Filtering input for learning constrained grammatical variability: The case of Spanish word order

Shalinee Maitra
Department of Linguistics
University of California Los Angeles
Los Angeles, CA 90095-1543
shalinee30@ucla.edu

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

Abstract

Children learn basic word order from data in which both subjects and objects can appear in variable positions. Spanish learners acquire a word order that deterministically places objects after verbs, and allows variation only in subject position. We present a model for acquiring this type of constrained variability from messy data. Our model expects that (1) its data contain a mixture of signal and noise for canonical word order, and (2) subjects control agreement on verbs. We find that this model can learn to filter noise from its data to identify the canonical word order for Spanish while a model that does not track subject-verb agreement cannot. These results suggest that having expectations about the types of regularities that the data will contain can help learners identify variability that is constrained along certain dimensions.

1 Introduction

Children acquire the canonical word order of their language at young ages, from input that contains a mixture of canonical and non-canonical word orders whose structure they cannot yet represent (Hirsh-Pasek and Golinkoff, 1996; Perkins and Lidz, 2021, 2020). Non-canonical sentences like *wh*-questions introduce perceived variability into learners' data, which they must abstract away from in order to identify basic subject and object position. However, some types of variability are part of the core grammatical phenomenon to be acquired. In Spanish, full lexical objects canonically must occur after verbs, but subject position is not fully deterministic: subjects can occur both pre- and post-verbally in basic clauses (1-2) (Lozano, 2006; Domínguez and Arche, 2008; De Prada Pérez and Pascual y Cabo, 2012). Learners must identify that this variability is a property of the language's basic clause syntax, whereas other variability is due to subject or object displacement in non-canonical sentence types (3). How do learners identify that basic subject position varies, but object position is fixed, if both argument positions appear to be variable in their data?

(1) *Mariela tiró la pelota.* (basic SVO)
Mariela throw-PAST-SG the-SG ball-SG
'Mariela threw the ball.'

(2) *Entró Mariela.* (basic VS)
Enter-PAST-SG Mariela
'Mariela entered.'

(3) *¿Cuál pelota tiró Mariela?* (*wh*-Q, OVS)
Which-SG ball-SG throw-PAST-SG Mariela
Which ball did Mariela throw?

On one proposal, learners might avoid being misled by messy data by assuming that some portion of their data is "noise," introduced by grammatical processes they cannot yet account for. Successful learning arises when learners are able to infer which portion of their data to treat as noise, and which portion to treat as signal for the rules governing the phenomenon they are trying to acquire (Perkins and Hunter, 2023; Perkins et al., 2022; Schneider et al., 2020). This can be seen as a mechanism for "regularization" in learning (Hudson Kam and Newport, 2005, 2009; Culbertson et al., 2013) whereby learners acquire a system that allows less variability than the data that they are learning from. But the case of Spanish word order poses a challenge for this approach. Here, learners must abstract away from certain *types* of variability— for instance, the noise introduced by non-canonical sentence types— while treating other types of variability as informative about the phenomenon to be acquired. That is, learners must identify that they should "regularize" along only certain dimensions.

We propose that learners might solve this problem by using knowledge about the specific types of regularities that grammars tend to exhibit. In the case of word order acquisition, learners might expect that subjects and objects will enter into different sorts of grammatical dependencies— for instance, that subjects tend to control agreement on verbs. We present a learner that looks for evidence of subject-verb agreement in its data, and uses this information to infer which portion of its data to treat as signal for underlying basic word order. We show that this learner is able to identify constrained variation in Spanish word order. We also show that our learner performs substantially better than a learner that does not track subject-verb agreement. This suggests that for certain types of grammatical generalizations, successful learning requires knowledge of the sorts of dependencies that grammars make available, along with mechanisms for detecting relevant evidence in noisy data.

2 Acquiring word order in Spanish

Cross-linguistically, children learn basic word order in infancy (Perkins and Lidz, 2020; Hirsh-Pasek and Golinkoff, 1996; Franck et al., 2013; Gavarró et al., 2015; Zhu et al., 2022). They do so at ages even before they have adult-like representations for non-canonical clause types where this basic word order is distorted. For instance, infants learning English identify that their language is canonically SVO even before they can identify that arguments have been moved in *wh*-questions (Hirsh-Pasek and Golinkoff, 1996; Perkins and Lidz, 2021). This suggests that learners have a way to implicitly "filter" the messiness introduced by non-canonical clause types when learning basic clause syntax (Pinker, 1984; Gleitman, 1990; Lidz and Gleitman, 2004).

On one proposal, learners might infer how to separate "signal" for the grammatical phenomenon being acquired from "noise" introduced by various other processes (Perkins et al., 2022; Perkins and Hunter, 2023). This inference is possible even if learners do not know ahead of time which of the utterances they hear should be treated as noise—for instance, because they have not yet learned what basic vs. non-basic clauses look like. Perkins and Hunter (2023) show that a learner can use the distributions of imperfectly-identified noun phrases and verbs in child-directed speech to determine which data to treat as signal for basic word order, without prior expectations about where noise will occur. Their model successfully filtered its noisy input in order to infer that French and English have canonical SVO word order. A similar mechanism has been applied to model the successful acquisition of verb transitivity classes (Perkins et al., 2022).

Here, we ask whether this same type of filtering mechanism can succeed in cases of more variable word order. In Spanish, full lexical objects are obligatorily postverbal, but subjects can occur both before and after the verb in basic clauses.[1] But a variety of constructions obscure evidence for these basic word orders. For instance, *wh*-dependencies and topic and focus constructions introduce frequent argument displacement. Furthermore, Spanish has frequent null subjects, which cause a unique ambiguity for learning basic word order. For a child at early stages of syntactic development, sentences like (4) and (5) may be structurally ambiguous. If the child does not know the meaning of these words and whether null subjects are present, it is unclear whether the noun phrase after the verb is the subject or the object.

(4) *Traen los regalos.*
pro bring-PL the-PL gift-PL
'(They) bring the gifts.'

(5) *Llegan los profesores.*
arrive-PL the-PL teacher-PL
'The teachers arrive.'

On the basis of ambiguous data like (4) and (5), we can imagine at least two erroneous conclusions that the learner may reach. On the one hand, the learner might conclude that both of these sentences are transitive with null subjects, making the postverbal noun phrases both objects. This would mean that the learner is missing relevant evidence for postverbal subjects in the language. On the other hand, the learner might decide that both of these sentences are intransitive, and the postverbal noun phrases are both subjects. This would mean that the learner is missing relevant evidence for postverbal objects in the language. If this type of data is prevalent, the learner may need additional information to draw the correct conclusion that the language has both postverbal subjects and postverbal objects.

One possible source of information that could help children reach the correct conclusion is subject-verb agreement. Because objects do not agree with verbs while subjects do, postverbal nominals do not always match verbs in number (6). This

[1] In basic clauses with broad focus, postverbal subjects typically occur in intransitive clauses with unaccusative rather than unergative verbs (De Prada Pérez and Pascual y Cabo, 2012). There is also debate regarding the canonical clausal position of subjects in Spanish (Villa-García, 2012). We abstract away from these issues in the current discussion.

agreement asymmetry reflects a cross-linguistic tendency: in languages where verbs agree with an argument, that argument is typically a subject (Moravcsik, 1974, 1978; Gilligan, 1987).[2]

(6) *Trae los regalos.*
pro bring-SG the-PL gift-PL
(He) brings the gifts.

If children expect subjects to control agreement on verbs, and can find evidence for these agreement dependencies in their data, then number mismatches like the one in (6) could help them identify the postverbal noun phrase as an object and not a subject. Furthermore, a proliferation of postverbal noun phrases that agree with verbs could provide evidence for postverbal subjects, particularly if these occur at a rate higher than would be expected if they were all objects.

In languages that morphologically mark subject-verb agreement, there is evidence that infants can track these patterns from very young ages (Nazzi et al., 2011), along with other types of morphologically-marked dependencies (Van Heugten and Shi, 2010; Soderstrom et al., 2007; Hohle et al., 2006; Santelmann and Jusczyk, 1998). It is not clear how abstractly children represent these types of dependencies at young ages (Culbertson et al., 2016), but these sensitivities make it plausible that they might use them in the process of word order acquisition, particularly in a language like Spanish that has rich and transparent agreement morphology.

Can a filtering mechanism of the sort proposed in previous literature successfully acquire the constrained variability in Spanish word order, given the range of noise in the data that children will encounter? We present a model that learns from strings of imperfectly-represented noun phrases and verbs. It learns to filter noise from its data in order to identify canonical word order, using evidence for subject-verb number agreement but no further cues to sentence structure. We find that the learner is able to successfully identify that Spanish has postverbal objects and variation in subject position. Moreover, this learner performs substantially better than a learner that relies on the distributions of noun phrases and verbs alone, without expecting subjects and verbs to agree. Thus, solving this problem may require not only the ability to learn in a noise-tolerant way from distributions in data, but also expectations about the types of agreement dependencies that clause arguments enter into.

[2]Some languages mark object as well as subject agreement, while others do not mark subject verb agreement. Two relevant questions for future work are (i) how a learner would identify multiple agreement dependencies in languages with more complex agreement systems and (ii) how a learner would fare in a language with fewer agreement dependencies.

3 Model

We adapt a Bayesian learner from Perkins and Hunter (2023). The model observes strings of noun phrases and verbs tagged for number features. The model assumes that its observed strings have been generated by some mixture of canonical and non-canonical grammatical processes. Specifically, the learner chooses among discrete composite probabilistic context-free grammars (PCFGs) that contain different sets of "core" rules governing canonical word order (e.g., SVO, SOV, etc.), and a shared set of "noise" rules that introduce additional variability into the data. We compare two models whose hypothesis spaces contain different sets of composite PCFGs, one that expects subject-verb number agreement ('Agreement Model') and one that does not ('No-Agreement Model'). The model seeks to divide its data into signal and noise in order to identify which combination of core and noise rules best explains the distributions it observes.

3.1 Generative Model

The grammars in the *Agreement Model* generate strings with exactly one verb, either singular or plural (v-sg or v-pl), and up to two noun phrases, either singular or plural (np-sg or np-pl). Two of the grammars in the Agreement Model's hypothesis space are shown in Table 1: one whose canonical word order is SVO, and one whose canonical word order requires objects to occur after verbs but allows subjects to vary in their position ('VO', the target word order of Spanish). In these grammars, NP-pl is deterministically rewritten as np-pl, NP-sg as np-sg, V-pl as v-pl, and V-sg as v-sg; these are not shown for purposes of space.

These grammars enforce subject-verb argreement in their core rules by requiring, for S expansions, that only an NP-pl occurs with a VP-pl and only an NP-sg occurs with a VP-sg. However, for VP expansions, both NP-pl and NP-sg are allowed to occur with a V-sg or V-pl, so verbs are not required to agree with direct objects in number.

The learner chooses among nine possible grammars of this sort, whose core rules correspond to

SVO Core Rules	VO Core Rules	Shared Noise Rules	
S $\rightarrow$ NP-pl VP-pl	S $\rightarrow$ NP-pl VP-pl	S $\dashrightarrow$ NP-pl VP-pl	S $\dashrightarrow$ VP-pl
S $\rightarrow$ NP-sg VP-sg	S $\rightarrow$ NP-sg VP-sg	S $\dashrightarrow$ NP-sg VP-sg	S $\dashrightarrow$ VP-sg
	S $\rightarrow$ VP-pl NP-pl	S $\dashrightarrow$ VP-pl NP-pl	
	S $\rightarrow$ VP-sg NP-sg	S $\dashrightarrow$ VP-sg NP-sg	
VP-pl $\rightarrow$ V-pl NP-pl	VP-pl $\rightarrow$ V-pl NP-pl	VP-pl $\dashrightarrow$ V-pl NP-pl	VP-pl $\dashrightarrow$ NP-pl V-pl
VP-pl $\rightarrow$ V-pl NP-sg	VP-pl $\rightarrow$ V-pl NP-sg	VP-pl $\dashrightarrow$ V-pl NP-sg	VP-pl $\dashrightarrow$ NP-sg V-pl
VP-pl $\rightarrow$ V-pl	VP-pl $\rightarrow$ V-pl	VP-pl $\dashrightarrow$ V-pl	
VP-sg $\rightarrow$ V-sg NP-pl	VP-sg $\rightarrow$ V-sg NP-pl	VP-sg $\dashrightarrow$ V-sg NP-pl	VP-sg $\dashrightarrow$ NP-pl V-sg
VP-sg $\rightarrow$ V-sg NP-sg	VP-sg $\rightarrow$ V-sg NP-sg	VP-sg $\dashrightarrow$ V-sg NP-sg	VP-sg $\dashrightarrow$ NP-sg V-sg
VP-sg $\rightarrow$ V-sg	VP-sg $\rightarrow$ V-sg	VP-sg $\dashrightarrow$ V-sg	

Table 1: SVO and VO grammars, Agreement Model

nine distinct word order options. We model the learning process as a choice among these nine discrete grammars; see Perkins and Hunter (2023) for discussion of the role of discreteness in the learner's hypothesis space in this type of model. These grammars include the four most restricted word orders, where subjects deterministically occur either before or after the verb phrase and objects before or after the verb: SVO, SOV, OVS, and VOS (the four options arising from a 2x2 choice of subject and object position). The hypothesis space also includes a 'Free' word order that allows any ordering of subjects and objects, and four word orders that allow some degree of variation: two that fix object position as either OV or VO and allow subjects on either side of the verb phrase; and two that fix subject position as either SV or VS and allow objects on either side of the verb. Note that each of these last four grammars essentially combine two of the more restricted grammars. In particular, the VO grammar (the target word order for Spanish) is the union of the VOS and SVO grammars. See the Appendix for full details.

In addition to the core rules that generate canonical word order, each grammar has a set of noise rules (represented by dashed arrows in Table 1) that manipulate the same set of terminal and nonterminal symbols as the core rules, but allow for all possible permutations and deletions of clause arguments. Each of the nine grammars in the learner's hypothesis space has the same set of noise rules. This allows all of the grammars to generate any of the strings in the dataset. For example, the SVO grammar can generate the string v-pl np-sg np-pl via the trees in Fig. 1. In the first tree, two noise rules are used: the noisy S expansion places the subject after the VP, and the noisy VP expansion places the object after the verb. Notice that it is also possible for a tree to be generated by a mixture of

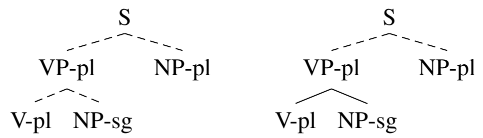

Figure 1: Two possible analyses of v-pl np-sg np-pl (suppressing NP-sg $\rightarrow$ np-sg, NP-pl $\rightarrow$ np-pl and V-pl $\rightarrow$ v-pl rewrites) where solid lines indicate core rules and dashed lines indicate noise rules

core and noise rules, as in the second tree: here, the S expansion is noisy, but the VP can be expanded according to the core rules of the SVO grammar.

The core rules of these grammars do not contain the rules S $\rightarrow$ VP-sg and S $\rightarrow$ VP-pl, meaning that the learner expects canonical clauses to have subjects. These expansions of S only occur in the noise rules; subject-drop is assumed to be a process that introduces noise for basic word order learning.

The *No-Agreement Model* is just like the Agreement Model, except that the grammars in its hypothesis space do not encode subject-verb number agreement. These grammars generate strings that contain exactly one v and up to two nps, not marked for number. The SVO grammar and the VO grammar are shown in Table 2. In these grammars, NP is deterministically rewritten as np and V is deterministically rewritten as v; these are again omitted for the sake of space.

The No-Agreement Model has the same nine word order options as the Agreement Model in its hypothesis space: the four most deterministic word orders, four that allow variability in either subject or object position, and one that allows both subject and object position to vary. Each of these nine grammars again shares the same set of noise rules, which allow any word ordering as well as argument deletion. Just as in the Agreement Model, subjectless clauses are only allowed via the grammars' noise rules.

SVO Core Rules	VO Core Rules	Shared Noise Rules
S → NP VP	S → NP VP	S ⇢ NP VP
	S → VP NP	S ⇢ VP NP
		S ⇢ VP
VP → V NP	VP → V NP	VP ⇢ V NP
VP → V	VP → V	VP ⇢ V
		VP ⇢ NP V

Table 2: SVO and VO grammars, No-Agreement Model

Agreement	No Agreement
0.38 v-sg	0.5 v
0.18 v-sg np-sg	0.25 v np
0.14 v-pl	0.12 np v
0.08 np-sg v-sg	0.06 np v np
0.04 np-sg v-sg np-sg	0.05 v np np
0.04 v-pl np-sg	0.02 np np v
0.03 v-sg np-sg np-sg	
0.02 v-pl np-pl	
0.02 np-sg v-pl	

Table 3: Proportions of most frequent string types

For both models, the prior distribution over the nine grammars G in the learner's hypothesis space is uniform, meaning each of the nine grammars has the same prior probability. This means that none of the canonical word orders is preferred *a priori*. Each of the allowable core and noise rules in these grammars has some probability associated with it. To work with these rule probabilities, we recast the composite grammars illustrated in Tables 1 and 2 into standard PCFGs, following Perkins and Hunter (2023).[3] For every nonterminal N in these grammars, we add additional nonterminals N^+ and N^-. The expansions for N^+ and N^- are determined by the grammar's core and noise rules, respectively. We also add the rules $N \rightarrow N^+$ and $N \rightarrow N^-$, whose weights represent the probabilities for using a core vs. noise expansion of N. Let $\vec{\theta}_{n_G}$ be the vector of probabilities for expanding a nonterminal n in the resulting standard PCFG for G. The prior distribution over $\vec{\theta}_{n_G}$ is a Dirichlet distribution with parameters α_{n_G}. We set all components of α in these distributions to 1, which results in a uniform distribution over the rule probabilities. This means that all core expansions of a given nonterminal are equally likely *a priori*, as are all noise expansions.

Each grammar conditions a distribution over trees and strings. Just as for any standard PCFG, the probability of generating a string via a particular tree under grammar G is the product of the rule probabilities $\vec{\theta}_G$ used in that tree. To calculate the overall probability of a string under grammar G, we sum over the probabilities of all of the ways that it could be generated.

3.2 Inference

Our model infers the posterior probability distribution over its grammars G and an approximation of trees $\vec{t}$ given its observed strings $\vec{w}$. Following Perkins and Hunter (2023), instead of inferring a distribution over $\vec{t}$ directly, we sample approximations of trees, which we call 'coarse structures', $\vec{s}$. These coarse structures abstract away from the core vs. noise distinctions in the trees. For example, both trees in Fig. 1 would share the same coarse structure: the same tree without a distinction between dashed and solid lines. Abstracting away from this distinction means that all grammars in the learner's hypothesis space can generate every coarse structure, using either noise rules, or core rules, or some combination. This allows for feasible sampling of grammars given a sample of coarse structures.

We use Gibbs Sampling to estimate the joint posterior probability of grammars and coarse structures, $P(G, \vec{s} \mid \vec{w})$, summing over all combinations of core and noise options in $\vec{s}$ and integrating over $\vec{\theta}$. The steps of sampling work as follows. First, G is randomly initialized to one of the nine grammars in the hypothesis space. Then, we alternate between drawing samples from the posterior probability of a grammar given a set of coarse structures for the observed strings, $P(G \mid \vec{w}, \vec{s})$, and the posterior probability of coarse structures given a grammar and the observed strings, $P(\vec{s} \mid \vec{w}, G)$.

Via Bayes' Rule, the posterior probability of a grammar given coarse structures and strings, $P(G \mid \vec{w}, \vec{s})$, is proportional to the likelihood of the strings and coarse structures given the grammar, times the prior probability of that grammar:

$$(1) \quad P(G \mid \vec{w}, \vec{s}) = \frac{P(\vec{s}, \vec{w} \mid G)P(G)}{\sum_{G'} P(\vec{s}, \vec{w} \mid G')P(G')}$$

We assume that all grammars have equal prior probability, and calculate the likelihood $P(\vec{s}, \vec{w} \mid G)$ following Perkins and Hunter (2023). After sampling a new grammar from the posterior distribution in Eq. (1), we sample a new set of coarse structures from $P(\vec{s} \mid \vec{w}, G)$ using a Hastings pro-

[3] This formalization bears resemblance to a latent variable PCFG (Cohen, 2017), in which the choice between noise (−) vs. non-noise (+) at each nonterminal node could be recast as a choice of a particular latent state. We thank an anonymous reviewer for pointing this out.

posal, following a method introduced in Johnson et al. (2007). These steps are repeated until the chain converges to a stable distribution which estimates the joint posterior $P(G, \vec{s} \mid \vec{w})$. We refer the reader to Perkins and Hunter (2023) for more details of the sampling procedure.

For the results reported below, 20,000 iterations of Gibbs Sampling were performed. Every tenth sample of the last 10,000 iterations was analyzed.

4 Simulations

4.1 Data

We tested our learners on datasets of child-directed Spanish created from the Fernandez/Aguado corpus in CHILDES (Fernandez Vazquez and Gerardo Aguado). The corpus includes a total of 45,610 utterances directed to 47 different children between the ages of approximately 3;0 and 4;0. This corpus was chosen because of its large size and the large number of children included, allowing for more reliable estimates of the distributions that any given child might hear.

The dataset for the Agreement Model consisted of strings of noun phrases and verbs annotated with number features. We conducted an automatic search of the corpus, using a heuristic that aimed to approximate the immature grammatical category knowledge of an infant learning basic word order. Because young infants can differentiate nouns from verbs using determiners, auxiliaries, and pronouns (Babineau and Christophe, 2022; Shi and Melançon, 2010; Hicks et al., 2007) we noisily identified noun phrases and verbs in the corpus using these functional cues. All full pronouns were included as np's, with their number determined by the form of the pronoun. Any word occurring after a determiner was counted as the head of an np, and its number was determined based on the inflection of the determiner. Any word occurring after an auxiliary was counted as a v, and its number was determined by the inflection on the auxiliary. Proper names were counted as np-sg's. *Wh*-words and clitics were not counted as np's, because there is no evidence that children identify these as nominals before learning basic word order (Perkins and Lidz, 2021; Brusini et al., 2017).

After these strings were extracted, only strings with exactly one verb and up to two noun phrases where at least one noun phrase matched the verb in number were retained. From this subset of the corpus, we calculated the proportion of each string type, and sampled 25 strings according to these proportions. This resulted in 9 string types included in the dataset for the Agreement learner (see Table 3). This dataset is substantially noisy: nearly 60% of these strings cannot not be generated by the core rules of the VO grammar, which is the target word order of Spanish, without using noise rules.

The dataset for the No-Agreement learner was generated by the same process and heuristics for finding noun phrases and verbs, but number features were not tagged.[4] We sampled 25 strings according to their proportions in the corpus, resulting in the 6 string types in Table 3. Just as in the dataset for the Agreement Model, almost 60% of these strings cannot be generated by the core rules of the VO grammar without the option of noise.

4.2 Results

Figure 2 shows the posterior distribution over grammars inferred by the Agreement and the No-Agreement Model, averaged across 10 runs of each learner. In these graphs, the dashed line represents no substantial learning: a learner that maintains its prior belief that all of its 9 grammars are equally probable would infer a distribution with all bars hovering around 0.11.

The No-Agreement Model inferred roughly this flat distribution. The target VO grammar, along with most other grammars, was assigned posterior probability around 0.11. VOS and OVS were assigned slightly higher posterior probability (both a mean of 0.14); overall, the model gave slightly higher probability to the more restrictive grammars. The fact that all grammars were assigned low and approximately equal probability suggests that the No-Agreement Model did not learn much useful information about Spanish word order.

The Agreement Model, by contrast, inferred a substantially different distribution. Three of the learner's grammars received much higher probability than the other six. These three grammars are VO (mean posterior probability: 0.23), SVO (mean: 0.20), and VOS (mean: 0.28). The other grammars

[4] There are certain strings that were present in the No Agreement dataset that were not present in the Agreement dataset. For example, the string np-sg v-pl np-sg would not be included in the Agreement dataset because the Agreement grammars cannot generate this string (since neither np agrees with the verb in number), but this string would be tagged as np v np under the heuristics for the No Agreement dataset, and thus would be included. This is why the proportions in the Agreement dataset in Table 3 do not add up to the relevant proportions in the No Agreement dataset.

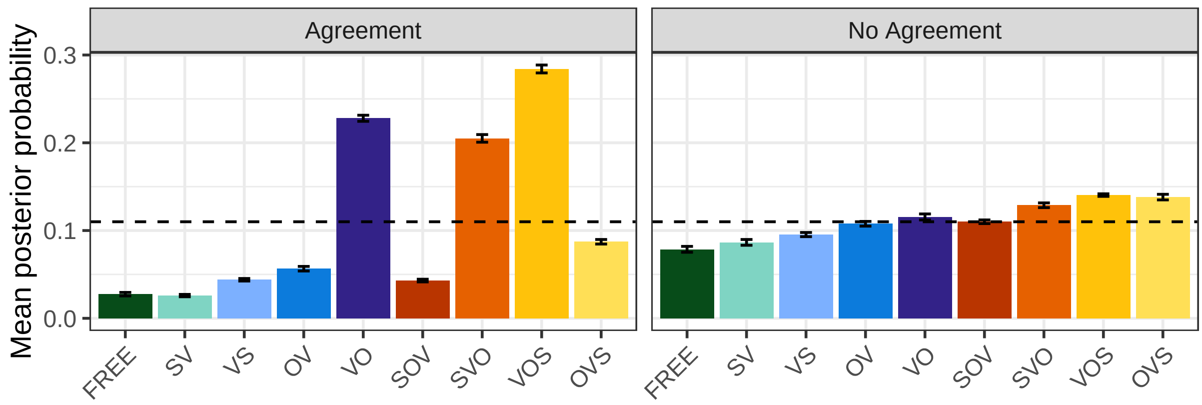

Figure 2: Posterior distribution over word-order grammars (G)

were all assigned much lower probability, ranging from 0.03 (SV) to 0.09 (OVS).

While the target VO grammar is among the three that the learner identified as having highest posterior probability, it did not identify this grammar as the single most probable. However, looking more closely at these results, we see that the learner's inference was fairly sensible. All three grammars with stand-out posterior probability only allow postverbal objects, which indicates that the learner successfully identified Spanish object position. Furthermore, the strings that the target VO grammar can generate are exactly the combination of the strings generated by the SVO and VOS grammars. So, the fact that these three grammars were assigned the highest posterior probability indicates that the learner had success in determining that object position is fixed, but subject position varies. This inference is striking given the degree of noise that the learner needed to overcome: nearly 60% of the strings in its data were not consistent with the canonical word order options that it successfully identified, without taking noise into account.

Interestingly, we see that the learner's inferred distribution favors VOS by a small amount. Why would this be the case? One reason may be that this type of Bayesian learner prefers more restrictive hypotheses. This is a phenomenon known as "Bayesian Occam's Razor," under which the hypothesis with the fewest degrees of freedom that can explain all the data will be preferred (Griffiths et al., 2008). In the case of these models, the SVO and VOS grammars correspond to hypotheses with fewer degrees of freedom than the more flexible VO grammar. Spanish allows both of these word orders, so the combination of explaining the data well and having fewer degrees of freedom gives VOS a small advantage over VO, and gives SVO high probability as well. The same preference for restrictive hypotheses is visible in the No Agreement Model, where the four most constrained grammars tended to receive higher posterior probability than the more flexible ones.

The learner's slight preference for VOS points to an additional limitation in its search for subject-verb agreement. The strings that provide the best evidence for VOS are the v-initial strings in which there is at least one postverbal np that matches the v in number: our learner will tend to take this match as evidence for subject-verb agreement, and analyze these strings as having postverbal subjects. These strings make up 23% of the learner's dataset, lending support to grammars in which the subject is fixed postverbally. However, because Spanish allows null subjects, a number of these postverbal np's are likely to be objects rather than subjects: this is the ambiguity demonstrated in (4-5) in Section 2. If a child were only tracking number agreement, like our learner, perhaps that child would likewise mis-analyze many of these sentences.

Possible extensions of this learner might leverage other information in order to overcome this bias towards VOS. One of the potential cues that is available in Spanish, but is removed by our preprocessing of the data, is person agreement. Tracking person features would give the learner an additional way to disambiguate between subjects and objects. Of the v-initial strings in which the v and a potential subject np match in number, approximately 25% *mismatch* in person features (see Table 4). These person mismatches could help a more sophisticated learner identify that many of these strings are underlyingly verb-object, not verb-subject, just as mismatches in number features can disambiguate

V-initial string type	Prop. person mismatches
v-sg np-sg np-sg	0.28
v-sg np-sg	0.25
v-pl np-pl	0.16
Overall	**0.25**

Table 4: Person mismatches in relevant v-initial strings

these parses in cases like (6). An example is shown in (7), where the 3rd-person postverbal object *a ella* mismatches the 1st-person inflected verb *veo*.

(7) *La veo a ella.*
pro her-3SG see-1SG to her-3SG
'(I) see her.'

So, a learner that makes use of a wider range of evidence for subject-verb agreement might overcome its bias towards determinism, and infer with higher probability that subject position is variable.

In sum, our results show that the Agreement Model was able to use its expectation of subject-verb agreement to abstract away from a great degree of noise in its data and infer a canonical word order in which objects are obligatorily postverbal, with some variation in subject position. By contrast, the No-Agreement Model failed to infer that any of its hypothesized canonical word orders were more probable than any of the others. Thus, tracking subject-verb number agreement helped substantially in this learning problem. A learner that expected subjects to agree with verbs was able to draw reasonable inferences about Spanish word order on the basis of noisy data; a learner with no awareness of agreement could not.

5 Discussion

We present a model for learning constrained variability in Spanish word order. Spanish learners need to acquire a word order with obligatorily postverbal objects and variable subject position from messy data, in which both subjects and objects might appear to vary in position. We extend an approach introduced by Perkins and Hunter (2023) to model this learning as a case of separating "signal" for basic word order from "noise" from non-canonical clause types. We pursue the hypothesis that, in solving this problem, learners may make use of knowledge that subjects and verbs will tend to agree. We compare a learner that attempts to identify a grammar of canonical word order using subject-verb number agreement to a learner that relies entirely on noun phrase and verb distributions. We find that the model that tracks subject-verb agreement is able to infer Spanish word order, whereas the model with no knowledge of agreement cannot. This suggests that knowledge of the types of dependencies that clause arguments enter into may helpfully guide word order learning.

Our case study demonstrates how tolerant this learning mechanism is to noise: the learner succeeds at identifying the target canonical word order even though approximately 60% of the data appears inconsistent with that order. The learner's noise-tolerance comes in part from its ability to find useful information in sub-parts of strings, instead of treating each string as either entirely signal or entirely noise. The learner assumes that noise can occur in any of the internal nodes in a tree individually, so it entertains the possibility that a string could be generated with a mixture of core vs. noise rules, as shown in Figure 1. This allows the learner to look within strings to find evidence for the grammatical regularities it expects, thereby making use of more of its data.

Thus, if Spanish-learning children are reliably able to track subject-verb agreement at the age when they are learning word order, then they might be able to use agreement to aid in this task, even in the absence of other reliable cues to sentence structure (e.g., from meaning or prosody; Pinker, 1984; Christophe et al., 2008). However, this depends on children knowing the morphological forms of number and potentially person inflection in the language. Prior work shows that French learners track subject-verb dependencies in infancy (Nazzi et al., 2011), and learners in various languages track similar dependencies at young ages (Van Heugten and Shi, 2010; Soderstrom et al., 2007; Hohle et al., 2006; Santelmann and Jusczyk, 1998). However, we do not know precisely when children begin to track these dependencies, and how reliably and abstractly they represent them (Culbertson et al., 2016). Further work could explore whether our filtering mechanism would succeed even if learners have noisy or incomplete representations of these dependency types. These findings also invite further behavioral work on the acquisition of agreement in Spanish and similar languages.

Our model provides a window into the mechanisms for acquiring basic clause syntax in a language with frequent argument-drop and complex argument realization patterns. Subject pro-drop is a frequent and basic property of Spanish; however, our model treats this as a type of noise to ignore, and expects that canonical clauses will have overt subjects. While learners must eventually acquire pro-drop in Spanish, it may make sense for a learner to only attempt to learn canonical subject position from overt arguments, setting aside subject-drop as a phenomenon to be acquired independently. Indeed, in exploratory simulations, we find that allowing null subjects in the learner's core grammar rules does not help it identify Spanish word order; what helps is knowledge of subject-verb agreement. Our model therefore makes the prediction that knowledge of subject-verb agreement, but not necessarily pro-drop, may need to be acquired prior to the acquisition of word order in Spanish— a prediction that could be tested in future behavioral work. Beyond Spanish, many languages with argument-drop and more variable word orders also have rich case and agreement systems. The model presented here could therefore be extended to explore how case and agreement dependencies may inform learning in languages with diverse argument structure profiles.

These results have broader implications for our understanding of when and how learners regularize variable data (Hudson Kam and Newport, 2005, 2009; Reali and Griffiths, 2009; Ferdinand et al., 2019). We highlight a distinction between forms of regularization in which learners (i) abstract away from variability in data in order to draw fully deterministic generalizations, and (ii) draw generalizations that are not fully deterministic, but are still more constrained than the data would appear to support. For the current case study, we propose that learners use knowledge about the kinds of regularities that grammars tend to exhibit in order to identify which types of variability they should learn from, and which types they should treat as noise. This mechanism may generalize to other areas in language acquisition and learning in other domains, in which learners' regularization tendencies arise from the expectation that their data will noisily reflect a richly structured underlying system.

Acknowledgments

We thank Xinyue Cui, Tim Hunter, Victoria Mateu, the UCLA Psycholinguistics/Computational Linguistics Seminar, and three anonymous reviewers for helpful feedback and assistance.

References

Mireille Babineau and Anne Christophe. 2022. Preverbal infants' sensitivity to grammatical dependencies. *Infancy*, 27(4):648–662.

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

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

Shay B Cohen. 2017. Latent-variable pcfgs: Background and applications. In *Proceedings of the 15th Meeting on the Mathematics of Language*, pages 47–58.

Jennifer Culbertson, Elena Koulaguina, Nayeli Gonzalez-Gomez, Géraldine Legendre, and Thierry Nazzi. 2016. Developing knowledge of nonadjacent dependencies. *Developmental Psychology*, 52(12):2174.

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

Ana De Prada Pérez and Diego Pascual y Cabo. 2012. Interface heritage speech across proficiencies: unaccusativity, focus, and subject position in spanish. In *Selected Proceedings of the 14th Hispanic Linguistics Symposium*, pages 308–318. Cascadilla Proceedings Project Somerville, MA.

Laura Domínguez and María J Arche. 2008. Optionality in L2 grammars: the acquisition of SV/VS contrast in Spanish. In *Proceedings of the 32nd Annual Boston University Conference on Language Development*, Cambridge, MA. Cascadilla Press.

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

Marta Fernandez Vazquez and Alonso Gerardo Aguado. CHILDES Database Spanish Fernandez/Aguado Corpus.

Julie Franck, Severine Millotte, Andres Posada, and Luigi Rizzi. 2013. Abstract knowledge of word order by 19 months: An eye-tracking study. *Applied Psycholinguistics*, 34(2):323–336.

Anna Gavarró, Maya Leela, Luigi Rizzi, and Julie Franck. 2015. Knowledge of the OV parameter setting at 19 months: Evidence from Hindi–Urdu. *Lingua*, 154:27–34.

Gary Martin Gilligan. 1987. *A cross-linguistic approach to the pro-drop parameter.* Ph.D. thesis, University of Southern California, Los Angeles, CA.

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

Thomas L Griffiths, Charles Kemp, and Joshua B Tenenbaum. 2008. Bayesian models of cognition. In R. Sun, editor, *The Cambridge handbook of computational psychology*, page 59–100. Cambridge University Press.

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

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

Barbara Hohle, Michaela Schmitz, Lynn M Santelmann, and Jurgen Weissenborn. 2006. The recognition of discontinuous verbal dependencies by german 19-month-olds: Evidence for lexical and structural influences on children's early processing capacities. *Language Learning and Development*, 2(4):277–300.

Carla L Hudson Kam and Elissa L Newport. 2005. Regularizing unpredictable variation: The roles of adult and child learners in language formation and change. *Language Learning and Development*, 1(2):151–195.

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

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

Jeffrey Lidz and Lila R Gleitman. 2004. Argument structure and the child's contribution to language learning. *Trends in Cognitive Sciences*, 8(4):157–161.

Cristóbal Lozano. 2006. Focus and split-intransitivity: the acquisition of word order alternations in non-native spanish. *Second Language Research*, 22(2):145–187.

Edith Moravcsik. 1974. Object-verb agreement. *Working Papers on Language Universals*, 15:25–140.

Edith A Moravcsik. 1978. Agreement. In Joseph H Greenberg, Charles A Ferguson, and Edith A Moravcsik, editors, *Universals of Human Language. IV: Syntax*, pages 331–74. Stanford University Press, Stanford.

Thierry Nazzi, Isabelle Barrière, Louise Goyet, Sarah Kresh, and Géraldine Legendre. 2011. Tracking irregular morphophonological dependencies in natural language: Evidence from the acquisition of subject-verb agreement in french. *Cognition*, 120(1):119–135.

Laurel Perkins, Naomi H Feldman, and Jeffrey Lidz. 2022. The power of ignoring: filtering input for argument structure acquisition. *Cognitive Science*, 46(1):e13080.

Laurel Perkins and Tim Hunter. 2023. Noise-tolerant learning as selection among deterministic grammatical hypotheses. In *Proceedings of the 6th meeting of the Society for Computation in Linguistics (SCiL 2023)*.

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

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

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

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

Lynn M Santelmann and Peter W Jusczyk. 1998. Sensitivity to discontinuous dependencies in language learners: Evidence for limitations in processing space. *Cognition*, 69(2):105–134.

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

Rushen Shi and Andréane Melançon. 2010. Syntactic categorization in french-learning infants. *Infancy*, 15(5):517–533.

Melanie Soderstrom, Katherine S White, Erin Conwell, and James L Morgan. 2007. Receptive grammatical knowledge of familiar content words and inflection in 16-month-olds. *Infancy*, 12(1):1–29.

Marieke Van Heugten and Rushen Shi. 2010. Infants' sensitivity to non-adjacent dependencies across phonological phrase boundaries. *The Journal of the Acoustical Society of America*, 128(5):EL223–EL228.

Julio Villa-García. 2012. Spanish subjects can be subjects: Acquisitional and empirical evidence. *Iberia: An International Journal of Theoretical Linguistics, 4 (2), 124-169.*

Jingtao Zhu, Julie Franck, Luigi Rizzi, and Anna Gavarró. 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.

A Complete List of Grammars

VO Core Rules	**OV Core Rules**	**SV Core Rules**	**VS Core Rules**	**Free Core Rules**
S → NP-pl VP-pl	S → NP-pl VP-pl	S → NP-pl VP-pl	S → VP-pl NP-pl	S → NP-pl VP-pl
S → NP-sg VP-sg	S → NP-sg VP-sg	S → NP-sg VP-sg	S → VP-sg NP-sg	S → VP-pl NP-pl
S → VP-pl NP-pl	S → VP-pl NP-pl			S → NP-sg VP-sg
S → VP-sg NP-sg	S → VP-sg NP-sg			S → VP-sg NP-sg
VP-pl → V-pl NP-pl	VP-pl → NP-pl V-pl	VP-pl → NP-pl V-pl	VP-pl → NP-pl V-pl	VP-pl → NP-pl V-pl
VP-pl → V-pl NP-sg	VP-pl → NP-sg V-pl	VP-pl → NP-sg V-pl	VP-pl → NP-sg V-pl	VP-pl → NP-sg V-pl
		VP-pl → V-pl NP-pl	VP-pl → V-pl NP-pl	VP-pl → V-pl NP-pl
		VP-pl → V-pl NP-sg	VP-pl → V-pl NP-sg	VP-pl → V-pl NP-sg
VP-pl → V-pl	VP-pl → V-pl	VP-pl → V-pl	VP-pl → V-pl	VP-pl → V-pl
VP-sg → V-sg NP-sg	VP-sg → NP-pl V-sg	VP-sg → NP-pl V-sg	VP-sg → NP-pl V-sg	VP-sg → NP-pl V-sg
VP-sg → V-sg NP-pl	VP-sg → NP-sg V-sg	VP-sg → NP-sg V-sg	VP-sg → NP-sg V-sg	VP-sg → NP-sg V-sg
		VP-sg → V-sg NP-pl	VP-sg → V-sg NP-pl	VP-sg → V-sg NP-pl
		VP-sg → V-sg NP-sg	VP-sg → V-sg NP-sg	VP-sg → V-sg NP-sg
VP-sg → V-sg	VP-sg → V-sg	VP-sg → V-sg	VP-sg → V-sg	VP-sg → V-sg

SVO Core Rules	**SOV Core Rules**	**VOS Core Rules**	**OVS Core Rules**
S → NP-pl VP-pl	S → NP-pl VP-pl	S → VP-pl NP-pl	S → VP-pl NP-pl
S → NP-sg VP-sg	S → NP-sg VP-sg	S → VP-sg NP-sg	S → VP-sg NP-sg
VP-pl → V-pl NP-pl	VP-pl → NP-pl V-pl	VP-pl → V-pl NP-pl	VP-pl → NP-pl V-pl
VP-pl → V-pl NP-sg	VP-pl → NP-sg V-pl	VP-pl → V-pl NP-sg	VP-pl → NP-sg V-pl
VP-pl → V-pl	VP-pl → V-pl	VP-pl → V-pl	VP-pl → V-pl
VP-sg → V-sg NP-pl	VP-sg → NP-pl V-sg	VP-sg → V-sg NP-pl	VP-sg → NP-pl V-sg
VP-sg → V-sg NP-sg	VP-sg → NP-sg V-sg	VP-sg → V-sg NP-sg	VP-sg → NP-sg V-sg
VP-sg → V-sg	VP-sg → V-sg	VP-sg → V-sg	VP-sg → V-sg

Shared Noise Rules	**Shared Terminal Rules**
S $\dashrightarrow$ NP-pl VP-pl	NP-pl → np-pl
S $\dashrightarrow$ VP-pl NP-pl	NP-sg → np-sg
S $\dashrightarrow$ VP-pl	V-pl → v-pl
S $\dashrightarrow$ NP-sg VP-sg	V-sg → v-sg
S $\dashrightarrow$ VP-sg NP-sg	
S $\dashrightarrow$ VP-sg	
VP-pl $\dashrightarrow$ NP-pl V-pl	
VP-pl $\dashrightarrow$ NP-sg V-pl	
VP-pl $\dashrightarrow$ V-pl NP-pl	
VP-pl $\dashrightarrow$ V-pl NP-sg	
VP-pl $\dashrightarrow$ V-pl	
VP-sg $\dashrightarrow$ NP-pl V-sg	
VP-sg $\dashrightarrow$ NP-sg V-sg	
VP-sg $\dashrightarrow$ V-sg NP-pl	
VP-sg $\dashrightarrow$ V-sg NP-sg	
VP-sg $\dashrightarrow$ V-sg	

Table 5: All Agreement Grammars

VO Core Rules	**OV Core Rules**	**SV Core Rules**	**VS Core Rules**	**Free Core Rules**
S $\rightarrow$ NP VP	S $\rightarrow$ NP VP	S $\rightarrow$ NP VP	S $\rightarrow$ VP NP	S $\rightarrow$ NP VP
S $\rightarrow$ VP NP	S $\rightarrow$ VP NP			S $\rightarrow$ VP NP
VP $\rightarrow$ V NP	VP $\rightarrow$ NP V	VP $\rightarrow$ NP V	VP $\rightarrow$ NP V	VP $\rightarrow$ NP V
		VP $\rightarrow$ V NP	VP $\rightarrow$ V NP	VP $\rightarrow$ V NP
VP $\rightarrow$ V	VP $\rightarrow$ V	VP $\rightarrow$ V	VP $\rightarrow$ V	VP $\rightarrow$ V

SVO Core Rules	**SOV Core Rules**	**VOS Core Rules**	**OVS Core Rules**
S $\rightarrow$ NP VP	S $\rightarrow$ NP VP	S $\rightarrow$ VP NP	S $\rightarrow$ VP NP
VP $\rightarrow$ V NP	VP $\rightarrow$ NP V	VP $\rightarrow$ V NP	VP $\rightarrow$ NP V
VP $\rightarrow$ V	VP $\rightarrow$ V	VP $\rightarrow$ V	VP $\rightarrow$ V

Shared Noise Rules	**Shared Terminal Rules**
S $\dashrightarrow$ NP VP	NP $\rightarrow$ np
S $\dashrightarrow$ VP NP	V $\rightarrow$ v
S $\dashrightarrow$ VP	
VP $\dashrightarrow$ NP V	
VP $\dashrightarrow$ V NP	
VP $\dashrightarrow$ V	

Table 6: All No-Agreement Grammars