner's data consist of k observations of 'ka' and b observations of 'bo,' with both $k\ , b > 0$, and the learner's prior takes the form of a symmetric Beta with parameter α, the value of θ that

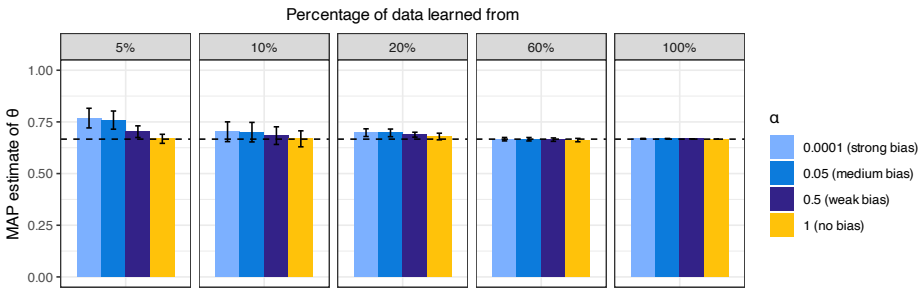


Figure 2. Maximum *a posteriori* estimates of θ across varying regularization biases, and varying amounts of data randomly sampled from 84 observations of 'ka' and 42 observations of 'bo'. Each bar represents an average across 10 randomly-generated subsets of a particular size. The dashed line indicates the proportion of 'ka' in the training data.

In this approach, the learner's prior bias will play less and less of a role as the amount of data increases, eventually resulting in posterior estimates that are very close to the maximum-likelihood estimate. Given the amount of training data that children observe in Austin et al.'s (2022) experiment, accounting for their regularization behavior with this approach requires a further assumption, which is that young learners' general cognitive limitations significantly reduce 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 numerical regularization bias exert its influence. Following Perfors (2012), if we suppose that a learner applies the inference above to a certain randomly sampled subset of the available training data in the Austin et al. experiment, then Figure 2 shows the maximum *a posteriori* values of θ that the learner would arrive at, for several possible sizes of subsets. When the learner's regularization bias is weak or a large amount of data is learned from, the posterior estimates of θ are around 0.67, matching the actual proportion of 'ka' in the data. But with an increasing bias towards regularization and less data learned from, the learner's posterior estimates of θ are more extreme. Thus, with a skewed prior and limitations on the amount of data to learn from, a

maximizes the posterior probability density under (4) is $\hat{\theta}_{\text{MAP}} = \frac{k+\alpha-1}{k+b+2\alpha-2}$. 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.

child would infer that the true posterior probability of 'ka' in the language is greater than the rate of 'ka' in the training data, and would be likely to produce 'ka' at a higher rate at test.

2.3 Summary

In summary, a prominent previous approach to modeling learners' regularization behavior applies Bayesian inference with a skewed prior distribution (Culbertson et al., 2013; Perfors, 2012; Reali & Griffiths, 2009; K. Smith et al., 2017). 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. 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 degree of skew in a learner's prior can alter the strength of the 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: in each case, regularization is achieved through a symmetrical skew in the learner's prior, 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 asymmetries in the data.

3 Noisy Grammar Learners

In this section, we introduce our novel computational architecture for modeling learning from messy data, distinct from the previous numerical regularization bias approach. On our approach, a learner selects among restrictive grammatical hypotheses, which interact in opaque ways with other unknown processes in the language to produce distortions—"noise"— in the data. We will first apply our approach to the same artificial-language learning experiment modeled in the previous section, and show how this learning task can be construed as a choice between simple sets of rules for combining determiners and nouns. This

approach thereby offers an alternative mechanism for understanding learners' regularization behavior in experimental contexts. We will then show how to scale up this approach to the learning of grammars: finite systems that generate unbounded collections of sentences. The grammar-based system that we introduce serves as the basis for the detailed case studies of natural language acquisition in Sections 4 and 5. We will see that the task of the learner is to consider three questions: (i) What do the data from core grammatical rules look like? (ii) What do the distortions introduced by noise look like? (iii) How is responsibility for the observed data distributed among the core rules and the distortions? The learner considers these three questions simultaneously in seeking the best explanation of the observed data.

The mathematical properties on which this learning framework relies are compatible with many different choices about the format of the grammatical rules in a learner's hypothesis space. To illustrate the core foundations of this framework, it is helpful to imagine grammars as mechanisms that generate trees and strings through flips of biased coins. We'll first introduce the important intuitions through a simple illustration in which a grammar is a "bag" containing coins that, when flipped, produce a determiner in an artificial language. We will apply this coin-flipping model to the Austin et al. (2022) experiment, and show that this model can be recast in more familiar linguistic terms as a choice between two very simple sets of grammatical rules. We'll then introduce the general form of this learning architecture, which we call a *Noisy Grammar Learner*, and show how this same coin-flipping idea scales up to a learner acquiring complex grammars made up of collections of interacting rules.

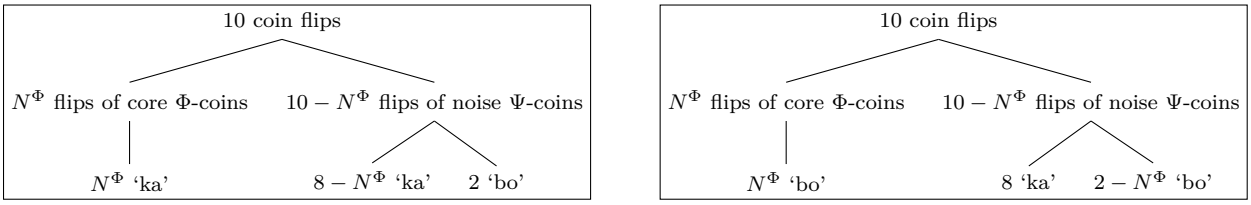
3.1 Modeling regularization as selection among noisy hypotheses

Suppose a participant is construing the task in Austin et al. (2022)'s experiment not as a task of estimating a gradient parameter θ that governs the probability of observing a particular determiner in the language, but rather as a task of choosing between two discrete "grammars," which provide different restrictive options for which determiner is canonical in the language. In one of these grammars, which we'll call G_{ka}, the canonical determiner is

'ka'; in the other, G_{bo}, the canonical determiner is 'bo.' The learner attempts to select between G_{ka} and G_{bo} based on the observed training data. To account for the fact that the data contains a mixture of both 'ka' and 'bo', we will consider that the rules for producing canonical determiners are embedded within a noisy system, in which other "noise" processes can introduce non-canonical determiners into the data.

3.1.1 Illustration with coin flips. We'll temporarily leave aside familiar linguistic notions of what grammar rules might look like, and imagine that these grammars are two bags containing coins that, when flipped, produce a determiner. G_{ka} contains "core" coins that always produce the canonical determiner 'ka'— we can think of these coins as having two sides both labeled 'ka.' G_{bo} similarly contains "core" coins where both sides are labeled 'bo'. We can think of the learner's task as trying to decide which bag's coins were responsible for a sequence of coin flips that generated some observed numbers of 'ka' and 'bo'. 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 producing 'ka', and $1 - \psi$ of producing 'bo'. To decide which bag provides a better explanation of the observed occurrences of 'ka' and 'bo', 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 — how many of the observed coin flips were "signal" bearing on the decision between G_{ka} and G_{bo}, 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 (grammar) over the other. In later sections, the grammars that a learner entertains will be comprised of sets of interacting rules, but the mathematical foundations of the learner's inference will follow the same reasoning.

We'll start by focusing attention on G_{ka}, where the core coins — we'll call these Φ-coins — always produce 'ka'. Suppose that ten times, a coin is drawn from this bag (with replacement) and flipped, producing eight 'ka' and two 'bo'. How many of these flips should we guess came from the core Φ-coins, and how many from the noise Ψ-coins? There is a wide



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

Figure 3. Partitioning 8 'ka' and 2 'bo' into core and noise

range of options, including the possibility that all ten flips came from the noise coins. But given the observed skew towards 'ka', there is a clear intuition that the core 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 core coins. Under the hypothesis that all ten flips came from noise coins ($N^\Phi = 0$), eight of the noise flips would need to produce 'ka' and two to produce 'bo' in order to generate the observed data. Contrast this with the more intuitively plausible hypothesis that there were six core flips and four noise flips ($N^\Phi = 6$). Under this hypothesis, the six core flips need to produce 'ka', which is guaranteed to happen; so, generating the observed data just amounts to having the four noise flips produce two 'ka' and two 'bo'. This is less "costly" than the first hypothesis's requirement that ten noise flips produce eight 'ka' and two 'bo'. By positing six core flips, six of the 'ka' observations that we need to generate come for free; with only noise flips, however, we get no such head start.

More precisely, given a particular probability ψ that a flip of a noise Ψ-coin produces 'ka', the likelihood of the data under the hypothesis that relies on only four noise flips ($N^\Phi = 6$) is $\binom{4}{2}\psi^2(1-\psi)^2$. Under the hypothesis that leaves all the work to ten noise flips ($N^\Phi = 0$), this likelihood is $\binom{10}{8}\psi^8(1-\psi)^2$. Because we don't know the value of ψ, we calculate the likelihood under each hypothesis by considering (marginalizing over) all possible values of that parameter. These details are given in Appendix A, but the key result can be understood intuitively as follows. If we assume that nothing is known about how frequently a noise flip will produce 'ka', then this means that we have no prior belief that any particular

value of ψ is more likely than any other value; each value of ψ has equal prior probability. With this assumption, our likelihood calculations have simple solutions. If there are four noise flips, then each of the possible numbers of 'ka' that these flips might produce (ranging from 0 to 4) is equally likely, and has probability $\frac{1}{5}$. So, under the $N^{\Phi} = 6$ hypothesis, the likelihood of observing eight occurrences of 'ka' out of ten flips, just like any other number out of ten flips, is $\frac{1}{5}$. On the hypothesis that there are ten noise flips, the likelihood of the data is smaller: there are eleven possible numbers of 'ka' that could have been observed, so the likelihood of seeing this particular number of 'ka' under the $N^{\Phi} = 0$ hypothesis is only $\frac{1}{11}$.

This is the central point to our approach: when we make no commitments about the probability that noise coins will produce 'ka', then *all* that matters about a particular hypothesis is how many noise flips it must appeal to. A learner of this sort considers how to partition the data into portions coming from core vs. noise processes, without any prior assumptions about whether noise processes are more likely to produce certain types of data over others: here, whether noise is more likely to produce 'ka' vs. 'bo'; in what follows, whether it is more likely to produce subject-initial vs. subject-final clauses. In each case, the hypotheses that will be favored are those that invoke noise processes as few times as possible.

We've seen that four noise flips is better than ten, but two is even better: the very best hypothesis is that there were eight core flips and two noise flips,[3] for a likelihood of $\frac{1}{3}$. But this is as far as we can go. With more than eight core flips, the likelihood of the data under G_{ka} is zero, because there are only eight occurrences of 'ka' in the data. The observed number of 'ka' puts a cap on the degree to which core flips can be used to achieve high likelihoods.

Now let's compare how the same observed data could have been generated by G_{bo}, in which the core Φ-coins always produce 'bo' rather than 'ka'. As with G_{ka}, the hypothesis

[3] If this is surprising, it may be because of an intuition that we expect the noise flips to yield a roughly equal number of 'ka' and 'bo', contrary to our assumption of a uniform prior on ψ. But if we have no reason to suspect that the probability of a 'ka' is around $\frac{1}{2}$, then we have no reason to suspect that the number of 'ka' observations from ten flips will tend to be around the middle of the $\{0, 1, 2, \ldots, 10\}$ range.

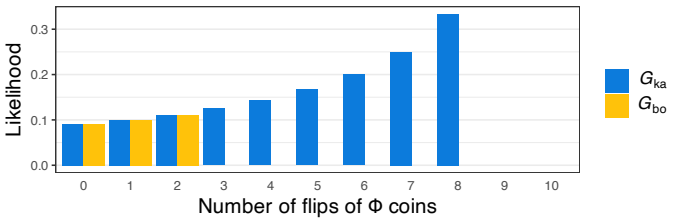


Figure 4. Likelihood of 8 'ka' and 2 'bo'

that there were zero core flips is the most costly, and alternatives that make use of larger numbers of core flips provide higher likelihoods. But since core flips in G_{bo} always produce 'bo', the best we can do is to suppose that both of the observed occurrences of 'bo' came from the core coins, and rely on eight noise flips to do the rest of the work (likelihood of $\frac{1}{9}$). So, there is no way for the core coins in G_{bo} to contribute to particularly good explanations of the observed high proportion of 'ka' in the data. Figure 4 shows the full range of ways that G_{ka} and G_{bo} can explain the observed data. The explanations made available by G_{ka} range from very good ones that invoke more flips of core Φ-coins, to more costly ones that invoke fewer flips. Since G_{bo} makes available a subset at the costly end of that range, it is intuitive that G_{ka} is a better explanation overall — even though the presence of noise coins ensures that any observed combination of 'ka' and 'bo' is compatible with both bags.

To calculate the overarching likelihood of the data under each bag, we need to consider all possible hypotheses about the number of core flips (N^{Φ}). This means marginalizing (summing) over the full range of possibilities, from zero core flips to ten — in other words, adding up the heights of the associated bars in Figure 4, weighted by the probability of each choice of N^{Φ}. See Appendix A for full details. Intuitively, if we know nothing about the ratio of core to noise coins in these bags, then the probability of any choice of N^{Φ} is the same across the eleven possible options from 0 to 10, and equals $\frac{1}{11}$. This means that each bag's likelihood is proportional to the sum of its corresponding bars in Figure 4. We find that the likelihood of the data under G_{ka} is 0.138, and the likelihood of the data under G_{bo} is 0.027.

Bayes' Rule tells us that the posterior probability of a particular bag is proportional to the likelihood of the data times the prior probability of the bag, analogous to the reasoning

in (4) in the previous section. Assuming that we have no reason to prefer G_{ka} or G_{bo} *a priori*, then the higher likelihood of G_{ka} will lead to a correspondingly higher posterior probability. We'll conclude that G_{ka} is about five times more likely to have generated the data than G_{bo}.

In sum, the key idea is that a learner's choice between competing restrictive grammatical hypotheses can be formally modeled as essentially a choice between different types of restrictive coins, embedded in a system ("bag") where flexible 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, through inference that takes the same form as the mathematical details we stepped through in this section. 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— an instance of what is often called "Bayesian Occam's Razor" (Tenenbaum & Griffiths, 2001).[4] 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 'ka' vs. 'bo').

3.1.2 Modeling the Austin et al. results. Figure 5 shows the result of applying this model to the actual data from the Austin et al. (2022) experiment, where participants observed 84 occurrences of 'ka' and 42 of 'bo'. To set up a direct comparison with the simulation in Section 2.2, we plot the inferred posterior probabilities of G_{ka} and G_{bo} for the previously-created intake datasets of different sizes. For the majority of these datasets, the model arrives at a 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'. Thus, like we found under

[4] This is conceptually similar to a related notion in Bayesian learning called the "Size Principle" (Xu & Tenenbaum, 2007), where a learner evaluates two hypotheses, one of which is strictly more flexible than the other, in that it can generate a superset of the outputs of the other. Given data perfectly consistent with both this "superset" hypothesis and the smaller "subset" hypothesis, such a learner will come to prefer the subset hypothesis. This can be seen as a simpler variant of the more complicated situation that we consider here: for instance, one in which a learner sees data consisting of only 'ka' and considers only the choice between attributing *all* data to the restrictive core component or *all* data to the flexible noise component.

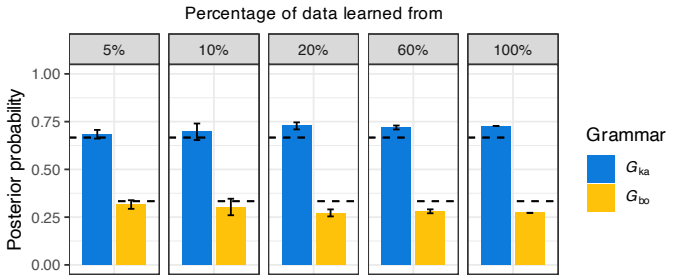


Figure 5. Posterior distribution over G_{ka} and G_{bo} across varying amounts of data randomly sampled from 84 observations of 'ka' and 42 observations of 'bo'. The dashed lines indicate the proportions of 'ka' and 'bo' in the training data.

the numerical regularization bias approach earlier, the learner arrives at a grammatical hypothesis that strengthens the dominance of 'ka' over 'bo' observed in the training data.

However, we see a difference in how learning interacts with the size of data. In the numerical regularization bias approach, a sharper skew in favor of 'ka' is a property of learning with very *small* amounts of data. We saw this prominently in the leftmost bars in Figure 2, where we modeled learning from very little data. This skew diminishes when more data are available to learn from, as the likelihood overcomes the learner's prior regularization bias. In our approach, a sharper skew in favor of G_{ka} only emerges as *more* data are 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 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 this approach could explain 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; alternatively, we might imagine that children choose either G_{ka} or G_{bo} on each production trial, with these same probabilities. Either case would lead to children's production of 'ka' at a higher rate than observed in the training data. The important point is that 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 computational approach outlined in Section 2.2.

Perkins et al. (2022) applied a similar approach to model a naturalistic phenomenon in language acquisition that resembles regularization: how learners identify 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 G_{ka} provides a good explanation for data that do not consist entirely of 'ka'.

In the general architecture that we now introduce, we scale up this approach to what we call a Noisy Grammar Learner: a learning framework for acquiring systems of interacting rules. In the more complex linguistic case studies that follow, the analogues of G_{ka} and G_{bo} will be the restrictive core grammatical rulesets that a learner is entertaining. Importantly, the learner's inference in these case studies takes exactly the same form as the details we stepped through carefully above. Similar to how the "core" coins in G_{ka} and G_{bo} always produce 'ka' or 'bo', the learner might consider the hypothesis that the core basic word order of a natural language is SVO or SOV. The analogues of the unpredictable "noise" coins will be other grammatical processes that may introduce distortions into the learner's data, such as non-canonical constructions that produce divergences from a core word order. Choosing between G_{ka} and G_{bo} will be analogous to choosing which grammatical hypothesis provides the best explanation for data that are a distorted reflection of these core grammatical rules.

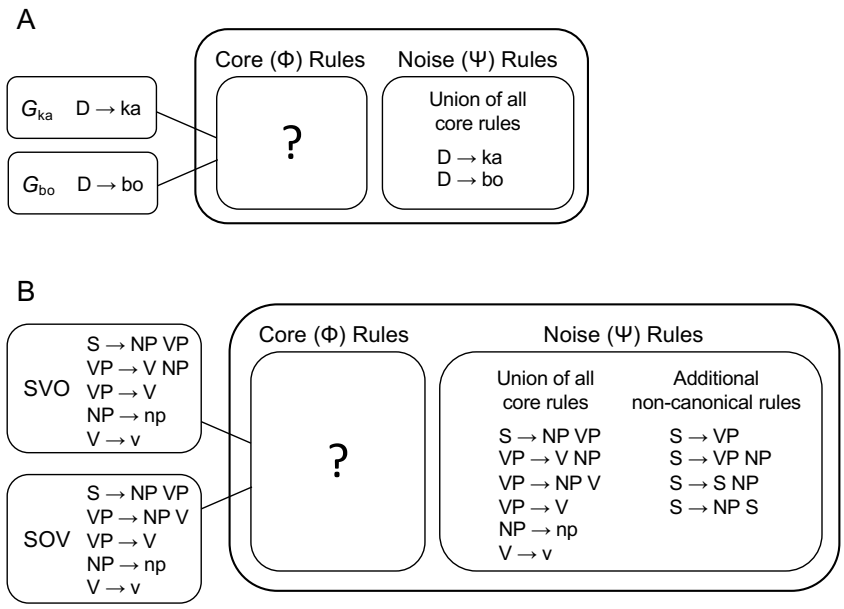


Figure 6. The hypothesis spaces of two example Noisy Grammar learners

3.2 Noisy Grammar Learners: From surface forms to grammar rules

3.2.1 Recasting "bags of coins" as grammars. To step up to the architecture of a Noisy Grammar Learner, we'll first recast our account of the Austin et al. (2022) results in terms of a choice between two extremely simple sets of grammatical rules, whose probabilities are analogous to the parameters of the bags of coins in the previous section. We can represent the relationship between the G_{ka} and G_{bo} hypotheses as in Figure 6a, where the two possible realizations of a determiner, 'ka' and 'bo', are expressed as two possible rules for rewriting the nonterminal symbol D. More precisely, these are rules of the sort that appear in a Context-Free Grammar (CFG).[5] Both of these rules are available as "noise" (Ψ) rules under either hypothesis; where the two hypotheses differ is which of the two rules they choose to include in their core (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

[5] 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.

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

(a) For G_{ka}

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

(b) For G_{bo}

Figure 7. Compiled-out PCFGs for the two grammars in Fig. 6a

form of a 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 7. 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., $\psi_{\mathrm{D}\rightarrow\mathrm{ka}} + \psi_{\mathrm{D}\rightarrow\mathrm{bo}} = 1$). Similarly, because these examples each only include one possible rewrite of D_{Φ}, the parameter $\phi_{\mathrm{D}\rightarrow\mathrm{ka}}$ in G_{ka} and the parameter $\phi_{\mathrm{D}\rightarrow\mathrm{bo}}$ in G_{bo} will necessarily be 1. But the generic notation in Figure 7 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.

These parameters are analogous to the probabilities that we saw in our "bags of coins" example. The parameter ϵ_{D} is analogous to the probability of selecting a noise vs. core coin for a particular coin flip: it expresses the ratio of Φ-coins to Ψ-coins in a bag. The ϕ parameters express the probabilities of the Φ-coins producing 'ka' vs. 'bo,' which are always 1 or 0 under a given grammar. The ψ parameters express the probabilities of the Ψ-coins producing 'ka' vs. 'bo,' which can vary between 1 and 0. So, the rule-based system that we introduce here expresses the same probabilistic model as the earlier bags-of-coins system.

Each of the two compiled-out PCFGs in Figure 7 allows exactly three derivations. Those allowed by G_{ka} are shown in (5), along with their probabilities. Note that these rules implement precisely the branching structure for the bags of coins in Figure 3. 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. The total probability assigned to the realization of D as 'ka' is the sum of the two corresponding derivations' probabilities.

(5)

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

The example we worked through in the previous section can be thought of as asking which of the two PCFGs in Figure 7 provides a better explanation for a corpus of 126 observed words: 84 'ka', and 42 'bo'. The system we propose below extends this idea to allow for noise not only in choices of rules that introduce single words (e.g. D $\rightarrow$ 'ka' or D $\rightarrow$ '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 $\rightarrow$ NP VP or S $\rightarrow$ 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 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 7 provides the best explanation for the observed strings. Crucially, this inference follows the same logic that we saw when the grammars were construed as bags of coins: just as the choice of noisily or non-noisily rewriting a single nonterminal (e.g., D) is analogous to the choice of flipping a noise or a core coin from a bag, the choices involved in rewriting a sequence of nonterminals are analogous to the choices involved in a sequence of coin flips, each of which could be a noise or a core coin.

3.2.2 Noisy CFG Learners. Here we instantiate the notion of a Noisy Grammar Learner for the particular case where the grammatical rules involved take the form of CFG rules. We'll define the learner below, and step through its inference mechanism using a

simple example hypothesis space; this provides the core mathematical proposal that we then test in simulations on naturalistic data in Sections 4 and 5. The mathematical properties on which our learning architecture relies are compatible with many different choices of grammatical formalism.[6] We use the CFG formalism here for the purposes of illustration, and because this is sufficient for formalizing the case studies in Sections 4 and 5.

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 $R_i \subseteq R$).[7] 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 example above, $R_1 = \{\text{D} \rightarrow \text{ka}\}$, $R_2 = \{\text{D} \rightarrow \text{bo}\}$, and $R = \{\text{D} \rightarrow \text{ka}, \text{D} \rightarrow \text{bo}\}$. However, in general there may be additional noise rules in R that are not part of any candidate Φ component, which allow possibilities that are non-canonical under all of the learner's hypotheses. The learner infers the probabilities of the different hypothesized Φ components: the different forms that canonical, non-noisy sentences can take.

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. Just as before, we will assume that the learner has no prior commitments to these parameters taking any particular value over another, which means that the same logic that

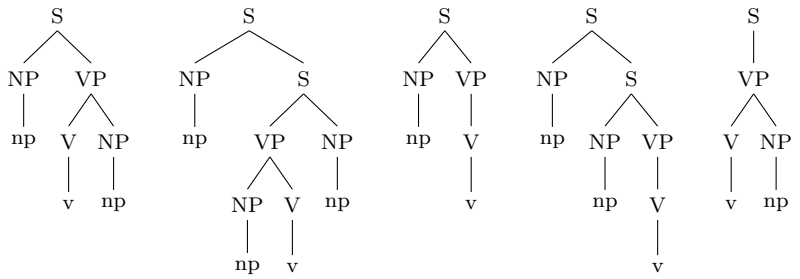
[6] For example, finite-state grammars (which are most likely appropriate for phonological and morphological acquisition) correspond to a sub-case of CFGs that only allows right-branching. Many different kinds of "mildly context-sensitive" grammars (Joshi, 1985; Stabler, 2004) can be characterized in ways that share the same abstract branching structure of CFG derivations, while extending their linguistic expressivity in a way that might be needed to characterize the learning of more complex syntactic structure (e.g. D. Weir, 1988; Seki, Matsumara, Fujii, & Kasami, 1991; Stabler, 2011).

[7] 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.

applied in working out the likelihood of the data under a bag of coins will also apply here.

The choice of what goes into the hypothesis space of a Noisy CFG learner reflects theoretical commitments about the knowledge that a learner brings to the language acquisition task. We will use the very small hypothesis space in Figure 6b to illustrate the basics of the model here. This amounts to equipping the learner with two basic word orders to choose from— SVO and SOV— corresponding to the two candidate Φ components. In either case, a sentence is made up of a noun phrase (NP) followed by a verb phrase (VP). A verb phrase may be a bare verb (V) or may in addition contain another NP. Where the two candidate Φ components differ is in whether this optional NP occurs before or after the V. The labels SOV and SVO for these two respective options are based on the purely structural classifications of the optional VP-internal NP as an object, and of the VP-external NP as a subject. In deciding between these options, the learner assumes that its data will come not only from the to-be-identified Φ component but also from the noise rules in the Ψ component. In Figure 6b, these noise rules generate sentences where the VP either has no subject (no sister NP) or has a subject on its right, and sentences with additional NPs. The learner is aware that these are the divergences from the core SVO/SOV word order that will complicate the task of using its data to make the SVO/SOV decision. The hypothesis spaces for our case studies below use an expanded version of the range of possibilities in Figure 6b.

All of these assumptions correspond to hypotheses about the learner's prior expectations about linguistic structure: the learner's *inductive biases*, at least some of which may be domain-specific in nature. The linguistics literature often uses the term "Universal Grammar" (e.g. Chomsky, 1965, 1980, 1986) to refer to the innate, domain-specific components of the capacities that a learner's prior expectations derive from.[8] Our framework is agnostic about the sources of the hypotheses in the learner's hypothesis space: whether these derive from Universal Grammar, from linguistic knowledge gained at an earlier stage of a learner's development, or from another area of cognition. What the framework provides is



(a) Input perceived as trees

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

(b) Counts of rules in trees

Figure 8. Example of data, parsed as trees, for a Noisy CFG learner

a way to formally express these hypotheses, whatever their source.

3.2.3 Likelihoods of trees in a Noisy CFG Learner. In the simulations that follow, 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, and 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 the task of calculating the likelihood of a collection of *trees* under a given hypothesized Φ component; this serves as the basis for calculating string likelihoods (see Appendices A and B for detail). Imagine a learner observes the trees in Figure 8a. How likely are these data under each of the two Φ components in Figure 6b? Neither of these Φ components can generate this collection of trees directly, but there is an intuition that the trees look more "compatible" with SVO than SOV. We can explain this using the same reasoning as we saw with the bags of coins.

Via the standard independence assumptions of context-free grammars, observing the trees in Figure 8a amounts to observing the counts of rewrites in Figure 8b. Furthermore, we can treat the rewrites of S nodes independently from the rewrites of VP nodes: the likelihood of the trees is the product of the probability of seven S rewrites breaking down as shown and the probability of five VP rewrites breaking down as shown. (The rewrites of the

[8] The term "Universal Grammar" is often associated with more intricate kinds of grammatical rules than the ones shown in Figure 6b, but the general computational framework that we describe here is compatible with a wide range of choices regarding the format of the Φ rules and Ψ rules, including many that resemble those more common in the contemporary syntax literature and would be necessary for modeling the learning of more complex grammatical phenomena. We return to this point in Section 6.

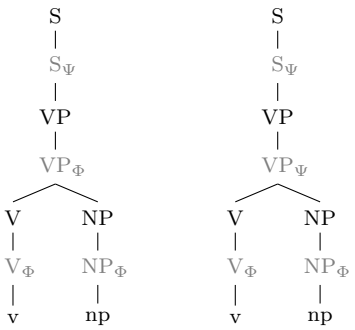
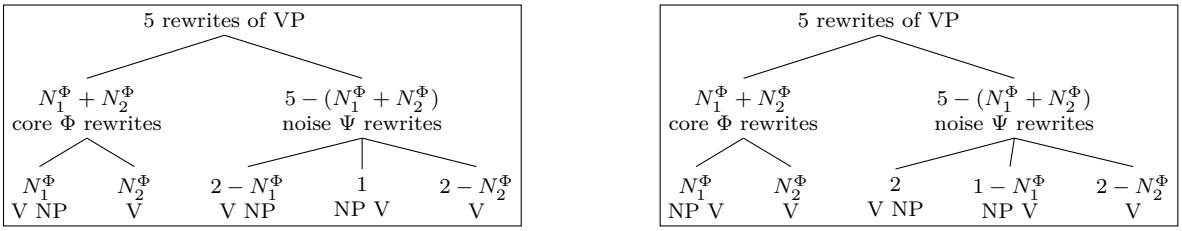


Figure 9. Two possible articulated trees for the rightmost tree shown in Fig. 8

NP and V nodes have probability one, and can be ignored.)

Underlyingly, each choice of Φ component corresponds to a compiled-out PCFG of the sort in Figure 7. These PCFGs generate trees that include additional layers of structure encoding the choice to treat each rewrite as either core or noise. Two examples are shown in Figure 9; we call these *articulated trees*. The probability of a standard tree like the ones in Figure 8 is the sum of the probabilities of its articulated trees, just as the probability of 'ka' in (5) is the sum of the probabilities of two derivations, one via D_Φ and one via D_Ψ. What is more interesting here is that this choice between core and noise mechanisms can be made independently at each node, even within the same tree. In both of these trees, the $S \rightarrow VP$ 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 they 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 PCFG for, say, the Φ component representing SVO word order, and given values for the rule probabilities — one for $VP \rightarrow VP_\Psi$, which we could call ϵ_{VP}, and one for $VP_\Phi \rightarrow V\ NP$, which we could call $\phi_{VP \rightarrow V\ NP}$, etc.— we could calculate the probability of each tree in Figure 8a by adding the probabilities of all of the possible core and noise choices at each node. But since we do not know the specific values of the PCFG's rule probabilities, we'll calculate the likelihood by marginalizing over all of these possible values, just as we did



(a) Under hypothesis G_{SVO} (b) Under hypothesis G_{SOV}

Figure 10. Partitioning the VP rewrites in Figure 8 into core and noise

when calculating likelihoods under a bag of coins. We assume a uniform prior over all possible values of these rule probabilities: we make no assumptions about the probability that a nonterminal will rewrite one way vs. another among its various possibilities, just as we made no assumptions about the probability that a noise coin would produce 'ka' vs. 'bo'. This means that we can calculate this likelihood via the same logic as in the coins example.

Focusing on the VP nodes, we see that the five rewrites of VP resulted in two occurrences of V NP, one of NP V and two of just V. Under the SVO hypothesis, the NP V rewrite would necessarily be noise, and the other four could be either core or noise. Under the SOV hypothesis, both of the V NP rewrites would need to be noise, and the other three could be either core or noise. Figure 10 shows the possible divisions under each hypothesis, where N_1^Φ and N_2^Φ are the counts of the transitive and intransitive core options. Let's consider the likelihood that these VP rewrites arose via a specific combination of core and noise under each hypothesis, i.e. via specific values for N_1^Φ and N_2^Φ. This likelihood is the product of the probabilities of this specific way of dividing the total rewrites into core vs. noise (at the top of Figure 10a), dividing the core rewrites into two groups (on the left), and dividing the noise rewrites into three groups (on the right); see Appendix A for details.

Just as we saw with the bags of coins, it turns out that the "best" hypotheses about how observations might break down into core vs. noise tend to be ones that invoke as many core mechanisms as possible. As the total possible number of core rewrites increases under a particular grammar, in general so does the likelihood of these VP rewrites. This example differs from the bags of coins example in having two core ways to rewrite VP, unlike the G_{ka}

and G_{bo} hypotheses where there is only one core outcome. But what matters is that the range of core Φ 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 core mechanisms will be preferred, because the core options share probability mass with fewer competitors: in most cases, there are fewer ways to divide a given number of rewrites into two groups than into three. The earlier examples illustrated the special case where the Φ component was completely deterministic, so there was no competition among core options.

To calculate the overall likelihood of the VP rewrites, with no particular choice of N_1^Φ and N_2^Φ, we need to marginalize (sum) over the full range of possible values for each of these variables; see Appendix A. Because there are more ways that the observed VP rewrites could have come from the Φ rules under G_{SVO} than under G_{SOV}, the two grammars differ in the range of values contributing to this sum. In particular, N_1^Φ (the number of core transitive VPs) could be any value from 0 to 2 under G_{SVO}, but for G_{SOV} it is capped at 1. This results a higher likelihood for G_{SVO}: 6.07×10^{-2}, compared to 3.71×10^{-2} for G_{SOV}.

An analogous calculation can be done for the likelihoods of the S rewrites in Figure 8. The likelihood of the trees in Figure 8 under a grammar is the product of the likelihoods of the VP and the S rewrites. Assuming that neither grammar is favored *a priori*, we find that G_{SVO} has higher posterior probability given these trees: 0.621, vs. 0.379 for G_{SOV}.

3.2.4 Full inference 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, $P(G \mid \vec{w})$.

In what follows, 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. As described above, we set a uniform prior over the probabilities associated with the rewrites of each nonterminal, so the model has no preference for or against assigning probability mass to any particular rewrite. In the simulations below we will compare this against an implementation of the numerical regularization approach, where the prior over rewrite

probabilities is numerically skewed to favor rule probabilities close to zero or close to one.

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 $\vec{t}$, e.g., Figure 8), it was possible to analytically calculate the posterior probability of a grammar given these trees and their strings, $P(G \mid \vec{t}, \vec{w})$, by marginalizing over the rule probabilities ($\vec{\theta^G}$) and the possible divisions of rewrites into signal and noise. But calculating the posterior probability of a grammar given only the strings, $P(G \mid \vec{w})$, would require further summing over all the possible ways to choose a tree for each string in the data. This calculation is intractable. So we instead estimate a joint posterior distribution over both trees and strings, $P(G, \vec{t} \mid \vec{w})$, using a technique for approximate probabilistic inference called 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 the posterior probability of grammars given the trees and strings, $P(G \mid \vec{t}, \vec{w})$, and then sampling new trees according to the posterior probability of trees given a grammar and the strings, $P(\vec{t} \mid G, \vec{w})$. This process will produce a sample from the joint posterior distribution over G and $\vec{t}$; ignoring the trees, we are left with a sample from the posterior distribution over G. See Appendix B for details.

3.3 Summary

In this section we first 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. We then showed how this approach generalizes to more complex systems of interacting grammatical rules, to define what we call a *Noisy Grammar Learner*. Given data that reflect a mixture of core and noise mechanisms, the learner evaluates three questions, corresponding to the branch-points in a choice for how to partition the data into core vs. noise: (i) What do the data from the core rules look like? (ii) What do the data from the noise rules look like? (iii) What is the right division into core 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 related aspects of natural language syntax acquisition can be set up as a problem of the form in Figure 6, with a choice among particular sets of core rules against the backdrop of a particular set of noise rules.

4 Simulation 1: Learning basic word order

In the following simulations, we show that the approach of deciding among competing Noisy CFGs can be applied to model naturalistic phenomena in language acquisition that involve drawing systematic generalizations from messy data. Our first simulation models the problem of acquiring the basic positions of subjects and objects despite opacity introduced by non-canonical word orders. Our second simulation models the problem of additionally identifying the affixes that mark subjects vs. objects in a language with case-marking, from data that contain not only non-canonical word orders, but also case-marker optionality.[9]

To recapitulate the learning problem introduced in Section 1, children identify the canonical word order of their language in infancy, at ages 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; Gagliardi et al., 2016; Perkins & Lidz, 2020, 2021). These "non-basic" clause types are common— *wh*-questions comprise approximately 15% of the input to English-learning 1-year-olds— and therefore create substantial distortions in the body of evidence that children have available to identify the structure of basic clauses[10] (Cameron-Faulkner, Lieven, & Tomasello, 2003; Stromswold,

[9] Code and data for all simulations can be found at `https://github.com/perkinsl/noisy-grammar-learner`.

[10] The notion of a "basic clause" has various characterizations. Pinker (1984), following Keenan (1976), characterizes basic clauses as "roughly, those that are simple, active, affirmative, declarative, pragmatically neutral, and minimally presuppositional." Here, we use the term "basic word order" to mean the word order in something similar to Chomsky (1957)'s notion of a kernel sentence, where no optional transformations have applied. In languages like English, French, and Japanese, learners must come to identify that one primary order of clause arguments is possible in these sentences, although other orders may be possible through optional transformations such as *wh*-movement and scrambling. In a more modern framework, this notion of "basic word order" could be taken to mean the surface word order that most transparently reflects the dependencies that arguments obligatorily enter into (e.g. theta role assignment and Case).

1995). How do learners manage to abstract away from this messiness in order to accurately identify basic word order? Some accounts assume that learners can "filter" non-basic sentences, ignoring them for early syntax learning (Pinker, 1984). But if learners do not know what counts as basic, it is not obvious how they identify which sentence types are *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 issues of circularity.

We tested our model on child-directed English, French, and Japanese. Both English and French are canonically SVO, but allow different types of argument movement. Japanese is canonically SOV, but has a large amount of argument movement along with argument-drop. Our model successfully identifies the target grammars for its noisy data in all three languages. Moreover, our model fares better than a learner that is capable of transparently 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 succeeds in this learning problem where the previous numerical regularization approach does not.

4.1 Noisy CFG learner for word order

Our learner's hypothesis space is an expanded version of the example in Figure 6b. It consists of four sets of Φ-rules and one shared set of Ψ-rules, giving rise to the four Noisy CFGs in Figure 11. Here, we explain more fully the structure of this hypothesis space.

Because the modeled learner is attempting to identify the canonical positions of subjects and objects, the learners' Φ-rules generate the core predicate-argument structure of basic clauses with subjects and optionally objects. We use the terms "subject" and "object" to express a structural asymmetry between core clause arguments that we assume is represented by a learner, and which is again cashed out in the CFG rules by defining "subject" as the NP daughter of S and sister of VP, and "object" as the NP daughter of VP and sister

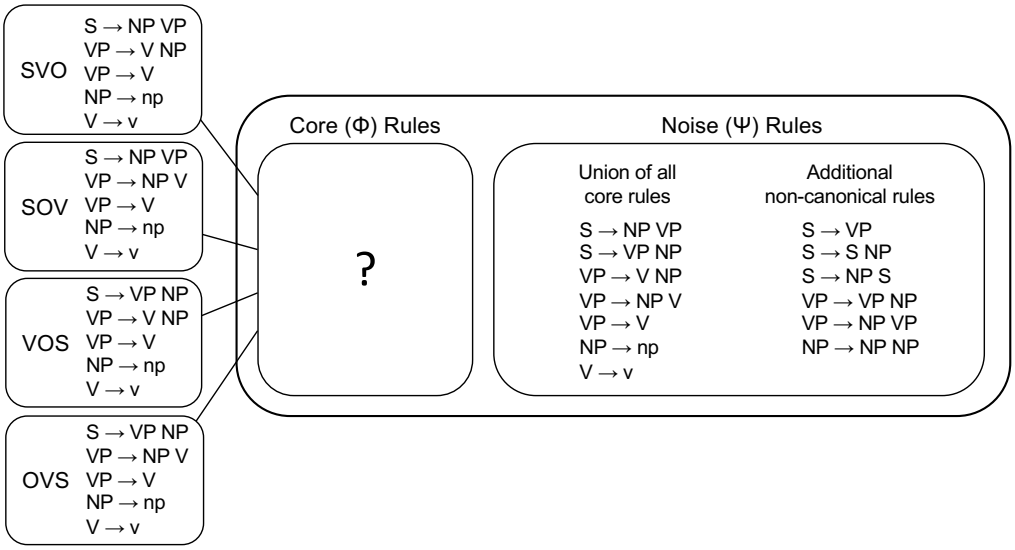


Figure 11. Hypothesis space of Noisy CFG learner for word order

of V. This abstracts away from many details of the ways in which these grammatical relations can be formally characterized, and we remain agnostic as to whether these relations and structural positions are derived or are grammatical primitives; the important point for our purposes is that the learner's hypothesis space encodes a particular structural relation, where objects form a constituent with the verb independent of the subject (Baker, 2001).

These Φ-rules have two important properties, encoding the learner's assumptions about the grammatical phenomenon to be acquired. In what follows, we show that these properties play a key role in the learner's ability to draw the correct word order generalizations. The first property is a set of substantive, domain-specific expectations about the nature of the rules governing basic clause syntax, encoded in all four sets of Φ rules. These include the expectation that a sentence will obligatorily have an NP that is the sister of a VP, whereas a VP may or may not contain an NP that is the sister of its V; this encodes a widely-held assumption that subjects in canonical clauses are required in ways that objects are not (e.g. Keenan, 1976; Chomsky, 1982).[11] Furthermore, noun phrases must occupy either canonical

[11] We abstract away from the various proposals for why subjects may be required, and note that this is also debated. As McCloskey (1997) puts it, "It has sometimes been claimed that every clause must have a subject. This is not obviously correct, but it is clearly correct in some broad sense for some languages. There is no

subject or canonical object position, reflecting the learner's belief that adjunction is in some sense non-canonical. The second important property of the Φ rules is the ways in which the four candidate rule sets differ, which express the choice points available to the learner. Each of the four options deterministically puts subjects before or after verb phrases and objects before or after verbs, yielding a four-way choice of restrictive canonical word order: SVO, SOV, VOS, OVS. This leaves aside questions about natural languages where the canonical word order separates the verb from the object (VSO and OSV), which we return to Section 6.

All four grammars share the same set of Ψ-rules, encoding the learner's hypotheses about the non-canonical processes that will introduce noise into the data. These allow non-canonical clauses to contain exceptions to the learner's expectations about basic clause syntax: for instance, the rule S → VP allows non-canonical clauses to lack subjects, and the rule S → NP S allows NPs to be adjoined to a S rather than occupying argument positions. These are analogous to some of the Ψ-rules that we saw in Figure 6b in Section 3. These Ψ-rules do not encode the full range of non-basic constructions that the mature grammar will eventually include (e.g., *wh*-movement, relativization, raising); instead, they encode "placeholder" rules that simply allow for all permutations and deletions of NP arguments, and for additions of non-argument NPs.[12] This models a specific stage in the incremental process of grammar acquisition, motivated by the previously-reviewed empirical findings, in which a learner is only acquiring basic word order, leaving the full details of these other grammatical constructions to be acquired at a later developmental stage. In Section 6, we return to the question of how this development might proceed. 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 flexible noise coins and restrictive core coins.

Crucially, while the learner's rules contain hypotheses about which canonical and

---

other argument-type or syntactic position for which this claim can be made with even remote plausibility. There are no languages, as far as I know, for which it has ever occurred to anyone to claim that every clause must have a direct object, or an indirect object or a prepositional complement or whatever" (p. 198).

[12] We talk about these intuitively as additions, permutations, and deletions of NP arguments, but note that noisy analyses of strings are not derived from core rule analyses in any respect.

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. Our learner marginalizes over all possible choices of these parameters to infer which Noisy CFG best explains its messy data.

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. However, in restricting our attention to this type of distributional learning, we do not imply that children ignore other potential correlations between meaning and syntax, between prosody and syntax, or even between different syntactic phenomena (for instance, knowing that postpositions tend to correlate with pre-verbal objects) (Christophe, Millotte, Bernal, & Lidz, 2008; Morgan & Demuth, 1996; Pinker, 1984; Greenberg, 1970). We abstract away from these other sources of information in order to ask how much ground a learner might be able to gain by applying this learning mechanism to imperfectly-perceived string distributions, noting that a similar mechanism could be generalized to make use of correlated information from other domains.

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; for full details see Appendix C.[13] Based on empirical evidence, we assume that infants at this age

[13] We are operating with the simplifying but potentially tenuous assumption that children at this age can represent verbs and noun phrases in their input *as such*: that is, not only have they managed to categorize

	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

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; Marquis & Shi, 2012; Haryu & Kajikawa, 2016; He & Lidz, 2017); see Dye, Kedar, and Lust (2019) for a review. For example, we assume that an English learner can recognize *you* as a pronoun, *the* as a member of a functional category that precedes nouns (which we call "determiner"), *will* as a member of a functional category that precedes verbs (which we call "auxiliary"), and *-ed* as a verbal suffix; these are four of the elements that met our criteria for being recognizable, which included being among the 100 most frequent tokens in the corpus.

We use some simple heuristics to guess at the positions of verbs and noun phrases in the corpus on the basis of only the recognizable functional elements. 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 positions are those preceding a case-marker, and v-cue positions include those preceding the negation

words into two classes based on their distributions with functional elements, but they also know which of these classes head noun phrases (can act as clause arguments) and which head verb phrases (act as core clausal predicates). How children would arrive at these category labels, and whether they can do so before learning word order, is an important question that we leave for future work. For discussion of this issue, see Gutman, Dautriche, Crabbé, and Christophe (2015), He and Lidz (2017), Mintz (2003), and Pinker (1984), among many others.

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

marker *-nai*. *Wh*-words and object clitics were not identified as np's, because they may not be recognized as such by infants at this age (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 the learner does not know which case-markers mark subjects vs. objects (a problem considered in Section 5).

This allows us to map each corpus utterance to a string consisting of any number 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 of a single clause are arranged. Table 2 shows the proportions of the relevant string types for each language.

We created datasets of strings for each language sampled according to the distributions in Table 2. For computational tractability, our largest dataset was 50 strings. We also tested a range of smaller datasets (10-30 strings) in order to explore the effects of dataset size on learning; recall that in the numerical regularization simulations in Section 2.2, regularization was a property of learning from small datasets. 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). No dataset can be directly 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 12 displays our model's inferred posterior probability distribution over the four Noisy CFGs in its hypothesis space, averaged over 10 runs of the model on each dataset. Within each language, the model's posterior distributions take the same qualitative shapes across datasets of varying sizes. Just as in the artificial language 'ka'/'bo' simulations in Section 3.1, 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 successfully overcame 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 evidence for the correct target word order.

We find an interesting additional result in Japanese. With sufficient data, the Japanese learner 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?

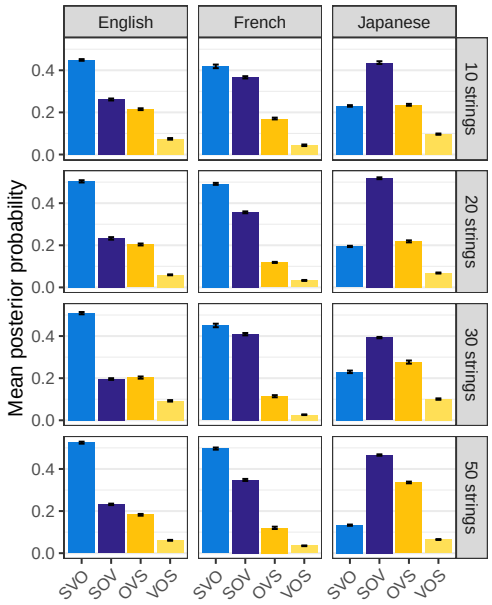


Figure 12. Posterior distribution over word-order grammars, Noisy CFG learner

This difference reveals some subtleties in how the learner makes use of its data. 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 much less 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 → 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. That is, the more the learner gained confidence that the noise rule S → VP was the most probable way to rewrite an S, the more it gained confidence that the string np v should be analyzed with this noisy S rule in combination with the core rule VP → NP VP. This VP rule is included in the core rules of the OVS grammar but not the SVO grammar, 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. 12 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. Above we saw that the Japanese learner came to prefer OVS over SVO, even through the core rules of SVO could account for more strings. This illustrates 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 13, 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 hypothesis space in English and French, and SOV highest posterior probability in Japanese (Fig. 13, 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? Our learner considers a tradeoff between fit to the data and restrictiveness of its hypotheses,

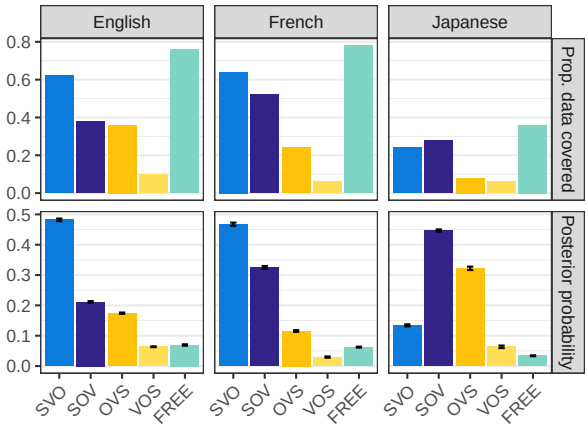


Figure 13. Five-way hypothesis space: Proportion data coverage vs. model's posterior distribution (50-sentence datasets)

again an instance of "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 restrictive core coin rather than a flexible noise coin.

These findings demonstrate the flexibility and robustness of this learning mechanism. Our learner identifies the target strict word order 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 12 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 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 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 simulation in Section 2.2, thereby assessing how our model compares to the general regularization bias approach taken in previous literature.

There are two important properties that distinguish our Noisy CFG model's hypothesis space from that of the fully-flexible learner. First, our model's hypothesis space encodes an expectation that the grammatical systems generating its data comprise a mixture of restrictive core rules and more flexible noise processes. Second, it also encodes substantive expectations about the content of the core rules: specifically, it expects canonical clauses to have subjects and not to have adjuncts. Comparing our model to a learner that lacks these properties provides insight into the role that they played in our model's success.

We tested two variants of the fully-flexible 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 2.2. For each collection of probabilities for the rules expanding a given nonterminal, we use a symmetric Dirichlet prior with parameter α; the Dirichlet distribution is a generalization of the Beta distribution to more than two outcomes. When $\alpha < 1$, this again has again 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 one very extreme value of α (0.0001) 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 distribution over $\vec{\theta}$ displays these properties, we examine the learner's posterior distribution over trees for its data. We estimate this posterior via 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 trees 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 14-15, 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 both the unbiased and biased model results only for the 50-sentence dataset (Figure 14). Results from the biased model for the smaller datasets are plotted in Figure 15.

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 our Noisy CFG learner, we analyze the learner's samples for the 50-sentence dataset (Figure 14). For the unbiased learner, the number of sampled treesets that we are able to analyze is high: 99% of the English samples, 90% of the French samples, and 94% of the Japanese samples 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

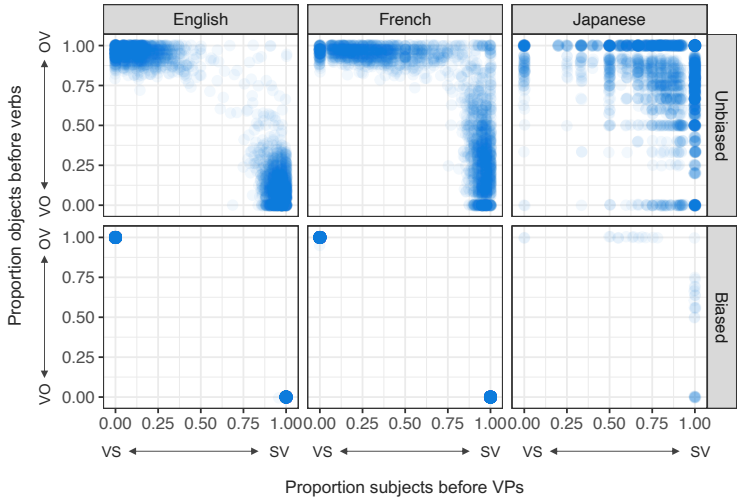


Figure 14. Posterior distribution over subject and object position in sampled treesets ($\vec{t}$), fully-flexible learner, 50-sentence dataset

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 (Figure 14, 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. In 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 corners corresponding to canonical word orders (Figure 14, 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 restrictive corners rather than the flexible middle, but it did not identify SVO as a better corner than OVS. With smaller datasets (Figure 15), the learner put additional probability mass on the SOV corner, but still did not successfully discriminate SVO from OVS.

In Japanese, the number of sampled treesets that we can analyze for the biased learner

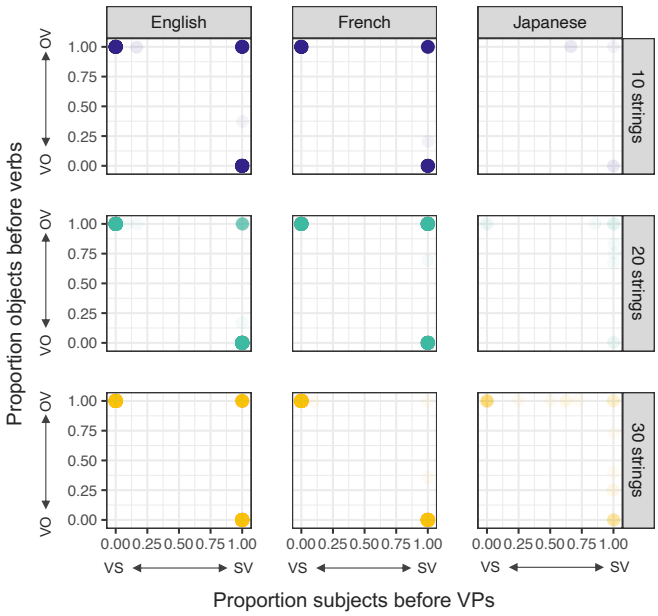


Figure 15. Posterior distribution over subject and object position in sampled treesets ($\vec{t}$), fully-flexible biased learner, smaller datasets

is surprisingly low, resulting in very few data points plotted for that learner for any size of dataset in Figures 14–15. 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 → 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? Recall that there are two 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: (i) an expectation that its data are generated via a mixture of restrictive core rules and more flexible "noise" rules, and (ii) an ability to encode substantive expectations about the nature of those core rules.

The specific 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 should emerge for the canonical-subject analysis, again paralleling our preference to analyze a sequence of 'ka' as coming from a restrictive core 'ka' coin, rather than a flexible noise coin. In our comparison against the "data-coverage" heuristic, we saw that this preference for restrictive hypotheses can inform the learner's choice *across* grammars that vary in restrictiveness. Here, this same mechanism should apply *within* each grammar, informing the choice to attribute 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. The prevalence of bare v strings in Japanese provides evidence that unary-branching structures (S dominating only VP, VP dominating only V) are common. For the fully-flexible learner, strings that do contain np's can make use of these common substructures if the np's are analyzed 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 flexible noise rules. The original Japanese learner concluded that its core rules for introducing subjects were likely less helpful than the noise rule $S \rightarrow VP$, which is needed to account for bare v strings. But these strings leave open the option to use the core rules for introducing objects, which the learner indeed uses to analyze common strings with np's. Our learner's pressure to use its restrictive core rules whenever possible is likely what prevented it from analyzing all np's as adjuncts, allowing it to draw firmer conclusions about subject and object positions than the fully-flexible learner did.

Thus, the learner's two sorts of substantive expectations work in formally similar ways. The asymmetry in the hypothesis space between restrictive core rules and flexible noise rules leads the learner to prefer core rule analyses. By allowing only particular sorts of core rules, we are able to express expectations that canonical clauses have particular shapes. This likely contributed in important ways to the learner's ability to draw inferences from its noisy data.

We test this possibility by focusing specifically on the learner's expectation that canonical clauses require subjects. Does our learner's success in English and French depend on its substantive expectations about the core rules for subjects, or does it only depend on the distinction in the hypothesis space 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 work in formally similar ways, investigating 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 grammar 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, because the Japanese learner already needed to analyze much of its data as subjectless. But it should lead to worse performance in English and French, because it will encourage these learners to distribute probability among their core rules in a different way than before: now these learners will not face pressure to use rules for introducing subjects over this rule for subjectless clauses.

Figure 16 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; the content of those rules also plays a large role.

4.4 Summary

This case study shows that the approach of deciding among competing Noisy CFGs can be applied to successfully model a learning problem in natural language syntax acquisition which resembles the phenomenon of regularization. Our model learns basic word order from the noisy sentence representations available to a young infant. Using distributions

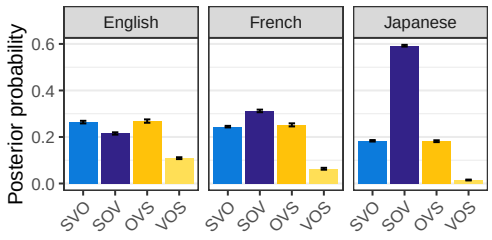


Figure 16. Posterior distribution over word-order grammars, Noisy CFG learner without subjects required canonically (50-sentence datasets)

of imperfectly-identified noun phrases and verbs, our model successfully infers that English and French are SVO and Japanese is SOV. It does so by partitioning its data into portions demonstrating core word order and noise coming from non-canonical structures, effectively implementing the idea that young learners "filter" non-basic clauses from their data (Pinker, 1984). Because the learner's hypotheses contain specific restrictive core rules, a preference emerges to use those core rules to explain skews in the data, 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.

We further find that our model succeeds where the general regularization bias approach does not. 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. Second, it encodes substantive expectations about the nature of those core rules: canonical clauses have subjects, and noun phrases canonically occupy argument positions. Without these properties, the learner is unable to draw accurate inferences from its noisy data. This suggests that, for this learning problem, it is important for learners to have specific expectations about the nature of basic clause syntax. By embedding these expectations within a noise-tolerant system, our learning architecture makes it possible for learners to recover the target basic clause structure for the language from data that appear inconsistent on the surface.

5 Simulation 2: Learning case-marking

In our second case study, we demonstrate how our approach generalizes to a related learning problem within early morphosyntax acquisition. Whereas our first case study explores the problem of acquiring word order, our second simulation explores the problem of acquiring word order in tandem with acquiring a case-marking system where 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), and also does not know the canonical order of subjects and objects in the language. We show that our learning architecture can be successfully applied to this scenario.

We conducted simulations on a new dataset generated from child-directed Japanese, in which noun phrases sometimes 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 model a learner that is attempting to identify the grammatical relations marked by *ga* and *o* in tandem with 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 succeeds where learner implementing a numerical regularization bias does not.

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 hypothesis space from

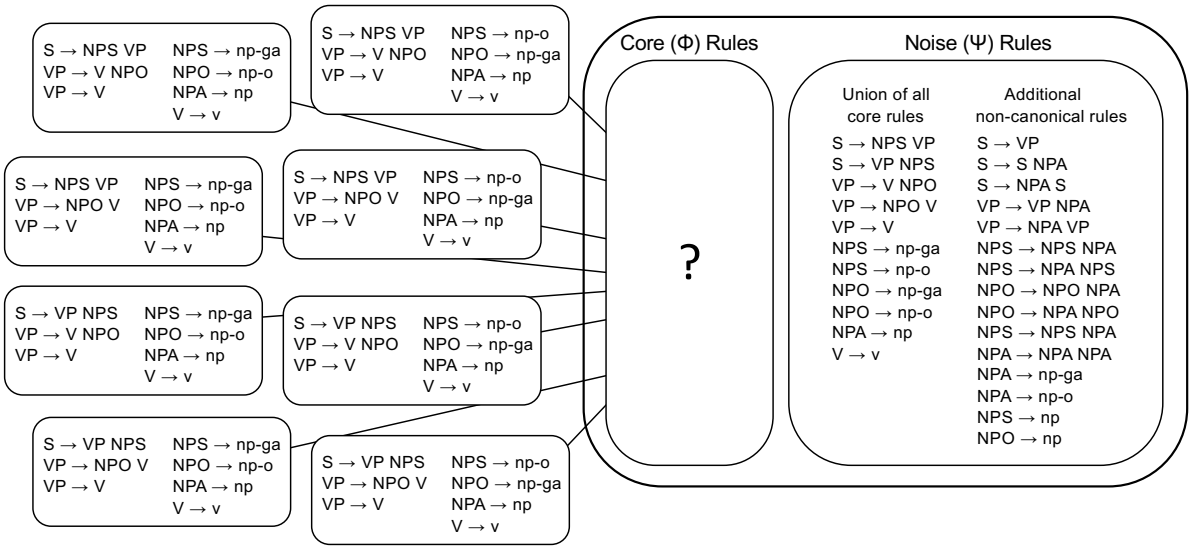


Figure 17. Hypothesis space of Noisy CFG learner for case-marking

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. The learner thus chooses among the eight Noisy CFGs in Figure 17. The Φ-rules are similar to those in Section 4: they generate basic transitive and intransitive clause structures, with subjects obligatory and objects optional. Subjects and objects are again defined by their structural positions, but here they are represented with distinct nonterminal symbols (NPS and NPO), which introduce distinct case markers. 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 the 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 restrictive core rules for case-marking and flexible

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

noise rules. 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.

5.2 Data

The learner again observes strings of imperfectly-identified verbs and noun phrases, some now affixed with *ga* and *o*. We 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 previously represented simply as np were here subdivided into three types: those followed by 'ga' (np-ga), those followed by 'o' (np-o), and all others (np).[14] For simplicity, we test our learner on one dataset of 50 strings sampled in their relevant proportions.

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. Because case-markers are omitted very frequently in Japanese, this restriction is needed in order to ensure that our learner is able to observe case-marked noun-phrases in sufficient proportions in its 50-string dataset. This restriction means that this model does not learn from strings consisting of bare v, unlike the learner in Section 4; whereas the bare v strings had a particular skewing effect

[14] This represents a simplification of the learning problem in Japanese, where children must identify the functions of *ga* and *o* among several other case-markers, and differentiate case-markers from postpositions, which have similar distributions. This also abstracts away from cases where certain predicates can mark objects with nominative *ga*, e.g., when the subject is marked with dative case. How Japanese learners find their way around these complications is an important question that we leave for future work.

in the previous simulations, here where the learning problem is more complex, a learner may have motivation to ignore them. However, it is an empirical question how many sentences Japanese-learning children use for acquiring word order and case-marking, and therefore whether they might observe enough evidence for the distributions of case-marked noun-phrases without filtering their data in this manner. It is also an open empirical question whether word order and case-marking are acquired at the same time by Japanese learners. To the extent that our learner successfully infers that Japanese is SOV in this simulation, this shows us that this inference is possible in two settings: first, on the basis of the "unfiltered" data in Section 4, and second, on the basis of data that are filtered to only include sentences that are relevant for also learning case-marking.

From the distributions in Table 3, we observe that 55% of the strings in this learner's dataset 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.

5.3 Results

5.3.1 Noisy CFG learner. Figure 18 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 compare our approach to the numerical regularization approach, we again constructed a comparison learner with a "fully-flexible" hypothesis space, consisting 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

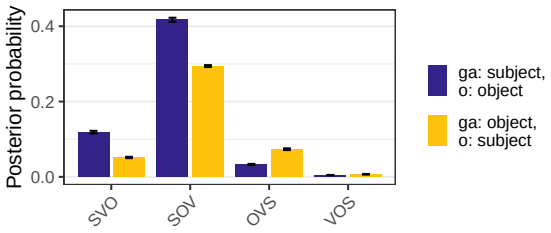


Figure 18. Posterior distribution over case-marking and word-order grammars, Noisy CFG learner

arguments and adjuncts. This is in contrast to our learner's hypothesis space, where the core rules for nominative/accusative case-marking encode the substantive expectation that *ga* and o each canonically mark a different argument, one marking NPS and the other marking NPO. The fully-flexible learner does not encode a distinction between canonical rules for case-marking and non-canonical noise processes, so 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 learner: one assumes that all rules are equally probable *a priori*, and one has a strong bias to use a single expansion for a given nonterminal (α in the Dirichlet prior set to 0.0001). For the learner to match our model's success, its posterior distribution would need to place 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 were parsed as NPS vs. NPO, and the proportion of np-o that were parsed as NPS vs. NPO (Figure 19a). 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.

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

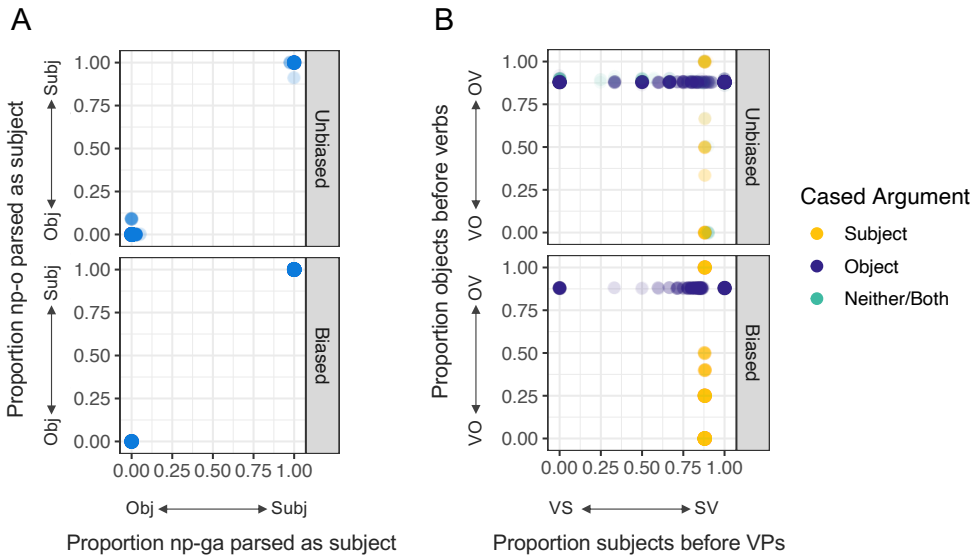


Figure 19. Posterior distribution in sampled treesets ($\vec{t}$) over (A) np-ga and np-o and (B) subject and object position, fully-flexible learner

system of Japanese. The learner's posterior distribution over case-marked np's lies in two different corners: analyses where np-ga and np-o both realize subjects (upper right corner), or where they both realize objects (lower left corner). Both the biased and unbiased behaved similarly, with the biased learner converging more strongly to these 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 roughly 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 $t < 2.04$, $p > 0.07$). Importantly, the learner did not treat np-ga and np-o differently in terms of the clause arguments that they realize; instead, it treated both as markers of the same argument.

Thus, while the learner's numerical regularization bias pushed it more strongly towards certain deterministic corners of its hypothesis space, the pattern that emerges is not a case-marking system. Instead, the learner appeared to infer that np-ga and np-o are members of the same category, with the label of that category either NPS or NPO. One possible

explanation might come from the fact that np-ga and np-o distribute similarly in the data, both tending to occur immediately to the left of v. These distributional similarities may have led the learner to categorize both of these case-marked np's together. Using NPS or NPO as the label for this category would then capture another pattern in the data: case-marked noun-phrases only occur once per sentence, with no strings containing both np-ga and np-o. Because the learner's rules allow only one NPS and one NPO to occur per derivation, treating one of these as the label of the "case-marked np" category provides a good account for why only one case-marked np occurs per string. This property of the data, which other learners may consider accidental, is one that the fully-flexible learner treats as meaningful.[15]

We then assessed the fully-flexible learner's inference about word order by performing the same analysis described in Section 4.3.3 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 19b. Compared to the fully-flexible Japanese learner in Section 4.3.3, 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 is SOV, we would expect to see most samples in 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 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 4.3.3. However, when

[15] 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.

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 occur before the verb.

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

5.4 Summary

We show that our approach of choosing among Noisy CFGs generalizes to another more complex learning problem in the acquisition of natural language morphosyntax: learning canonical word order along with which of two candidate case-markers marks subjects and which marks objects, from data that provide noisy and sparse evidence. We test our learner on distributions of verbs and noun phrases affixed with *ga* and *o* in child-directed Japanese. Japanese allows clause arguments and case-markers to be dropped, introducing a large amount of noise into the data. Our model nonetheless succeeds at separating out noise from signal for the correct grammar of nominative/accusative case-marking, along with canonical SOV word order. By contrast, a learner with a numerical regularization bias falls short: it does not identify a case-marking system, where *ga* and *o* mark different clause arguments.

To explain why our model succeeds in this learning problem, where the general regularization bias approach does not, we can again consider the two properties distinguishing 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, arising 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 thus 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 might in principle manage to draw systematic generalizations from messy data in the incremental 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 at the current stage of development, and other independent processes that conspire to introduce divergences from those target grammatical processes into the data. We model this inference as a choice among different *Noisy Grammars*: 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 likely generated by 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; K. Smith & Wonnacott, 2010). We compare our learner to a common computational implementation of this general regularization bias approach (Reali & Griffiths, 2009; Perfors, 2012), and show that both are able to account for results from a representative artificial language learning experiment (Austin et al., 2022). However, only our learner succeeds 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: that basic clauses require subjects (argument-drop is a non-canonical process), noun-phrases are arguments of basic clauses (adjunction is non-canonical), and in a nominative/accusative case-marking system, case-markers each canonically mark a different clause argument. A learner with a regularization bias operating over a fully-flexible hypothesis space, and no expectation that it is acquiring a grammar that is restrictive in particular ways, does not 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 expects to encounter, which is important for success in the learning problems that we model. In these case studies, 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 thus does not favor any particular extreme within the gradient hypothesis space. However, another logical possibility is that a learner's domain-specific knowledge (such as a preference for clauses to have subjects) could be expressed as a numerical bias that is asymmetrical, preferring certain extremes over others depending on the learning problem at hand.[16] That is, here we contrast an approach to learning with a discrete hypothesis space and rich domain-specific knowledge against an approach with a gradient hypothesis space and few domain-specific expectations, but these are not the only two possibilities in the space of theoretical options. We leave exploration of these issues to future work.

There are two sorts of questions that are raised by the particular domain-specific expectations that we posit here. One is whether these expectations about the structure of basic clauses can be empirically supported. This invites further empirical work investigating whether children possess these expectations at the stage of development that we model here. The second question is where these particular expectations might come from. One possibility is that they are part of the learner's initial knowledge about the forms that grammars can take (Universal Grammar). If this is the case, it would help explain how children manage to draw the correct word-order and case-marking inferences despite the opacity in their data. It would also be in line with the tendencies of grammars to have these properties cross-linguistically (e.g., Baker, 2001; Keenan, 1976; Chomsky, 1982). However, it is also possible that these expectations could reflect linguistic knowledge that has been acquired at some earlier stage of development. If so, this would require an account for how children could

[16] In particular, instead of adopting a symmetrical Dirichlet prior with a single α shape parameter, it is possible to encode asymmetrical preferences through a more general parameterization with distinct α values for each of the possible outcomes in the associated categorical distribution. A domain-specific expectation that subjects are obligatory could plausibly be encoded with an asymmetrical prior that prefers probabilities near 1 (over those near 0) for the rules S $\rightarrow$ NP VP and S $\rightarrow$ VP NP, and prefers probabilities near 0 for S $\rightarrow$ VP. But it is less obvious how this mechanism could express our learner's domain-specific assumptions about case-marking in Section 5, due to the interdependence between different nonterminals' rewrite options: the prior distribution would need to prefer points in the gradient hypothesis space that assign probabilities near 1 to both NPS $\rightarrow$ np-ga and NPO $\rightarrow$ np-o, and similarly for NPS $\rightarrow$ np-o and NPO $\rightarrow$ np-ga, while dispreferring points with high probabilities for both NPS $\rightarrow$ np-ga and NPO $\rightarrow$ np-ga, for example.

arrive at this knowledge about the nature of clauses before identifying the position of clause arguments. The results that we have reported here do not distinguish these possibilities.

In contrasting our approach with the numerical regularization approach, 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 more data are learned from, the prior regularization bias plays less of a role. Some previous accounts have 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 account, we see that limitations to the learner's intake are not necessary in the same way. Our learner identifies the correct canonical grammar with strikingly small 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.

The Noisy Grammar architecture can straightforwardly be applied to deal with more complex grammatical phenomena than we have addressed here, by examining learning within a hypothesis space where the candidate core rules can express those phenomena. We first introduced the architecture in Section 3.2.3 with a sample hypothesis space where the possible core rulesets differed only in whether objects precede the verb or follow it; in Section 4 we crossed this with a further two-way choice of whether subjects precede or follow the VP, presenting the learner with four possible basic word orders (SOV, SVO, OVS and VOS). This hypothesis space does not make available the two other logically possible orders, VSO and OSV, and would therefore have to be extended to investigate word order acquisition in a language like Irish, which is VSO. This could be done by crossing the OV/VO choice

with a three-way choice that includes the option of allowing the verb and object to be discontinuous and surround the subject, rather than occurring only to its left or to its right.

A further direction to explore is the acquisition of word order in a language like German, where the core word order is SOV, but is only observed in embedded clauses due to the "verb second" (V2) property of main clauses (e.g. Thiersch, 1978). The learning difficulties posed by these interacting mechanisms have been discussed by Gibson & Wexler, 1994, for example. To explore this scenario, we could modify the hypothesis space in Figure 11 to include rules allowing embedded clauses (e.g. VP $\rightarrow$ V S) in all four core rulesets, and then cross the four-way choice of word order with an additional two-way choice of V2 (SVO-with-V2, SVO-without-V2, SOV-with-V2, etc.). The distinctive core grammatical rules for V2 would displace elements in ways formally analogous to those that would allow a verb and its object to be separated in VSO and OSV word orders. These sorts of choices of what to put in a learner's hypothesis space reflect theoretical commitments about the content of the inductive bias that the learner brings to this particular stage of acquisition, raising the questions that we noted above about the source of this inductive bias.

These are just two of many phenomena that might require moving from the simple context-free grammar rules in this paper, to a more linguistically expressive formalism. The Noisy Grammar framework that we present, and the inference that our learner performs, is compatible with a wide range of grammar formalisms — specifically, 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. Analyses of discontinuous verb phrases and V2 word order from the contemporary syntax literature could both be formalized using "mildly context-sensitive" grammars that share the relevant mathematical characteristics with CFGs (e.g. Joshi, 1985; D. Weir, 1988; Stabler, 2011). With any 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 rules), and (ii) that these noise rules are a superset of the union

of all of the core rules in the learner's hypothesis space. It is these relationships between the core rules and the noise rules, not the content or specific format of the rules themselves, that are crucial to the inferential logic that we have emphasized in this paper.

These future directions point towards ways in which our approach could be applied to model the incremental growth of grammatical knowledge over the course of a child's development. The developmental stage that we model in our case studies is one in which a child only attempts to acquire the basic positions of clause arguments, without yet attempting to acquire the specific grammatical processes through which those arguments may be displaced in "non-basic" constructions, such as *wh*-questions, relative clauses, and passives. This is motivated by empirical evidence that, at least in English, children acquire these phenomena sequentially: they first identify the canonical positions of subjects and objects, and only later identify the forms that argument movement can take (Gagliardi et al., 2016; Hirsh-Pasek & Golinkoff, 1996; Lidz et al., 2017; Perkins & Lidz, 2021, 2020). In the framework that we put forward, a learner who has identified one of the current word-order hypotheses with some degree of certainty may then treat that knowledge as fixed, and proceed to consider other hypotheses about core rules to account for the data that had been considered "noise" for the purposes of learning word order. If such a learner settles on, say, SVO as its basic word order, it might discard the other three options in Figure 11 and move on to consider a new handful of candidate core rulesets that extend the chosen SVO basic word order with various different displacement possibilities. We pursue this in ongoing work.

This is one example of how our approach might account for the distinct stages of learning in a child's process of growing a grammar. This can be modeled as a process of introducing more grammatical rules into learners' core rulesets, allowing them to understand the processes that were previously considered "noise," with the result that less of their data is treated as noise over time. At no developmental stage do these core rules need to be fully deterministic; a learner may consider that these include variable processes that should not be "regularized away." Our framework provides a way for learners to evaluate how evidence for

any grammatical phenomenon to be acquired, whether deterministic or not, can be separated out from phenomena that are yet to be acquired at a particular stage of development.

In providing a way to explicitly account for the development of grammatical knowledge, our approach differs from other models of grammar acquisition, including those that also lean on the noise-tolerance of Bayesian inference to address ambiguity and messiness in the data (Abend, Kwiatkowski, Smith, Goldwater, & Steedman, 2017; Kwiatkowski, Goldwater, Zettlemoyer, & Steedman, 2012; Perfors et al., 2011; Y. Yang & Piantadosi, 2022). On these previous approaches, grammars are learned wholesale: a learner considers a grammar that is complete even in its initial form, in the sense that it explicitly encodes all of the interacting grammatical processes that could give rise to the data in any language. Learning involves identifying which rules have probability close to zero for the learner's language, similar to the process taken by our "fully-flexible" learners. The problem of being potentially misled by opaque interactions among grammatical phenomena is thus sidestepped by having the learner acquire all of these phenomena jointly. Many other learning models take joint-learning approaches to various degrees (e.g., Elsner, Goldwater, Feldman, & Wood, 2013; Feldman, Griffiths, Goldwater, & Morgan, 2013; Gibson & Wexler, 1994; Howitt et al., 2021; Sakas & Fodor, 2001, 2012; Niyogi & Berwick, 1996; C. Yang, 2002), and it is possible that some amount of joint learning is useful in certain domains. But the empirically-attested trajectory of word order acquisition does not sit well with full joint learning, and points towards the need for a different sort of mechanism in this domain: one that can model how pieces of a grammar are acquired incrementally, from representations of data that are immature and potentially misleading given the learner's current grammatical knowledge. Our approach provides a novel framework for characterizing this type of incremental learning.

Finally, returning to the two approaches to regularization that we compare in this paper, these relate more broadly 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; C. 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 children 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 Acknowledgments

We thank Xinyue Cui, Shalinee Maitra, and Hanyu Zhou for their assistance with testing the code for these simulations. We also thank Naomi Feldman, Avni Gulrajani, Jeff Lidz, Shalinee Maitra, Lisa Pearl, the UCLA Psycholinguistics/Computational Linguistics Seminar, the audiences at BUCLD 2022 and SCiL 2023, and three anonymous reviewers for helpful feedback and discussions on earlier versions of this work.

8 References

Abend, O., Kwiatkowski, T., Smith, N. J., Goldwater, S., & Steedman, M. (2017). Bootstrapping language acquisition. *Cognition*, *164*, 116–143.

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.

Baker, M. C. (2001). The Natures of Nonconfigurationality. In R. K. Baltin & C. Collins (Eds.), *The Handbook of Contemporary Syntactic Theory* (pp. 407–438). Malden, Mass: Blackwell Publishers Ltd.

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.

Chemla, E., Mintz, T. H., Bernal, S., & Christophe, A. (2009). Categorizing words using 'frequent frames': what cross-linguistic analyses reveal about distributional acquisition strategies. *Developmental science*, *12*(3), 396–406. (Publisher: Wiley Online Library)

Chomsky, N. (1957). *Syntactic Structures*. The Hague: Mouton.

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

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

Chomsky, N. (1980). *Rules and representations*. New York: Columbia University Press.

Chomsky, N. (1982). *Some concepts and consequences of the theory of government and binding*. MIT Press.

Chomsky, N. (1986). *Knowledge of language: Its nature, origins and use*. New York: Praeger.

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.

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.

Dye, C., Kedar, Y., & Lust, B. (2019, February). From lexical to functional categories: New foundations for the study of language development. *First Language*, *39*(1), 9–32. doi: 10.1177/0142723718809175

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.

Elsner, M., Goldwater, S., Feldman, N., & Wood, F. (2013). A joint learning model of word segmentation, lexical acquisition, and phonetic variability. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (pp. 42–54). Association for Computational Linguistics.

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.

Feldman, N. H., Griffiths, T. L., Goldwater, S., & Morgan, J. L. (2013). A role for the developing lexicon in phonetic category acquisition. *Psychological review*, *120*(4), 751. (Publisher: American Psychological Association)

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.

Franck, J., Millotte, S., Posada, A., & Rizzi, L. (2013). Abstract knowledge of word order by 19 months: An eye-tracking study. *Applied Psycholinguistics*, *34*(2), 323–336.

Frank, R., & Kapur, S. (1996). On the use of triggers in parameter setting. *Linguistic Inquiry*, 623–660.

Gagliardi, A., Mease, T. M., & Lidz, J. (2016). Discontinuous development in the acquisition of filler-gap dependencies: Evidence from 15-and 20-month-olds. *Language Acquisition*, *23*(3), 1–27.

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.

Gavarró, A., Leela, M., Rizzi, L., & Franck, J. (2015). Knowledge of the OV parameter setting at 19 months: Evidence from Hindi–Urdu. *Lingua*, *154*, 27–34.

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)

Gibson, E., & Wexler, K. (1994). Triggers. *Linguistic Inquiry*, *25*, 407–454.

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

Greenberg, J. H. (1970). *Language Universals*. De Gruyter.

Gutman, A., Dautriche, I., Crabbé, B., & Christophe, A. (2015). Bootstrapping the syntactic bootstrapper: probabilistic labeling of prosodic phrases. *Language Acquisition*, *22*(3), 285–309.

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.

Haryu, E., & Kajikawa, S. (2016). Use of bound morphemes (noun particles) in word segmentation by Japanese-learning infants. *Journal of Memory and Language*, *88*, 18–27. (Publisher: Elsevier)

He, A. X., & Lidz, J. (2017). Verb learning in 14-and 18-month-old English-learning infants. *Language Learning and Development*, 1–22.

Hicks, J., Maye, J., & Lidz, J. (2007). The role of function words in infants' syntactic categorization of novel words. *Linguistic Society of America Annual Meeting*.

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.

Howitt, K., Dey, S., & Sakas, W. G. (2021, January). Gradual syntactic triggering: The gradient parameter hypothesis. *Language Acquisition*, *28*(1), 65–96.

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.

Hyams, N. (1986). *Language Acquisition and the Theory of Parameters*. Dordrecht: D. Reidel Publishing Company.

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.

Jin, K.-S., & Fisher, C. (2014). Early evidence for syntactic bootstrapping: 15-month-olds use sentence structure in verb learning. In *Proceedings of the 38th Boston University Conference on Language Development.* Boston, MA: Cascadilla Press.

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.

Joshi, A. (1985). How much context-sensitivity is necessary for characterizing structural descriptions? In D. Dowty, L. Karttunen, & A. Zwicky (Eds.), *Natural language processing: Theoretical, computational and psychological perspectives* (pp. 206–250). New York: Cambridge University Press.

Jusczyk, P. W., & Aslin, R. N. (1995). Infants' detection of the sound patterns of words in fluent speech. *Cognitive psychology*, *29*(1), 1–23.

Keenan, E. L. (1976). Towards a Universal Definition of 'Subject'. In C. Li (Ed.), *Syntax and Semantics: Subject and Topic.* New York: Academic Press.

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.

Kwiatkowski, T., Goldwater, S., Zettlemoyer, L., & Steedman, M. (2012). A probabilistic model of syntactic and semantic acquisition from child-directed utterances and their meanings. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics* (pp. 234–244).

Labov, W. (1989). The child as linguistic historian. *Language variation and change*, *1*(1), 85–97. (Publisher: Cambridge University Press)

Lewis, J. D., & Elman, J. L. (2001). Learnability and the statistical structure of language: Poverty of stimulus arguments revisited. In *Proceedings of the 26th annual Boston University conference on language development* (Vol. 1, pp. 359–370). Citeseer.

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.

Manzini, M. R., & Wexler, K. (1987). Parameters, Binding Theory, and Learnability. *Linguistic Inquiry*, *18*(3), 413–444.

Marquis, A., & Shi, R. (2012). Initial morphological learning in preverbal infants. *Cognition*, *122*(1), 61–66. (Publisher: Elsevier)

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.

McCloskey, J. (1997). Subjecthood and subject positions. In L. Haegeman (Ed.), *Elements of grammar: Handbook of generative syntax* (pp. 197–235). Springer.

Miller, K. (2013). Acquisition of variable rules:/s/-lenition in the speech of Chilean Spanish-speaking children and their caregivers. *Language variation and change*, *25*(3), 311–340. (Publisher: Cambridge University Press)

Miller, K., & Schmitt, C. (2012). Variable input and the acquisition of plural morphology. *Language Acquisition*, *19*(3), 223–261. (Publisher: Taylor & Francis)

Mintz, T. H. (2003). Frequent frames as a cue for grammatical categories in child directed speech. *Cognition*, *90*(1), 91–117.

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.

Niyogi, P., & Berwick, R. C. (1996). A language learning model for finite parameter spaces.

Cognition, *61*(1-2), 161–193. (Publisher: Elsevier)

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.

Perfors, A., Tenenbaum, J. B., & Regier, T. (2011). The learnability of abstract syntactic principles. *Cognition*, *118*(3), 306–338. (Publisher: Elsevier)

Perfors, A., Tenenbaum, J. B., & Wonnacott, E. (2010). Variability, negative evidence, and the acquisition of verb argument constructions. *Journal of child language*, *37*(3), 607–642.

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., & Christiansen, M. H. (2005). Uncovering the richness of the stimulus: Structure dependence and indirect statistical evidence. *Cognitive Science*, *29*(6), 1007–1028. (Publisher: Wiley Online Library)

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

Roberts, J. (1997). Acquisition of variable rules: a study of (-t, d) deletion in preschool children. *Journal of child language*, *24*(2), 351–372. (Publisher: Cambridge University Press)

Roberts, J., & Labov, W. (1995). Learning to talk Philadelphian: Acquisition of short a by preschool children. *Language Variation & Change*, *7*(1).

Sagae, K. (2021). Tracking child language development with neural network language models. *Frontiers in Psychology*, *12*, 674402. (Publisher: Frontiers Media SA)

Sakas, W. G., & Fodor, J. D. (2001). The structural triggers learner. In S. Bertolo (Ed.), *Language acquisition and learnability* (pp. 172–233). Cambridge: Cambridge University Press. (Publisher:)

Sakas, W. G., & Fodor, J. D. (2012). Disambiguating syntactic triggers. *Language Acquisition*, *19*(2), 83–143. (Publisher: Taylor & Francis)

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.

Seki, H., Matsumara, T., Fujii, M., & Kasami, T. (1991). On multiple context-free grammars. *Theoretical Computer Science*, *88*, 191–229.

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.

Shin, N., & Miller, K. (2022). Children's Acquisition of Morphosyntactic Variation. *Language Learning and Development*, *18*(2), 125–150.

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, J., & Durham, M. (2019). *Sociolinguistic variation in children's language: Acquiring community norms*. Cambridge University Press.

Smith, J., Durham, M., & Richards, H. (2013, January). The social and linguistic in the acquisition of sociolinguistic norms: Caregivers, children, and variation. *Linguistics*, *51*(2).

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.

Song, J. Y., Shattuck-Hufnagel, S., & Demuth, K. (2015). Development of phonetic variants (allophones) in 2-year-olds learning American English: A study of alveolar stop/t, d/codas. *Journal of Phonetics*, *52*, 152–169. (Publisher: Elsevier)

Stabler, E. P. (2004). Varieties of crossing dependencies: structure dependence and mild context sensitivity. *Cognitive Science*, *28*, 699–720.

Stabler, E. P. (2011). Computational perspectives on minimalism. In C. Boeckx (Ed.), *The oxford handbook of linguistic minimalism.* Oxford: Oxford University Press.

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.

Thiersch, C. L. (1978). *Topics in german syntax* (Unpublished doctoral dissertation). MIT.

Torrego, E. (1989). Unergative-unaccusative alternations in Spanish. *MIT Working Papers in Linguistics*, *10*, 253–272.

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, D. (1988). *Characterizing mildly context-sensitive grammar formalisms* (Unpublished doctoral dissertation). University of Pennsylvania.

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.

Xu, F., & Tenenbaum, J. B. (2007). Word learning as Bayesian inference. *Psychological review*, *114*(2), 245.

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

Yang, Y., & Piantadosi, S. T. (2022, February). One model for the learning of language. *Proceedings of the National Academy of Sciences*, *119*(5), e2021865119.

Zhu, J., Franck, J., Rizzi, L., & Gavarró, A. (2022). Do infants have abstract grammatical knowledge of word order at 17 months? Evidence from Mandarin Chinese. *Journal of*

Child Language, *49*(1), 60–79.

Appendix A

Details of the likelihood calculations in Section 3

8.1 Likelihoods under bags of coins

Recall the scenario with just G_{ka}: this bag contains an unknown number of Φ-coins, which always produce 'ka', and an unknown number of Ψ-coins, which all have some single unknown probability ψ of producing 'ka'. Ten times, a coin is chosen from the bag and flipped; this produces eight 'ka' and two 'bo'. How many of these ten flips might we guess came from Φ-coins, and how many from Ψ-coins?

In the main text we contrasted two hypotheses: one where there are six Φ-flips and four Ψ-flips ($N_\Phi = 6$), and one where there are zero Φ-flips and ten Ψ-flips ($N_\Phi = 0$). Conditioned upon the unknown probability ψ, the likelihood of the $N_\Phi = 6$ hypothesis is $\binom{4}{2}\psi^2(1-\psi)^2$, and the likelihood of the $N_\Phi = 0$ hypothesis is $\binom{10}{8}\psi^8(1-\psi)^2$. To compare the likelihoods not conditioned on ψ, we can marginalize over this parameter, as in (6) and (7).

$$P(\tbinom{8\ \text{'ka'}}{2\ \text{'bo'}} \mid N_\Phi = 0, G_{\text{ka}}) = \int_0^1 P(\tbinom{8\ \text{'ka'}}{2\ \text{'bo'}} \mid N_\Phi = 0, G_{\text{ka}}, \psi)\ p(\psi)\ d\psi = \int_0^1 \binom{10}{8}\ \psi^8\ (1-\psi)^2\ p(\psi)\ d\psi \tag{6}$$

$$P(\tbinom{8\ \text{'ka'}}{2\ \text{'bo'}} \mid N_\Phi = 6, G_{\text{ka}}) = \int_0^1 P(\tbinom{8\ \text{'ka'}}{2\ \text{'bo'}} \mid N_\Phi = 6, G_{\text{ka}}, \psi)\ p(\psi)\ d\psi = \int_0^1 \binom{4}{2}\ \psi^2\ (1-\psi)^2\ p(\psi)\ d\psi \tag{7}$$

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 those values. On a numerical regularization approach, such a belief could in principle be expressed by assuming that $p(\psi)$ takes the form of a skewed Beta distribution. 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 assumption of a uniform Beta prior makes available particularly simple closed-form solutions for the integrals in (6) and (7). Specifically, for any m and k:

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

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. If we hypothesize n_Φ flips of core Φ-coins, then we must invoke $10 - n_\Phi$ flips of noise Ψ-coins, 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 $\frac{1}{3}$. But this is the highest we can go: if there are only eight occurrences of 'ka' in the data, then the Φ-coins could not have been flipped more than eight times. The cap on the number of Φ-flips is even lower under G_{bo}: here the Φ-coins produce 'bo', so the largest number of Φ-flips is two.

$$P(\begin{smallmatrix}8\text{ 'ka'}\\2\text{ 'bo'}\end{smallmatrix} \mid N_\Phi = n_\Phi, G_{\text{ka}}) = \begin{cases} \dfrac{1}{11-n_\Phi} & \text{if } n_\Phi \leq 8 \\ 0 & \text{otherwise} \end{cases} \tag{9}$$

$$P(\begin{smallmatrix}8\text{ 'ka'}\\2\text{ 'bo'}\end{smallmatrix} \mid N_\Phi = n_\Phi, G_{\text{bo}}) = \begin{cases} \dfrac{1}{11-n_\Phi} & \text{if } n_\Phi \leq 2 \\ 0 & \text{otherwise} \end{cases} \tag{10}$$

To compare the likelihoods of G_{ka} and G_{bo} not conditioned on a particular choice of n_Φ, we can marginalize over all 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:

$$(11)\qquad P(\,{}^{8}_{2}\,\text{`ka'}_{\text{`bo'}} \mid G_{\text{ka}}) = \sum_{n_\Phi=0}^{10}\Big[P(\,{}^{8}_{2}\,\text{`ka'}_{\text{`bo'}} \mid N_\Phi = n_\Phi, G_{\text{ka}}) \times P(N_\Phi = n_\Phi \mid G_{\text{ka}})\Big]$$
$$= \sum_{n_\Phi=0}^{8}\left[\frac{1}{11-n_\Phi} \times P(N_\Phi = n_\Phi \mid G_{\text{ka}})\right]$$

$$(12)\qquad P(\,{}^{8}_{2}\,\text{`ka'}_{\text{`bo'}} \mid G_{\text{bo}}) = \sum_{n_\Phi=0}^{2}\left[\frac{1}{11-n_\Phi} \times P(N_\Phi = n_\Phi \mid G_{\text{bo}})\right]$$

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 4 in the main text, following the logic from (8).[17]

$$(13)\qquad P(\,{}^{8}_{2}\,\text{`ka'}_{\text{`bo'}} \mid G_{\text{ka}}) = \sum_{n_\Phi=0}^{8}\left[\frac{1}{11-n_\Phi} \times \frac{1}{11}\right] = 0.138$$

$$(14)\qquad P(\,{}^{8}_{2}\,\text{`ka'}_{\text{`bo'}} \mid G_{\text{bo}}) = \sum_{n_\Phi=0}^{2}\left[\frac{1}{11-n_\Phi} \times \frac{1}{11}\right] = 0.027$$

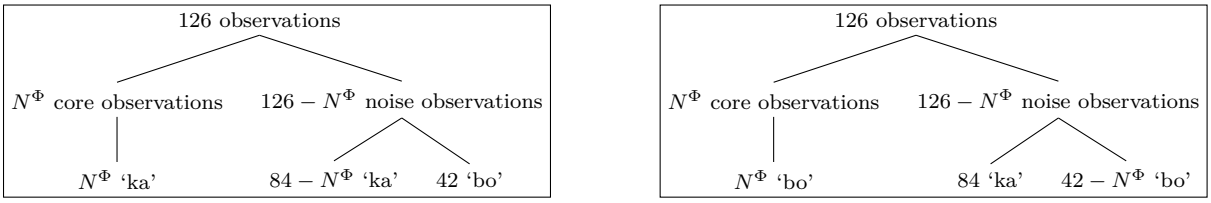
Given the likelihoods in (13) and (14) and a prior probability for each bag, we can apply Bayes' Rule to calculate each bag's posterior probability. Assuming an equal prior of 0.5 for each bag, we find that G_{ka} is about five times more likely to have generated the data than G_{bo}.

$$(15)\qquad P(G_{\text{ka}} \mid \,{}^{8}_{2}\,\text{`ka'}_{\text{`bo'}}) = \frac{P(\,{}^{8}_{2}\,\text{`ka'}_{\text{`bo'}} \mid G_{\text{ka}})P(G_{\text{ka}})}{P(\,{}^{8}_{2}\,\text{`ka'}_{\text{`bo'}} \mid G_{\text{ka}})P(G_{\text{ka}}) + P(\,{}^{8}_{2}\,\text{`ka'}_{\text{`bo'}} \mid G_{\text{bo}})P(G_{\text{bo}})}$$
$$= \frac{0.138 \times 0.5}{0.138 \times 0.5 + 0.027 \times 0.5}$$
$$= 0.834$$

$$P(G_{\text{bo}} \mid \,{}^{8}_{2}\,\text{`ka'}_{\text{`bo'}}) = 0.166$$

To extend this to the actual training data from the Austin et al. (2022) experiment, Figure A1 illustrates the ways in which the 84 observations of 'ka' and 42 of 'bo' could break down under each hypothesis, where again N^Φ is the number of observations for which the

[17] 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 4 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 G_{bo}'s total larger than G_{ka}'s.



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

Figure A1. Partitioning 84 'ka' and 42 'bo' into core and noise

core mechanisms were responsible. Then following (13) and (14), the likelihood the data under one of these hypotheses is a sum over the possible values of N^Φ, where each contribution to the sum represents a particular choice among 127 ways to split the total determiner observations, and among $(127 - n^\Phi)$ ways to split the noise observations. The two hypotheses differ only in the range of contributing values in the summation.

$$P(\begin{smallmatrix}84\text{ 'ka'}\\42\text{ 'bo'}\end{smallmatrix} \mid G_{\text{ka}}) = \sum_{n^\Phi=0}^{126} \left[P(\begin{smallmatrix}84\text{ 'ka'}\\42\text{ 'bo'}\end{smallmatrix} \mid N^\Phi = n^\Phi, G_{\text{ka}}) \times P(N^\Phi = n^\Phi \mid G_{\text{ka}}) \right] \tag{16}$$

$$= \sum_{n^\Phi=0}^{84} \left[\frac{1}{127 - n^\Phi} \times \frac{1}{127} \right] = 8.65 \times 10^{-3}$$

$$P(\begin{smallmatrix}84\text{ 'ka'}\\42\text{ 'bo'}\end{smallmatrix} \mid G_{\text{bo}}) = \sum_{n^\Phi=0}^{126} \left[P(\begin{smallmatrix}84\text{ 'ka'}\\42\text{ 'bo'}\end{smallmatrix} \mid N^\Phi = n^\Phi, G_{\text{bo}}) \times P(N^\Phi = n^\Phi \mid G_{\text{bo}}) \right] \tag{17}$$

$$= \sum_{n^\Phi=0}^{42} \left[\frac{1}{127 - n^\Phi} \times \frac{1}{127} \right] = 3.24 \times 10^{-3}$$

Using these likelihoods, and assuming equal prior probabilities for $P(G_{\text{ka}}) = P(G_{\text{bo}}) = 0.5$, then Bayes' Rule tells us that the posterior probability of G_{ka} is more than twice as high than the posterior probability of G_{bo}.

$$P(G_{\text{ka}} \mid \begin{smallmatrix}84\text{ 'ka'}\\42\text{ 'bo'}\end{smallmatrix}) = \frac{\left(8.65 \times 10^{-3}\right) \times 0.5}{\left(8.65 \times 10^{-3}\right) \times 0.5 + \left(3.24 \times 10^{-3}\right) \times 0.5} \tag{18}$$

$$= 0.728$$

$$P(G_{\text{bo}} \mid \begin{smallmatrix}84\text{ 'ka'}\\42\text{ 'bo'}\end{smallmatrix}) = 0.272$$

8.2 Likelihoods of trees

Considering the collection of trees in the main text Figure 8, let's work out the likelihood of the observed collection of VP rewrites arising via a specific combination of noise and non-noise under the SVO hypothesis— i.e. via specific values $n_1^\Phi \in \{0,1,2\}$ and $n_2^\Phi \in \{0,1,2\}$ in the main text Figure 10a. There are six possible ways that the five VP rewrites could have been split between noise and non-noise, so the particular split that has $n_1^\Phi + n_2^\Phi$ non-noise rewrites has a probability of $\frac{1}{6}$. Similarly, the probability of the $n_1^\Phi + n_2^\Phi$ non-noise rewrites breaking into two groups as they did (on the left of Figure 10a) is $\frac{1}{n_1^\Phi + n_2^\Phi + 1}$. For the breakdown of the noise rewrites (on the right of Figure 10a), 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 $n = 5 - (n_1^\Phi + n_2^\Phi)$ and $k = 3$, this is $\frac{(7-(n_1^\Phi+n_2^\Phi))!}{(5-(n_1^\Phi+n_2^\Phi))!2!}$. The overall likelihood is therefore

$$P\left(\begin{smallmatrix}2\ \text{VP}\to\text{V NP}\\ 1\ \text{VP}\to\text{NP V}\\ 2\ \ \text{VP}\to\text{V}\end{smallmatrix}, \begin{smallmatrix}N_1^\Phi=n_1^\Phi\\ N_2^\Phi=n_2^\Phi\end{smallmatrix} \,\middle|\, G_{\text{SVO}}\right) = \frac{1}{6} \times \frac{1}{n_1^\Phi + n_2^\Phi + 1} \times \frac{(5-(n_1^\Phi+n_2^\Phi))!2!}{(7-(n_1^\Phi+n_2^\Phi))!} \tag{19}$$

and from here we can marginalize over the unknown values of n_1^Φ and n_2^Φ to find a likelihood conditioned only on the choice of G_{SVO}.

$$\begin{aligned} P\left(\begin{smallmatrix}2\ \text{VP}\to\text{V NP}\\ 1\ \text{VP}\to\text{NP V}\\ 2\ \ \text{VP}\to\text{V}\end{smallmatrix} \,\middle|\, G_{\text{SVO}}\right) &= \sum_{n_1^\Phi=0}^{2} \sum_{n_2^\Phi=0}^{2} \left[\frac{1}{6} \times \frac{1}{n_1^\Phi + n_2^\Phi + 1} \times \frac{(5-(n_1^\Phi+n_2^\Phi))!2!}{(7-(n_1^\Phi+n_2^\Phi))!}\right] \\ &= 6.07 \times 10^{-2} \end{aligned} \tag{20}$$

The three fractions inside the summation in (20) correspond to the three branch-points in Figure 10a, just as the fractions inside the summations in (16) and (17) correspond to the branch-points in Figure A1.

The situation for G_{SOV} is similar, except that n_1^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 (16) and (17), this distinction has no effect on the fractions inside the summation, but constrains the range of values to sum over for n_1^Φ: rather than the cap of 2 in (20), for the SOV hypothesis it is capped at 1.

$$(21)\qquad P\left(\begin{smallmatrix}2\ \text{VP}\to\text{V NP}\\ 1\ \text{VP}\to\text{NP V}\\ 2\ \ \text{VP}\to\text{V}\end{smallmatrix}\,\middle|\, G_{\text{SOV}}\right) = \sum_{n_1^\Phi=0}^{1}\sum_{n_2^\Phi=0}^{2}\left[\frac{1}{6}\times\frac{1}{n_1^\Phi+n_2^\Phi+1}\times\frac{(5-(n_1^\Phi+n_2^\Phi))!2!}{(7-(n_1^\Phi+n_2^\Phi))!}\right]$$
$$= 3.71\times10^{-2}$$

The probabilities of the S rewrites can be calculated similarly, giving rise to the following likelihoods of the trees in Figure 8:

$$(22)\qquad P(\text{trees in Figure 8}\mid G_{\text{SVO}}) = P\left(\begin{smallmatrix}2\ \text{VP}\to\text{V NP}\\ 1\ \text{VP}\to\text{NP V}\\ 2\ \ \text{VP}\to\text{V}\end{smallmatrix}\,\middle|\, G_{\text{SVO}}\right)\times P\left(\begin{smallmatrix}2\ \ \text{S}\to\text{NP S}\\ 0\ \ \text{S}\to\text{S NP}\\ 3\ \text{S}\to\text{NP VP}\\ 1\ \text{S}\to\text{VP NP}\\ 1\ \ \text{S}\to\text{VP}\end{smallmatrix}\,\middle|\, G_{\text{SVO}}\right)$$
$$= (6.07\times10^{-2})\times(1.45\times10^{-3})$$
$$= 8.82\times10^{-5}$$
$$P(\text{trees in Figure 8}\mid G_{\text{SOV}}) = (3.71\times10^{-2})\times(1.45\times10^{-3})$$
$$= 5.39\times10^{-5}$$

Assuming equal prior probabilities for the two grammars, we find that the posterior probability of G_{SVO} given the trees in Figure 8 is $\frac{8.82}{8.82+5.39} = 0.621$, compared to 0.379 for G_{SOV}.

Appendix B

Details of Gibbs sampling

In the first step of sampling, we use Bayes' Rule to calculate the posterior probability of each grammar given the observed strings $\vec{w}$ and a collection of hypothesized trees $\vec{t}$ for those strings:

$$P(G|\vec{t},\vec{w}) = \frac{P(\vec{t},\vec{w}|G)P(G)}{\sum_{G'} P(\vec{t},\vec{w}|G')P(G')} \tag{23}$$

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 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).[18]

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

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 $n_1^{A^\phi} \dots n_m^{A^\psi}$:

$$(24)\quad P(\vec{t}|G) = \prod_A P(n_1^A \dots n_m^A \mid G) = \prod_A \Bigg[\sum_{n_1^{A^\phi}=0}^{n_1^A} \cdots \sum_{n_m^{A^\phi}=0}^{n_m^A} \Big[P(n_1^{A^\phi} \dots n_m^{A^\phi} | n^{A^\phi}, G) \times P(n_1^{A^\psi} \dots n_m^{A^\psi} | n^{A^\psi}, G) \times P(n^{A^\phi} | n^A, G) \Big] \Bigg]$$

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 $\vec{\phi}^{A_G}$. To represent the prior over $\vec{\phi}^{A_G}$, we use 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, which results in a uniform prior distribution: the model has no preference for or against assigning probability mass to any particular expansions of A. Because $\vec{\phi}^{A_G}$ is unknown, we integrate over all possible values of $\vec{\phi}^{A_G}$ to obtain

$$(25)\qquad \frac{B(\vec{\alpha_\phi}^{A_G} + (n_1^{A^\phi} \dots n_m^{A^\phi}))}{B(\vec{\alpha_\phi}^{A_G})}$$

for this first term, where $\vec{\alpha_\phi}^{A_G}$ represents the parameters of the 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, 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:

$$(26)\qquad \frac{1}{\binom{n^{A^\phi}+m^{A^\phi}-1}{m^{A^\phi}-1}} = \frac{n^{A^\phi}!(m^{A^\phi}-1)!}{(n^{A^\phi}+m^{A^\phi}-1)!}$$

The second term in the sum in (24) is analogous: this is the probability, given n^{A^ψ}

total noisy occurrences of A, of observing $n_1^{A^\psi} \ldots n_m^{A^\psi}$ noisy occurrences of each rewrite type, which follows a multinomial distribution with parameter $\vec{\psi}^{A_G}$. The third term is the probability of observing n^{A^ϕ} total core occurrences out of n^A overall occurrences of A. This follows a binomial distribution with parameter $(1 - \epsilon^{A_G})$. We again integrate over all possible values of $\vec{\psi}^{A_G}$ and ϵ^{A_G} assuming uniform priors over these parameters, obtaining results analogous to (25).

This allows us to calculate the likelihood $P(\vec{t} \mid G)$ for each hypothesized G, and (with 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

$$f(t_i) = P(w_i|t_i)P(t_i|\vec{t}_{-i}, G) \tag{27}$$

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

$$P(t_i|\vec{t}_{-i}, G) = \frac{P(\vec{t}|G)}{P(\vec{t}_{-i}|G)} \tag{28}$$

Both $P(\vec{t}|G)$ and $P(\vec{t}_{-i}|G)$ can be calculated according to to (24). In practice, $P(\vec{t}_{-i}|G)$ does not need to be calculated because it will cancel out in the acceptance function (29), 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 first initialize a new set of trees $\vec{t}$ for the Hastings sampler that we will then use to re-sample $\vec{t}$ given G and $\vec{w}$. We do so by randomly initializing the parameters θ for the compiled-out PCFG given by the current choice of grammar G, then generating full "articulated" trees for each string based on this choice of θ, and finally collapsing the articulated trees to remove the core/noise distinction. Given these initialized

trees, 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 core vs. 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

$$A(t_i') = min\left(1, \frac{f(t_i{}')P(t_i|w_i, \vec{\theta^s})}{f(t_i)P(t_i{}'|w_i, \vec{\theta^s})}\right) \tag{29}$$

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.

Appendix C

Appendix on data pre-processing

The procedure described here is a simple implementation of the idea that an infant has learned to recognize certain classes of functional elements, and can use those functional elements as cues to the positions of nouns and verbs. This idea receives substantial empirical support from findings that functional elements are recognized early in development and are used to classify content words by young infants, at or before the ages that we model here (approximately 15 months); see Table C1. These results primarily come from English, French, and German. We extrapolate from these results in modeling Japanese, noting that much less is known about infants' early word class and word order development in that language, and also noting that the possibility that this development may not proceed within the same timeframe across all of the languages that we model. There are many ways in which the idea of using functional elements to identify noun phrases and verbs could be fleshed out (see e.g., Mintz, 2003; Chemla, Mintz, Bernal, & Christophe, 2009; Gutman et al., 2015), and different proposals will in general make distinct fine-grained predictions about exactly what syntactic information an infant can extract from a given utterance. But at the level of granularity that is relevant for this study, where our goal is to explore what can be learned from messy distributions of strings like those in Table 2 in the main text, we assume that many of these different options would yield roughly similar results.

We begin with CHILDES's tokenization of corpus utterances, treating the "clitics" in the CHILDES annotation as separate tokens (e.g. CHILDES analyzes *you wanna* as `you want~to`, which we treat as three tokens: `you want to`). We also split off as separate tokens apparent occurrences of a small number of hand-identified affixes for 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`.

We then assume that an infant (at around the age of 15 months) can recognize certain functional elements from among these tokens. The full list of such recognizable elements, for

Determiners used to classify nouns	Shi and Melançon (2010); Hicks et al. (2007); Höhle et al. (2004); Babineau et al. (2020); He and Lidz (2017)
Pronouns treated as clause arguments (NPs)	Babineau et al. (2020); Jin and Fisher (2014)
Auxiliaries used to classify verbs	Hicks et al. (2007)
English *-s*, *-ing*, and *-ed* recognized as suffixes	Kim and Sundara (2021); Mintz (2013)
French *-é* recognized as suffix	Marquis and Shi (2012)
Japanese *ga* recognized as bound morpheme	Haryu and Kajikawa (2016)

Table C1
Summary of selected empirical findings suggesting infants' abilities at or prior to 15 months to use functional elements to classify nouns and verbs.

each language, is given in Table C2. All of these recognizable elements were among the 100 most frequent tokens in the corresponding language's corpus. We assume that some of these are recognized by the infant as belonging to some syntactically-meaningful category. In English, for example, we assume that an infant can recognize `you` as PRONOUN; `the` as a member of a functional category that precedes nouns, which we call DET (determiner); `will` as a member of a functional category that precedes verbs, which we call AUX; and `-ed` as a V-SUFFIX (verbal suffix). Others are simply given the "meaningless" category OTHER. Elements with messy or ambiguous distributions (such as `is`, which might be an auxiliary or a copula main verb, and `to`, which might be an auxiliary or a preposition), we tended to classify as OTHER, in order to be conservative in our assumptions about the infant's knowledge. 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's tagging.

Once these recognizable elements have been tagged in the corpus, we use some simple heuristics to guess at the positions of verbs and nouns on the basis of only those tags. For each language, we define a small set of "NOUN-cue" positions, such that any unrecognized element in one of these positions is tagged as a NOUN. For example, the position immediately following a DETERMINER is a NOUN-cue position in English and French. There

are also "VERB-cue" positions for each language. The full list is given in Table C3. With the initial tagging (based on Table C2) now supplemented with NOUN and VERB tags based on functional cues, the final step is to extract a sequence of `np` and `v` tokens. This final sequence includes an `np` wherever something has been tagged as a PRONOUN, NAME or NOUN, and includes a `v` wherever something has been tagged as a VERB.

The full process is illustrated with two English example sentences in Figure C1. Some example sentences that yielded, via this procedure, the most common string types for each language are shown in Table C4.

English	
AUX	do, did, will, can, does, would
DET	the, a, your, some, my, his
NAME	Sarah, Adam, Eve
NEG	not
N-SUFFIX	-s
PRONOUN	you, it, I, he, she, me, we, they, them, him
V-SUFFIX	-ing, -ed
OTHER	is, what, that, to, are, no, and, oh, in, on, yeah, there, this, one, here, yes, where, of, for, with, right, up, how, well, now, why, huh, who, was, all, out, her, alright, be, down, just, at, so, okay, am, hm, if, when, mhm, too, two

French	
AUX	va, faut, vas, ai, vais
DET	le, un, le, les, une, des, du, ton
NAME	maman, Nanou
NEG	pas
PRONOUN	tu, ça, il, je, on, elle, toi, moi, ils, lui
V-SUFFIX	-é(e)(s), -er
OTHER	est, ce, que, là, cm, et, la, de, ah, oui, a, non, hein, y, est-ce, oh, à, te, as, alors, voilà, en, qui, comme, mais, quoi, dans, ben, où, pour, bon, me, allez, la, ouais, se, hop, sur, si, mh, es, bah, mhm, ne, d'accord, sont, au, euh, oui

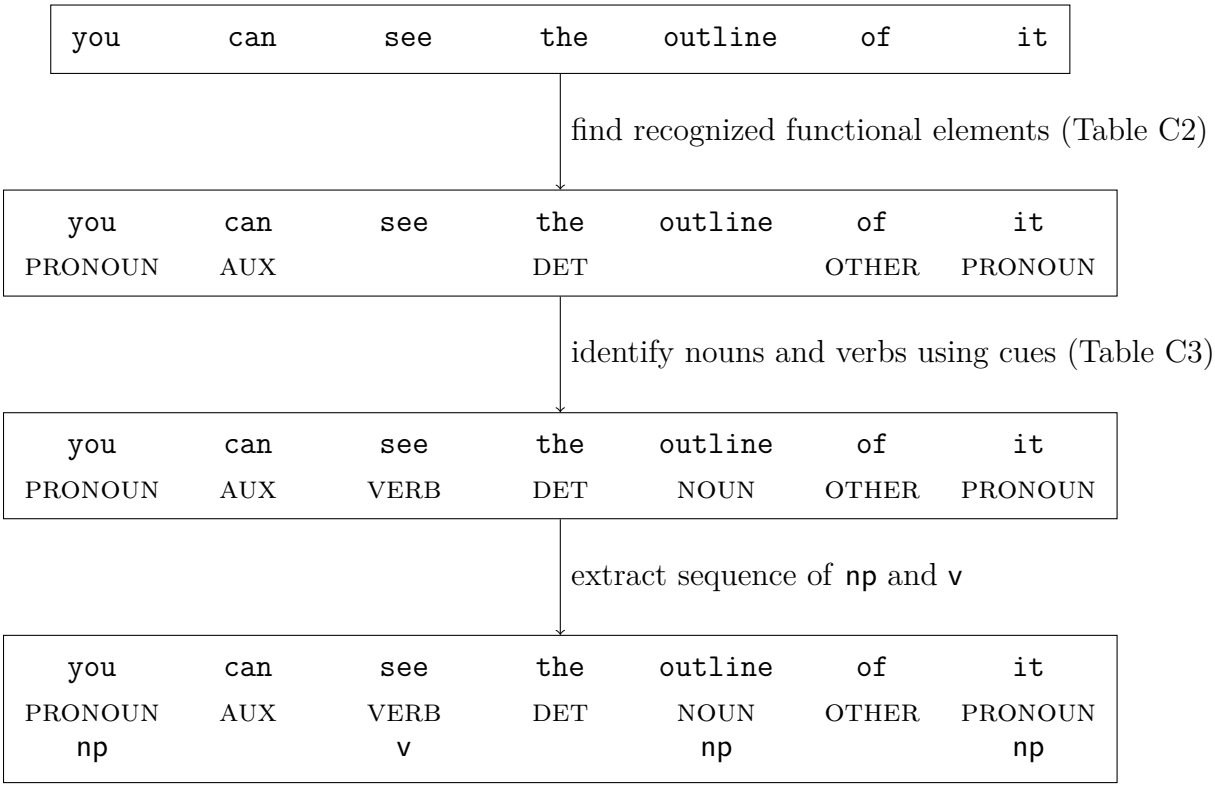
Japanese	
CM	ga, o, wa, no, ni, mo, de, to
NAME	Asatokun, Natchan, Nanami, Kakka, Okaasan, Jurichan, Natchi
NEG	-nai
PRONOUN	kore, koko, sore, kotchi
SFP	ne, yo, ka, kanaa, naa, sa, jan, yan, mon
V-SUFFIX	-te, -de, -ba
OTHER	no, da, nani, tte, ni, kara, na, doko, janai, desu, soo, koo, de, to, ne, deshoo, dore, dare, nande, kono, kedo, doo, ya

Table C2
The elements we assume are recognizable by the learner for each language.

English and French	
NOUN-cue positions	VERB-cue positions
following a DET preceding a N-SUFFIX	following an AUX following a NEG that follows an AUX preceding a V-SUFFIX

Japanese	
NOUN-cue positions	VERB-cue positions
preceding a CM (case-maker)	preceding a NEG preceding a SFP (sentence-final particle) preceding a V-SUFFIX

Table C3
The "cue positions" defined for each language


you can see the outline of it
find recognized functional elements (Table C2)
you can see the outline of it
PRONOUN AUX DET OTHER PRONOUN
identify nouns and verbs using cues (Table C3)
you can see the outline of it
PRONOUN AUX VERB DET NOUN OTHER PRONOUN
extract sequence of np and v
you can see the outline of it
PRONOUN AUX VERB DET NOUN OTHER PRONOUN
np v np np

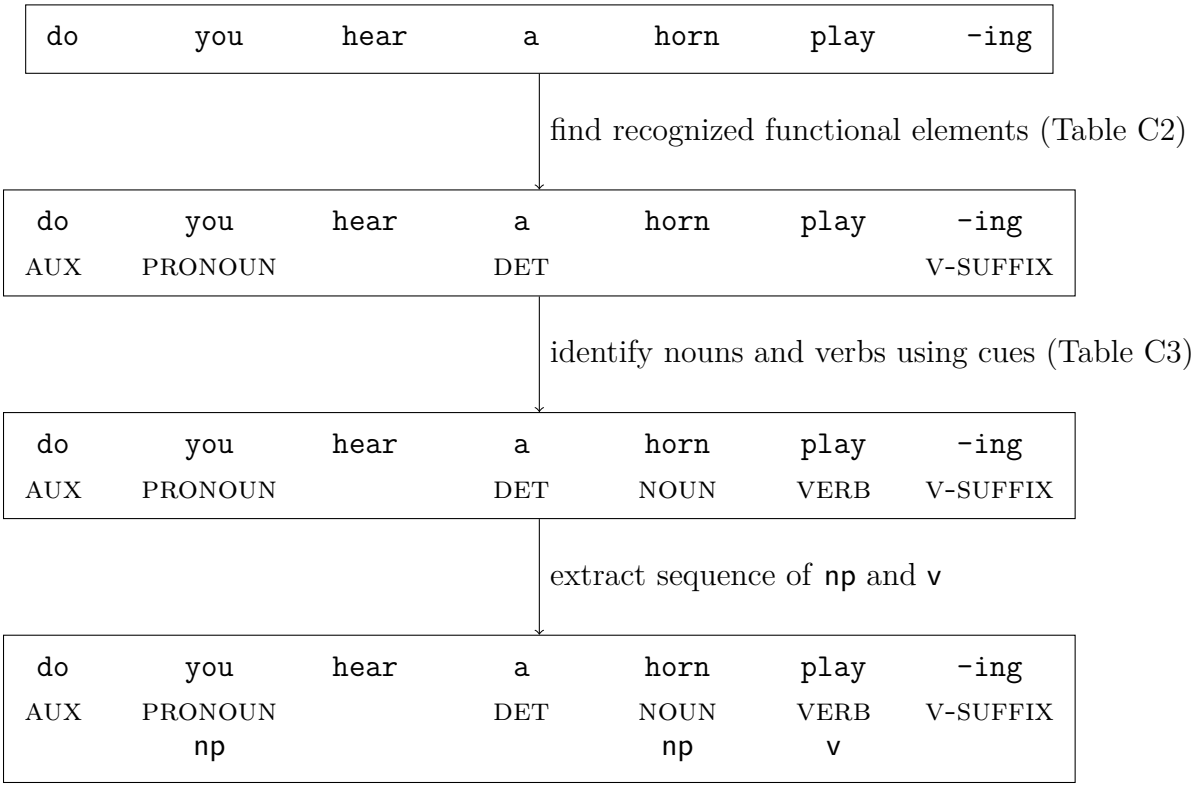

do you hear a horn play -ing
find recognized functional elements (Table C2)
do you hear a horn play -ing
AUX PRONOUN DET V-SUFFIX
identify nouns and verbs using cues (Table C3)
do you hear a horn play -ing
AUX PRONOUN DET NOUN VERB V-SUFFIX
extract sequence of np and v
do you hear a horn play -ing
AUX PRONOUN DET NOUN VERB V-SUFFIX
np np v

Figure C1

English	
np v	what is she do -ing
	you can stay out and play
	you can sit down
	do you still go to dance -ing school
np v np	we are try -ing it out
	you will have a zero
	the cat will get up on the chair
	he knock -ed three time -s
v	what happen -ed
	her hair curl -ed and everything
	anything miss -ing
	tattoo -ed man
v np	ask who that is knock -ing at your door
	tell her what happen -ed to your toy
	somebody knock -ed at my door two o'clock this morning
	that will make him go down

French	
np v	on va voir ailleurs
	laisse toi bien tomb -er
	tu as cass -é la boîte
	ça va être trop bas aussi
np v np	ce est moi qui ai fait ça
	on va voir si on trouve autre chose
	tu veux tir -er le sac
	on va faire un soleil
v	faut pas faire de bruit
	à aid -er
	eh ben fais chauff -er là-dedans
	ce était accroch -er où
np np v	ça y est tu as tout/adv/-- jet -é
	je vois que tu te amuses à saut -er
	oh je sais pas si je vais pouvoir hein
	moi je veux corrig -er maintenant

Japanese	
v	gohon yon -de
	ki ni naru kanaa
	mat -te
	haka -nai no
np v	kotchi no gohon yome -ba
	yuube mo yonda kara iya kanaa
	Asatokun gohan tabe -nai
	kore o mawashi -te yo
v np	gohon yon -de Asatokun
	nake -te kuru wa
	anoo sa kore dare no
	ja ne orenji no ne meemeehitsuji
np np v	sore dake wa yame -te
	kore Asatokun kitta ne
	sakanayasan mo koko aku jan
	Kakka ni kotchi kashi -te kureru

Table C4

Some example sentences corresponding to the most common string types for each language