Article

Acquiring Island Constraints through Efficient Structural Chunks

Niels Dickson [1]*,Lisa Pearl [2]* and Richard Futrell [2]

[1] Louisiana State University Health Sciences Center, Shreveport; nwd001@lsuhs.edu
[2] University of California, Irvine; lpearl@uci.edu, rfutrell@uci.edu
* Correspondence: [nwd001@lsuhs.edu, lpearl@uci.edu]

Abstract

How do children acquire knowledge of syntactic islands? Cross-linguistically, we see constrained variation (Sprouse et al., 2021), suggesting an interplay between child-internal factors and language-specific input. We use computational cognitive modeling to investigate a recent theory of the learning mechanism which relies on the child learning efficient chunks of syntactic structure from the input (Dickson, 2025; Dickson et al., 2024, 2022). Specifically, we adapt the Fragment Grammar (**FG**) chunking approach (T. O'Donnell et al., 2011; T. J. O'Donnell, 2015; T. J. O'Donnell et al., 2009) to syntactic islands, where the modeled child identifies both (i) chunks of hierarchical structure, and (ii) the probabilities of the learned chunks, in order to maximize the probability of the input. Following Dickson et al. (2022), the modeled child learns from cognitively-plausible input (realistic distribution and quantity of utterances) drawn from the CHILDES Treebank (L. Pearl & Sprouse, 2013). The modeled child is evaluated on its ability to replicate empirical data indicating human knowledge of syntactic islands (De Villiers et al., 2008; Liu et al., 2022; Sprouse et al., 2012), thereby demonstrating if the modeled child has acquired the relevant syntactic island knowledge. The FG-using modeled child performs better than several comparison modeled children as well as a current large language model, thus supporting the FG-chunk-based theory of acquisition. We discuss limitations and the potential of this efficient-chunking theory to explain the acquisition of additional empirical data, as well as implications for the relationship between acquisition and cross-linguistic variation with respect to syntactic islands.

Keywords: language acquisition; computational cognitive modeling; syntactic islands; structural chunking; efficiency; English

1. Syntactic islands and acquisition

Consider the English *wh*-questions in (1): (1a) and (1b) seem acceptable, while (1c) seems far less acceptable (Sprouse et al., 2012).

(1) a. What does Jack think $\_\__{what}$ is expensive?
b. Who does Jack think the necklace is for $\_\__{who}$?
c. * Who does Jack think the necklace for $\_\__{who}$ is expensive?

One explanation for this difference is that the *wh*-dependency in (1c) crosses a "syntactic island" (Ross, 1967), a latent structure that English speakers generate when processing this utterance. The island metaphorically has "no way off", so island-crossing *wh*-dependencies are typically found to be much less acceptable than non-island-crossing *wh*-dependencies

Received:
Revised:
Accepted:
Published:

https://doi.org/10.3390/1010000

(see e.g., Sprouse et al., 2012). Since Ross's seminal work on syntactic islands, there have been hundreds of articles across many languages trying to understand the nature of the observed constraints on *wh*-dependencies (see Boeckx (2012) for a valuable synthesis of much of the major debates, and Cuneo and Goldberg (2023), Momma and Dillon (2023), A. Goldberg et al. (2024), Winckel et al. (2025), and Matchin et al. (2025) (among others) for recent discussion).

Notably, much of this past work has been dedicated to understanding the adult representation of the observed language behavior. Less attention has focused on concrete theories for the acquisition of syntactic island knowledge. An influential generativist theory suggests that islands can be decomposed into building blocks (e.g., "bounding nodes" in the theory of Subjacency: Chomsky (1981); Chomsky et al. (1973); Huang (1982); see Boeckx (2012) for more recent adaptations); children must then have built-in knowledge about the inventory of potential island building blocks for human languages. Then, children learn which building blocks comprise islands for their language. How children learn which specific island building blocks are appropriate for their language is a process typically discussed less.

Here, we offer a concrete proposal for an alternative acquisition theory that draws inspiration both from this generativist approach as well as from usage-based approaches that leverage the statistical information available in children's input (A. E. Goldberg, 2006; McCauley & Christiansen, 2019; Tomasello, 2001). More specifically, the acquisition theory we implement here relies on children discovering the appropriate syntactic island building blocks (in line with the generativist approach) from the statistics of their input (in line with the usage-based approach), assuming some prior knowledge of the syntactic structure underlying observable utterances. Importantly, any theory must be able to account for the incredible efficiency of child acquisition in the face of what appear to be severe data ambiguity issues (a problem noted extensively in the research community as the "Poverty of the Stimulus": see L. Pearl (2022) for a recent overview).

Table 1 illustrates one aspect of the data ambiguity issue in children's input for learning about syntactic islands, drawing from a sample of English child-directed speech from the CHILDES Treebank (L. Pearl & Sprouse, 2013). In particular, the input is dominated by structurally-simple questions – five question types make up over 50% of the *wh*-dependency input (see L. Pearl and Sprouse (2013)) and L. Pearl and Bates (2022a) for similar findings about how skewed children's input seems to be). Thus, the data available are often ambiguous when it comes to the more complex *wh*-dependencies – how are children to know which are allowed (like (1a)) and which aren't (like (1c)), when neither type reliably occurs in their input? Children nonetheless reliably figure it out.

Example *Wh*-Dependency	Count	Percent of Stimuli	Cumulative Percent
What's that?	3,704	29.2%	29.2%
Who's that?	1,502	11.8%	41.0%
What are you doing?	696	5.5%	46.5%
What did you do?	466	3.7%	50.1%
What was that?	264	2.1%	52.2%

Table 1. Example *wh*-dependencies from the 5 most common *wh*-dependency types extracted from a sample of 12,704 child-directed *wh*-dependencies from the CHILDES Treebank

.

Moreover, children seem to be very data-efficient, even when compared to recent advances in language modeling. We now have large language models that produce language closely resembling human language (Futrell & Mahowald, 2025). However, these models require hundreds of billions of words to learn from. In contrast, children seem to accomplish much the same with less than 100 million words to work with (Warstadt et al., 2023). Children's remarkable data-efficiency is even more surprising because of the known cognitive limitations that children have (Behm et al., 2025; Fandakova et al., 2014; Gathercole et al., 2004; Paris, 1978). That is, children's ability to extract information from their input is immature, and so their *intake* can be quite different from the available *input* (L. Pearl, 2023a, 2023b).

Here, we investigate an acquisition theory for islands that can succeed under these conditions – that is, an acquisition theory capable of learning relevant syntactic islands knowledge as data-efficiently as children do. We begin by motivating the efficient-chunking approach at the heart of this acquisition theory, and reviewing chunking as a cognitive strategy, as well as prior chunking approaches in syntactic acquisition. We then discuss the implementation of the efficient-chunking strategy that we evaluate here, which relies on Fragment Grammars (**FGs**) (T. O'Donnell et al., 2011). We turn then to the specification of the FG-based modeled child's acquisition task, including the input, the intake, and the target behavior signaling acquisition of syntactic island knowledge.

We evaluate this FG-based modeled child by how well it can generate the target behavior, given realistic input, and compare its performance against several other modeled children as well as a neural network model whose architecture underlies high-performing large language models. We find that the FG-based modeled child can generate all target behavior patterns, and does so better than every other comparison modeled child and model. These results both support our proposed efficient-chunk-based acquisition theory, and highlight its superior performance for acquiring syntactic island knowledge from realistic input. We discuss key underlying assumptions of the current FG-based implementation of the efficient-chunking acquisition theory, and their potential impact on our findings here. We conclude by considering future directions about acquiring broader knowledge of *wh*-dependencies and the implications of our findings for the relationship between acquisition and constrained cross-linguistic variation of syntactic island constraints.

2. Efficient chunking for syntactic islands

2.1. Why chunking?

From a human cognition standpoint, a chunk is a series of units grouped together, typically based on some efficiency consideration (Chase & Simon, 1973; Miller, 1956; Ramkumar et al., 2016; Rosenbloom & Newell, 1982; Thalmann et al., 2019). Humans are limited-resource agents (Lieder & Griffiths, 2020), and so much of our cognitive success relies on efficiently organizing the input into units that are both useful and compact, such as with structured sequential information in the domains of language, vision, and motor planning (Ding, 2025). For instance, in motor planning e.g., speech articulation), deploying a chunk of action sequences that often occur together is more energy-efficient than planning each individual muscle twitch (Derrick et al., 2024; Ramkumar et al., 2016). This efficiency results because smooth, combined motions (i.e., motor chunks) expend less energy than jerky motions where each movement is planned separately. Moreover, planning a long sequence of actions expends computational energy in considering all possible ways to group individual actions. Chunked actions therefore require less computation to plan.

More specifically, useful chunks lead to "savings" on future input that can be broken into those useful chunks (e.g., in speech segmentation: M. C. Frank et al. 2010; Jessop et al. 2025; Perruchet et al. 2014; Perruchet and Vinter 1998). For syntactic acquisition, multi-word chunks appear to be a key unit of representation (Arnon, 2021; Arnon & Clark, 2011; A. E. Goldberg, 1995) and sensitivity to frequently-reused language chunks has been linked to success in second-language acquisition (Pulido, 2021). With this in mind, we turn now to previous models of syntactic acquisition that incorporate chunking.

2.2. *Previous models of syntactic chunking*

Two notable chunk-based acquisition theories for syntax involve the child creating useful multi-word chunks from the input, based on input processing considerations (Freudenthal et al., 2024; McCauley & Christiansen, 2019). In the Model of Syntax Acquisition in Children (**MOSAIC**) (Freudenthal et al., 2024, 2006, 2015), the modeled child tracks how often a multi-word sequence is encountered in the input – if the sequence occurs often enough, the sequence becomes a multi-word chunk (where individual words within that chunk can't be substituted). Similarly, the Chunk-Based Learner (**CBL**) (McCauley & Christiansen, 2019) creates multi-word chunks out of bigrams whose transitional probability is sufficiently high. Notably, the CBL-modeled child also reuses these chunks to process future input and decide which future sequences ought to be chunked. Both the MOSAIC and CBL modeled learners align with cross-linguistic child comprehension and production data. Notably, these chunking approaches were evaluated on their ability to parse and generate child language data in general (i.e., naturally-occurring utterances in child language interactions).

In contrast, the syntactic chunking approach of L. Pearl and Sprouse (2013) aimed specifically to explain the acquisition of syntactic island knowledge, as measured by behavior in controlled experiments. The modeled child's chunks incorporated syntactic structure, and were pre-specified as three units in size (i.e., trigrams). The units themselves primarily consisted of phrase structure pieces, such as "verb phrase" (VP), though the chunks could also include information about a single lexical item type (complementizers like "that"). The modeled child determined the relative frequency of the syntactic trigrams in its chunk inventory, based on the input, and then used these chunks to distinguish between *wh*-dependencies that crossed syntactic islands from those that didn't. The modeled child's output aligned with English *wh*-dependency judgments (see also L. Pearl and Bates (2022b)), suggesting these syntactic trigram chunks could explain acquisition of some syntactic island knowledge. From a theoretical standpoint, this means that "knowledge of syntactic islands" is distributed across the modeled child's inventory of chunks (and their resulting potential combinations). That is, the knowledge of any given syntactic island emerges from the child's inventory of syntactic trigram chunks and their associated probabilities.

In the syntactic trigrams approach, the efficiency of a chunk was captured via its probability, with higher-probability chunks yielding higher-probability parses of *wh*-dependencies. In other words, input that could be processed with higher-probability chunks was itself given a higher probability. One way to interpret input with a higher probability is as input that's easier (more efficient) to process, and that's the same link we'll use in our proposed chunking acquisition theory. That is, efficient chunks make the input have a higher probability for the modeled child. So, the modeled child's goal is to find efficient chunks that explain the input, i.e., give the input in general a high(er) probability.

2.3. *Finding efficient chunks: The Fragment Grammar learner*

We adapt and expand prior chunk-based approaches in order to explain more empirical data on the acquisition of syntactic islands. Like all prior chunking approaches, the child we model is looking for chunks that explain both current and future input with high probability. Like L. Pearl and Sprouse (2013), the child we model perceives syntactic structure in the input – i.e., the modeled child's intake includes phrase structure, as in Figure 1. A key difference for our modeled learner is flexibility in the nature of the chunks, in terms of both the units that comprise the chunks and how many units can be involved in a chunk.

More specifically, the units forming the chunks can be any phrase structure node or lexical item available in the input representation (instead of only words (Freudenthal et al., 2024; McCauley & Christiansen, 2019) or only phrase structure nodes with a single lexical item type (L. Pearl & Sprouse, 2013)). In addition, the chunks are free to vary in size, ranging from a single lexical item to the structure of an entire utterance (see Figure 1's minimal, intermediate, and maximal learner chunk representations).

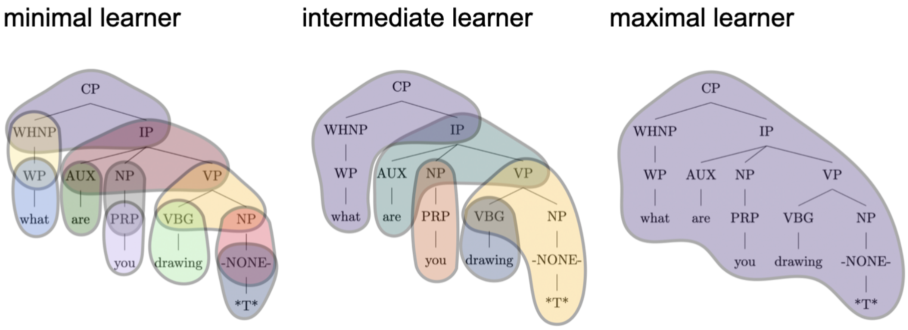


Figure 1. Potential chunk representations of the *wh*-dependency "What are you drawing?". The modeled child receives syntactic structure for the dependency, which is then chunked in different ways. The minimal learner (left) selects the smallest "chunks" available, effectively not chunking. The intermediate learner (center) makes several chunks of different sizes. The maximal learner (right) makes one chunk out of the entire *wh*-dependency structure.

We implement a modeled child's potential inventory of chunks using Fragment Grammars (**FGs**) (T. O'Donnell et al., 2011)), where a potential inventory (i.e., "grammar") of chunks (i.e., "fragments") has some probability assigned to each chunk. The FG-based modeled child searches the space of possible chunk inventories (i.e., FGs), using the input to identify which chunk inventories yield a high probability for the input data. The intuition for this search is similar to prior chunking approaches: if certain parts of the input appear together frequently, the modeled child will chunk these parts together.

The modeled child searches the hypothesis space of possible chunk inventories by using Bayesian inference, a principled reasoning approach that accords well with empirical data on human cognition, including language acquisition (e.g., Dowman, 2000; Feldman et al., 2013; Foraker et al., 2009; S. Frank et al., 2013; Goldwater & Griffiths, 2007; Griffiths et al., 2024; Gutman et al., 2015; Harmon et al., 2021; Kwiatkowski et al., 2012; T. J. O'Donnell, 2015; L. Pearl et al., 2010; L. S. Pearl & Mis, 2016; L. S. Pearl & Sprouse, 2019; Perfors et al., 2011; Perkins et al., 2017; Phillips & Pearl, 2015, among many others) – see L. S. Pearl (2021) for a recent overview. In particular, the modeled child uses Bayesian inference to balance how well the chunks explain the input (by giving the input a high probability) with how "simple" the chunk inventory is (e.g., the size of the chunk inventory: Chater and Vitányi 2007). The smaller the chunk inventory, the more space-efficient the chunk inventory is. So, the modeled child is searching for a space-efficient chunk inventory that can still give the input a high probability.

We can see how this balance would play out in the sample chunk representations in Figure 1, whose characteristics are summarized in Table 2. The maximal learner (Figure1: right) creates one chunk per *wh*-dependency type, and so gives any particular *wh*-dependency a very high probability. However, its chunk inventory is very large (as large as the number of *wh*-dependency types in the input), and so it's not very space-efficient.

In contrast, the minimal learner (Figure1: left) creates only tiny "chunks" (the smallest units possible), and so has a fairly space-efficient inventory. However, because no larger chunks are available, any "efficiency savings" from frequently-appearing structural sequences are lost. Every structural piece must be constructed every single time from the tiny chunks. So, the probability of the input will be lower.

In contrast to the minimal learner, the intermediate learner (Figure1: center) leverages frequently-appearing structural sequences to make larger chunks, in addition to the smallest "chunk" units that the minimal learner has. While the intermediate learner's chunk inventory is therefore larger (less space-efficient) than the minimal learner's, the intermediate learner's probability of the input data is higher – this is because the chunks, comprised of frequently-appearing structural sequences, yield higher probability for a given structural sequence than the combination of smaller units that comprise that structural sequence. In other words, the minimal learner must rely on combining the smaller units every time, while the intermediate learner can use the chunk. So, the intermediate learner yields a higher probability for the input than the minimal learner does, even though it has a less space-efficient chunk inventory than the minimal learner does. Bayesian inference allows the modeled child to find the chunk inventory that strikes the best balance between these two factors (input probability and space efficiency).

We note that the modeled child is using a computational-level implementation of Bayesian inference to approximate the mental computation children perform. More specifically, the modeled child samples a potential chunk inventory, and uses that chunk inventory to analyze the input. Then, the modeled child begins the following cycle: (i) sample a different potential chunk inventory, (ii) analyze the data with that inventory, (iii) adopt the new chunk inventory if the data have a higher probability with that inventory than with the previous chunk inventory. This cycle is repeated until the modeled child identifies a chunk inventory that yields a high probability for the input (and in particular, no other sampled chunk inventory yields a higher probability).

To be clear, we don't assume that children are capable of accomplishing this mental computation of Bayesian inference in this way – for one thing, it seems unlikely they can hold a detailed representation of all their input data over many years in mind. However, we *are* committed to children performing Bayesian inference, likely approximating this mental computation as best they can with the cognitive resources they have available. This is why we use "computational-level" to describe the modeled learners: we believe children perform the mental computation of Bayesian inference, but not necessarily using the algorithm the modeled children here use.

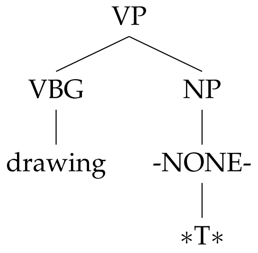
Chunk inventory type	**Space-efficient?**	**High input probability?**
Minimal	Yes	Somewhat
Intermediate	Somewhat	Yes
Maximal	No	Yes

Table 2. Characteristics of chunking strategies (minimal, intermediate, and maximal chunking), in terms of whether the chunk inventories are space-efficient and also yield a high probability for the modeled child's input data.

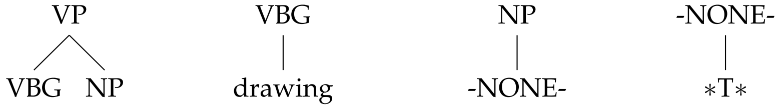
3. The FG modeled child: Implementation

The FG modeled child considers chunk inventories (i.e., Fragment Grammars) that are a type of a Probabilistic Context-Free Grammar (**PCFG**). A chunk can be represented as a rule that specifies how one unit can expand into other units. For example, the structure in (2) can be expressed with several PCFG chunks (3), which can be represented with the rules also shown in (3).

(2)



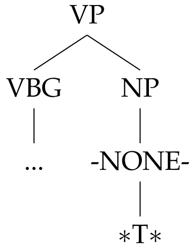
(3) PCFG chunks and rules



VP → VBP NP   VBG → drawing   NP → -NONE-   -NONE- → *T*

The FG chunk inventories differ from those that a PCFG can consider in two key ways. First, the FG allows larger chunks, as opposed to the minimal chunks (i.e., those that the minimal learner of Figure 1 uses). For example, the FG modeled child can consider the full VP chunk in (2), which might be represented in a rule as something like VP → (VBG → *drawing*) (NP → (-NONE- → *T*)). See Appendix A for the notation we implemented to capture FG chunks like this. The mathematical implementation that allows this larger-chunking process (known as an Adaptor Grammar: Johnson et al. 2007; T. J. O'Donnell 2015) uses a Pitman-Yor process (Pitman & Yor, 1997); notably, this implementation can only consider chunks that are fully expanded to the leaves. So, for example, a modeled child using this process could consider the chunk in (2) but not the chunk in (4), because the chunk in (4) leaves VBG unexpanded. See Appendix A.2 for details.

(4)



In contrast, a FG chunk inventory can include chunks like (4) that involve unexpanded nodes. For example, the chunk in (4) can be represented as something like VP → VBG (NP → (-NONE- → *T*)). This is a chunk that the intermediate learner of Figure 1 uses, which allows the specific verb to vary while keeping the *wh*-object position in the VP chunk. See Appendix A.3 for the "lazy evaluation" mathematical implementation of the Pitman-Yor process that allows this chunk option (T. J. O'Donnell, 2015).

Once the modeled child has inferred an efficient chunk inventory (and each chunk's associated probability) from the input, the modeled child can then generate a probability for any structure that can be comprised of the chunks available in that inventory. Here, we assume a structure's probability is the product of the chunks that comprise it (i.e., $p(structure) = \prod_{chunk\ c_i \in structure} p(c_i)$).

4. The FG modeled child's acquisition task

4.1. Input

The FG modeled child receives a realistic sample of child-directed *wh*-dependencies, derived from a distribution of 12,704 *wh*-dependencies from the CHILDES Treebank (L. Pearl & Sprouse, 2013). We follow L. Pearl and Bates (2022a) and estimate the total number of *wh*-dependencies that children encounter by considering their potential learning period, the average waking hours of a child at different ages (Davis et al., 2004), the utterances heard per hour (Rowe, 2012), and the relative frequency of *wh*-dependencies in children's input (see Table 3). We consider the learning period to start at 18 months when children seem capable of reliably recognizing *wh*-dependencies in their input (Perkins & Lidz, 2021). We consider the learning period to end at age 4, when children seem to demonstrate adult-like knowledge of several syntactic islands (De Villiers et al., 2008) (though there is evidence for adult-like knowledge of some islands even younger: Hirzel 2022).

min in learning period = waking hours / 60	utt/min	*wh*-dep/utt	= total *wh*-dep estimated in learning period
886,950	14.4	0.164	2,094,753

Table 3. Calculation of the *wh*-dependencies children encounter during the proposed learning period from 18 months to 4 years old. Values are derived from L. Pearl and Bates (2022b), which estimated the total number of *wh*-dependencies that children encounter by considering the average waking hours (Davis et al., 2004) per age, utterances encountered per minute (Rowe, 2012), and the proportion of *wh*-dependencies in the utterances in children's input.

The resulting estimate for the quantity of *wh*-dependencies children encounter is a little over 2 million (2,094,753, out of 2,094,753/0.164 = 12,772,884 utterances). The modeled child sees this quantity of *wh*-dependencies, distributed according to the sample of child-directed *wh*-dependencies from the CHILDES Treebank. As noted above in Table 1, the majority of *wh*-dependencies are of only a few types (though see (L. Pearl & Bates, 2022a; L. Pearl & Sprouse, 2013) for the complete distribution of *wh*-dependency types in different samples of the CHILDES Treebank).

4.2. A modeled child's intake

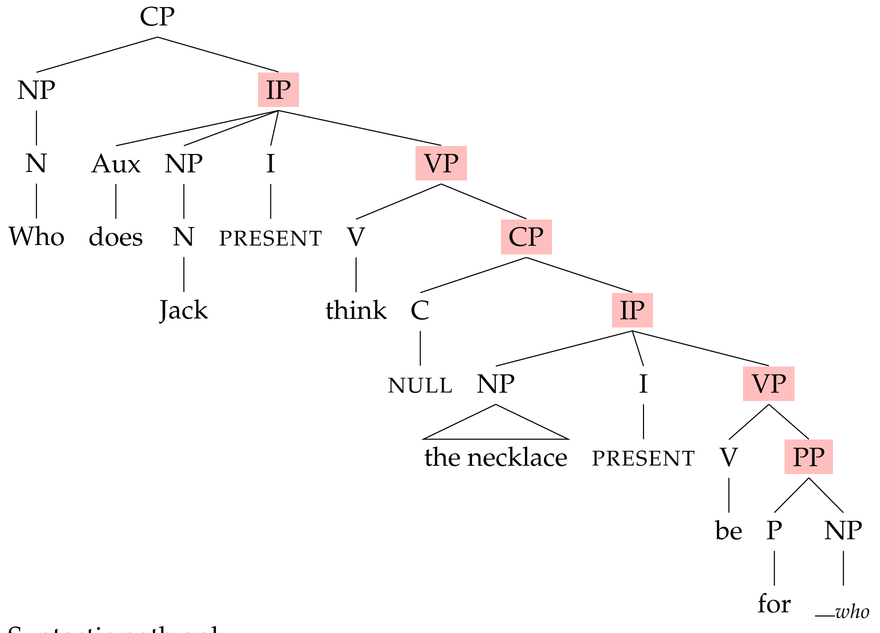
We follow L. Pearl and Sprouse (2013), and assume the modeled child's intake is a filtered subset of the available input (L. Pearl, 2023b). More specifically, the modeled child focuses on utterances containing *wh*-dependencies, ignoring other utterances. Then, for a given *wh*-dependency, the modeled child projects syntactic structure onto the *wh*-dependency, given its prior syntactic knowledge (see (5) for the *wh*-dependency "Who does Jack think the necklace is for?"[1]). With this syntactic structure in place, the modeled child then focuses on a subset of the available structure, which is the "syntactic path" connecting the *wh*-word to its gap, as in (5a).

More formally, L. Pearl and Sprouse (2013) define this syntactic path as the set of phrase structure nodes that contain the gap, until the phrase structure node that is parent to the *wh*-phrase is reached. In (5a), the gap $\_{who}$ is contained by PP, which is contained by VP, and so on, until IP is reached. Then, the parent of IP is CP, which is the parent of the *wh*-phrase. The syntactic path can be represented by a portion of the actual structure available, as in (5b), focusing on the phrase structure nodes and their accompanying heads in the syntactic path. A flattened version of the syntactic path can be represented as a sequence, such as $\text{IP}_{\textsc{present}}$-$\text{VP}_{think}$-$\text{CP}_{\textsc{null}}$-$\text{IP}_{\textsc{present}}$-$\text{VP}_{be}$-$\text{PP}_{for}$ for (5a).

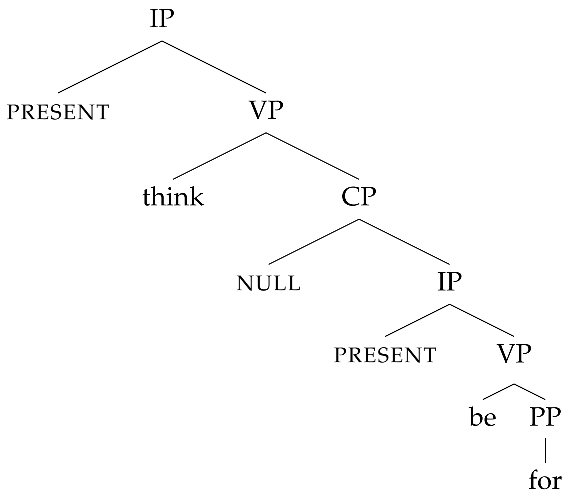
(5) Syntactic path representation for "Who does Jack think the necklace is for?"

[1] Our implementation assumes a particular syntactic phrase structure representation, but a syntactic path can be defined for any syntactic tree structure.

a. Projected syntactic structure and syntactic path



b. Syntactic path only



syntactic path sequence:
$\text{IP}_{\text{PRESENT}}$-$\text{VP}_{think}$-$\text{CP}_{\text{NULL}}$-$\text{IP}_{\text{PRESENT}}$-$\text{VP}_{be}$-$\text{PP}_{for}$

4.3. Modeled child target behavior

Linguistic knowledge is typically assessed by mapping observable behavior to underlying knowledge; we thus review three behavior patterns that serve as signals of syntactic island knowledge and so function as the target of acquisition for the modeled child here (L. Pearl, 2023b). The first two are acceptability judgment patterns for different *wh*-dependencies, while the third is an interpretation preference pattern for utterances ambiguous between two possible *wh*-dependencies.

4.3.1. Adult judgment data

The first pattern is a superadditive acceptability judgment pattern, as shown in the interaction plot in Figure 2 – the superadditive pattern itself appears as non-parallel lines (L. Pearl & Sprouse, 2013; Sprouse et al., 2012).

The judgment pattern arises from constructed stimuli sets like (6) that vary two factors: the length of the *wh*-dependency (matrix clause vs. embedded clause) and the absence/presence of a proposed island structure (non-island vs. island) (Sprouse et al.,

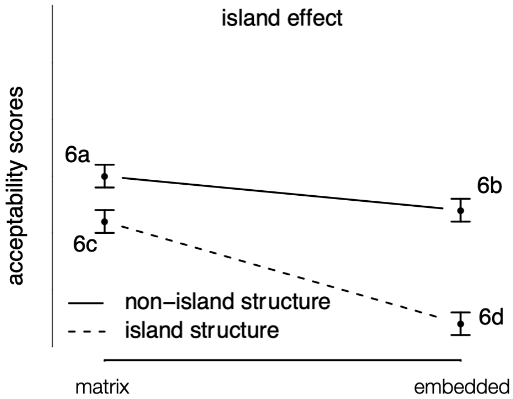


Figure 2. An interaction plot showing a pattern in the acceptability of *wh*-dependency judgments, which appears as non-parallel lines. The figure is annotated with examples of the four conditions from example (6).

2012). Examples (6a) and (6b) vary by the length of the dependency, but neither has an island structure. Square brackets surround the proposed island structure in (6c)-(6d), and the *wh*-dependency in (6d) crosses the island structure.

(6) a. Who __$_{who}$ thinks the necklace is expensive? MATRIX | NON-ISLAND
b. What does Jack think __$_{what}$ is expensive? EMBEDDED | NON-ISLAND
c. Who __$_{who}$ thinks [$_{CP}$ [$_{IP}$ [$_{NP}$ the necklace for Lily is expensive]]]? MATRIX | ISLAND
d. *Who does Jack think [$_{CP}$[$_{IP}$[$_{NP}$ the necklace for __$_{who}$] is expensive]]? EMBEDDED | ISLAND

Importantly, each factor is associated with a decrease in acceptability. First, an embedded dependency is less acceptable than a matrix dependency (length: (6b) is less acceptable than (6a) by some amount, *len*). Second, an utterance with an island structure in it is less acceptable than an utterance without one (absence/presence: (6c) is less acceptable than (6a) by some amount, *isl*). An additive effect for an embedded dependency with an island structure in it (6d) would be the simple addition of these two decreases in acceptability: (6d) = (6a) - (*len* + *isl*). So, there would be no interaction and Figure 2 would show parallel lines. A superadditive effect for an embedded dependency with an island structure in it (6d) would be an extra decrease beyond the simple addition of the two decreases due to length and presence of an island structure: (6d) = (6a) - (*len* + *isl* + *extra*). This extra decrease for (6d) causes the interaction plot to show non-parallel lines.

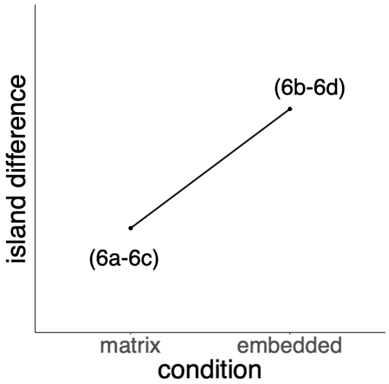


Figure 3. An acceptability difference plot showing the superadditive acceptability judgment pattern as a difference between the acceptability of the non-island and island structures, varying the factor of length. The difference is greater for the embedded *wh*-dependencies, which appears as a line with a positive slope on this plot between matrix and embedded stimuli.

https://doi.org/10.3390/1010000

This superadditive pattern can also be summarized by plotting the difference in acceptability between the non-island and island structures, as in Figure 3. In particular, the superadditive pattern is observed when the difference between the matrix *wh*-dependencies (rating(6a) - rating(6c)) is less than the difference between the embedded *wh*-dependencies (rating(6b) - rating(6d)). That is, the "island difference" is how the matrix *wh*-dependency difference compares to the embedded *wh*-dependency difference, with the idea that the increase in difference comes from the presence of the island-crossing dependency in examples like (6d). When the matrix difference is less than the embedded difference, we see a positive slope, as in Figure 3. So, a positive slope on this kind of "island difference plot" indicates a superadditive acceptability judgment pattern, which signals knowledge of syntactic islands (such as the one in (6d)). This qualitative pattern – the positive slope on an island difference plot – is the target output for the modeled child here, when given *wh*-dependency stimuli sets like (6).

The second target pattern involves an observed effect of lexical item frequency in adult acceptability judgments of *wh*-dependencies. In particular, Liu et al. (2022) found a positive correlation between the frequency of a certain *wh*-dependency's main verb and that *wh*-dependency's acceptability, as shown in in Figure 4. The x-axis of Figure 4 shows the log-transformed frequency of the main verb appearing in the linguistic context where the verb is followed by an embedded clause (e.g., "...*say/whine* that Jack saw.") – this is the verb frame.

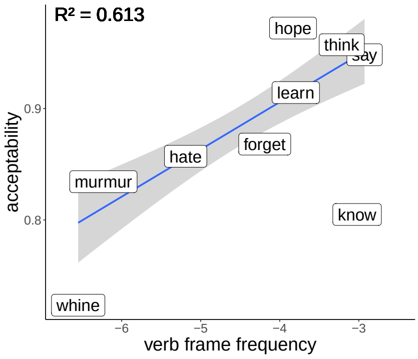


Figure 4. Behavioral data reported in Liu et al. (2022) showing a positive correlation between a *wh*-dependency's acceptability and the frequency of the main verb in a specific linguistic context (e.g., "What did Lily VERB that Jack saw?"). *Wh*-dependencies with verbs like "say" in this frame are rated as more acceptable than *wh*-dependencies with verbs like "whine".

We can see that as the verb frame's frequency increases, the *wh*-dependency's acceptability increases. For example, a dependency with a high-frequency main verb (e.g. "What did Lily *say* that Jack saw?") is judged to be more acceptable than an identical dependency with a lower-frequency main verb (e.g. "What did Lily *whine* that Jack saw?"). This positive correlation is the target behavior for the modeled child, representing this lexical effect on *wh*-dependencies.

4.3.2. Child judgment data

The third target pattern is child preferences for utterances that are potentially ambiguous between two *wh*-dependencies as in (7), derived from the behavioral data collected by De Villiers et al. (2008).

(7) Who did the police woman help to call?

a. *wh*-dependency 1 (main): Who did the police woman help $\_\_{\mathit{who}}$ [to call]?
Interpretation 1: Who was helped?

b. *wh*-dependency 2 (embedded): Who did the police woman help [to call $\_\mathit{who}$]?
Interpretation 2: Who was called?

In (7), the two logically possible interpretations correspond to different *wh*-dependencies, a main-clause one (7a) and an embedded-clause one (7b) (the embedded clause is indicated with [...]). Children then indicated which interpretation they preferred, and so which *wh*-dependency they preferred. Table 4 shows the full list of *wh*-utterance types used in De Villiers et al. (2008), and how often children preferred the embedded *wh*-dependency for each type, ordered by how often the embedded option was preferred.

Item	Question	Child emb pref
1	How did the boy say $\_\mathit{how}_{main}$ [he hurt $\_\mathit{how}_{emb}$ himself] ?	0.80
2	What did the mother say $\_\mathit{what}_{main}$ [she bought $\_\mathit{what}_{emb}$] ?	0.79
3	Who did the police woman help $\_\mathit{who}_{main}$ [to call $\_\mathit{who}_{emb}$] ?	0.48
4	Who did the little sister ask $\_\mathit{who}_{main}$ [how to see $\_\mathit{who}_{emb}$] ?	0.25
5	How did [the boy who sneezed $\_\mathit{how}_{emb}$] drink $\_\mathit{how}_{main}$ the milk?	0.20
6	What did the boy fix the cat [that was lying on the table with $\_\mathit{what}_{emb}$] $\_\mathit{what}_{main}$?	0.09
7	How did the girl ask $\_\mathit{how}_{main}$ [where to ride $\_\mathit{how}_{emb}$]?	0.04
8	Who did the boy ask $\_\mathit{who}_{main}$ [what to bring $\_\mathit{who}_{emb}$]?	0.04
9	How did the mom learn $\_\mathit{how}_{main}$ [what to bake $\_\mathit{how}_{emb}$]?	0.03

Table 4. Stimuli from De Villiers et al. (2008) used to probe child dependency preferences. Each question has a main and embedded clause (embedded clauses are marked in [...]), and is potentially ambiguous between a main-clause *wh*-dependency (wh_{main}) and an embedded-clause *wh*-dependency (wh_{emb}). Child preferences for the embedded-clause *wh*-dependency are shown, with items ordered by embedded-dependency preference.

As the target of acquisition, the modeled child will aim to generate these same preferences, when given the two *wh*-dependency options to choose between (i.e., the main-clause one vs. the embedded-clause one). Figure 5 plots the child preferences from De Villiers et al. (2008) against a modeled child able to perfectly reproduce those embedded-dependency preferences for each item.

5. Evaluating the FG modeled child

After learning from the input, the FG modeled child can generate a probability for any *wh*-dependency, using the most probable (highest probability) combination of chunks available from its learned chunk inventory. More specifically, we can calculate the maximum *a posteriori* (**MAP**) score under the inferred chunk inventory (Eisenstein, 2018), based on the highest-probability parse for the given syntactic structure (here: a *wh*-dependency's syntactic path). We first discuss how we link these generated probabilities to the target behavior patterns for syntactic islands. We then present other learning approaches that are given the same acquisition task as the FG modeled child and so can serve as a basis for comparison regarding acquisition performance.

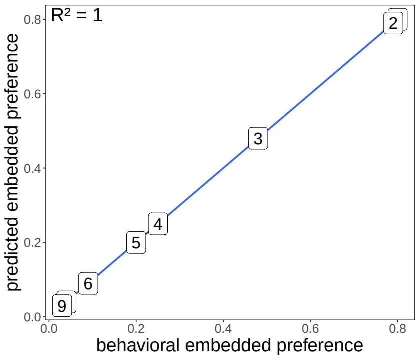


Figure 5. Example data depicting a perfect match between child preference data observed from De Villiers et al. (2008) on the x-axis and the modeled child's prediction on the y-axis for each item (1-9) in Table 4.

5.1. Linking probabilities with target behavior patterns

5.1.1. Probabilities to acceptability judgments

Recall that the structures the modeled child here considers are the syntactic paths of *wh*-dependencies, which can be rewritten as sequences. For instance, the syntactic path for "Who does Jack think the necklace is for?" (5) can be represented by the sequence $\text{IP}_{\text{PRESENT}}$-$\text{VP}_{think}$-$\text{CP}_{\text{NULL}}$-$\text{IP}_{\text{PRESENT}}$-$\text{VP}_{be}$-$\text{PP}_{for}$. When considering sequences, a sequence's probability is affected by length when the sequence's probability is calculated by multiplying together its individual pieces (i.e., chunks). The more pieces, the lower the probability, because probabilities are generally less than 1. However, we don't see this same reliable relationship between utterance length and human judgments of acceptability (Lau et al., 2015, 2017). So, utterances differing in length (e.g., "The necklace is for Lily" vs. "The necklace with the sparkly gems is for Lily" may still be comparably acceptable even though their probabilities are not (transparently) comparable.

To account for this potential effect of length, the modeled child's acceptability score for a given *wh*-dependency's syntactic path is impacted by the length of the *wh*-dependency, following Lau et al. (2015, 2017). More specifically, the modeled child's *length-factorized* score is the (log) probability for the *wh*-dependency's syntactic path divided by the length of the path, as in So, for the example syntactic path above from (5), the length would be 6 as there are 6 units in the sequence.

$$\text{(8)} \qquad \text{length-factorized(syn-path)} = P_{len\_fac}(\text{syn-path}) = \frac{\text{log(prob(syn-path))}}{\text{length(syn-path)}}$$

For the acceptability judgments from both Sprouse et al. (2012) and Liu et al. (2022), the modeled child calculates the length-factorized score for each of the stimuli in a given set. We can then assess if superadditivity is present for the data from Sprouse et al. (2012) and if a positive correlation is present for the data from Liu et al. (2022).

5.1.2. Probabilities to interpretation preferences

To generate an interpretation preference for a potentially-ambiguous *wh*-utterance from the data of De Villiers et al. (2008) in Table 4, the modeled child needs to generate a score for each *wh*-dependency associated with a potential interpretation of that utterance (i.e., the main-clause *wh*-dependency vs. the embedded-clause *wh*-dependency). We follow prior work linking probabilities to production frequencies (Mayer, 2021), and again use the length factorization in (8) to generate a score for each *wh*-dependency. We then calculate the preference for the embedded-clause dependency by normalizing these scores to calculate the probability of an embedded-clause preference, as in (9). So, the preference ranges between 0 (strongest main-clause *wh*-dependency preference) to 1 (strongest embedded-clause *wh*-dependency preference).

(9) $$\text{Pref}_{emb} = \frac{\text{P}_{\text{len\_fac}}(\text{emb-}wh)}{\text{P}_{\text{len\_fac}}(\text{emb-}wh) + \text{P}_{\text{len\_fac}}(\text{main-}wh)}$$

5.2. Comparison: Other modeled children and models

As a comparison to the FG-modeled child, we consider a set of modeled children that vary in the nature of the chunks they consider for their chunk inventories, all of which are more constrained than the chunks the FG-based modeled child considers. Because of this ability to consider a wider hypothesis space of possible chunks, the FG-modeled child is also more complex than the comparison modeled children (i.e., it has more free parameters that enable better data-fitting). It could be that simply having more parameters enables better acquisition performance. To test this possibility, we additionally use a neural network model with many free parameters as a comparison. We review each comparison modeled child and the neural network model below.

5.2.1. Modeled children with simpler chunks

Recall from Section 3 that FGs are more-flexible version of Adaptor Grammars (**AGs**), which are more-flexible versions of PCFGs. In particular, AGs allow larger chunks but require full expansion of non-terminal nodes, causing their hypothesis space to exclude chunks like (4) (i.e., VP → VBG (NP → (-NONE- → *T*))) that leave a non-terminal node open (i.e., VBG). Thus, an AG-based modeled child considers a more-restricted space of possible chunks than the FG-based child. Similarly, a PCFG-based child has a further-restricted space of possible chunks compared to the AG-based child, because the PCFG-based child can only consider minimal chunks like those in (3) (ie., VP → VBG NP, VBG → *drawing*, etc.). Implementation details for the AG-based and PCFG-based modeled children are found in Appendix B.1.

We additionally implement modeled children with fixed-size chunks, following L. Pearl and Sprouse (2013) for some options. Prespecifying the size of the chunks contrasts with the modeled children using FGs, AGs, and PCFGs, as the size of those chunks can vary (and is learned from the input). Instead, the modeled children using fixed-size chunks specifically use trigrams (i.e., 3-unit chunks) to form the syntactic path sequence of a *wh*-dependency. The trigram-based modeled children we consider vary the amount of lexical information included in the trigrams, as in (10). The syntactic path is then broken into successive sequences of trigrams, based on the information included from the syntactic path. See Appendix B.2 for details about how trigrams combine to form a syntactic path.

(10) Comparison representations for "Who did Jack think the necklace was for?"

a. Fully-lexicalized: all lexical information included with phrase label
syn path: START-IP$_{present}$-VP$_{think}$-CP$_{null}$-IP$_{present}$-VP$_{be}$-PP$_{for}$-END
trigrams: START-IP$_{present}$-VP$_{think}$, IP$_{present}$-VP$_{think}$-CP$_{null}$, ..., VP$_{be}$-PP$_{for}$-END

b. Phrasal-only: phrase label only
syn path: START-IP-VP-CP-IP-VP-PP-END
trigrams: START-IP-VP, IP-VP-CP, ..., VP-PP-END

c. Lexicalized CP: only the head of the CP (the complementizer) is included
syn path: START-IP-VP-CP$_{null}$-IP-VP-PP-END
trigrams: START-IP-VP, IP-VP-CP$_{null}$, ..., VP-PP-END

d. Lexicalized main verb: only the main verb is included
syn path: START-IP-VP$_{think}$-CP-IP-VP-PP-END
trigrams: START-IP-VP$_{think}$, IP-VP$_{think}$-CP, ..., VP-PP-END

The modeled child using fully-lexicalized trigrams (10a) would include all the information on the syntactic path in its trigrams, both phrasal and lexical (e.g., IP$_{present}$-VP$_{think}$-CP$_{null}$). In contrast, a modeled child using only phrasal information in its trigrams (10b) ignores all lexical information (e.g., IP-VP-CP) and was considered as a baseline in L. Pearl and Sprouse (2013). L. Pearl and Sprouse (2013) found that a modeled child that included lexical information only for complementizer phrases (CPs) in its trigrams (e.g., IP-VP-CP$_{null}$) was able to capture the superadditive adult acceptability judgment pattern in Figure 2 for several island types. However, this modeled child can't explain the variation by main verb found in the positive correlation pattern from Liu et al. (2022), so we also consider a modeled child that includes lexical information only for main verbs, as in (10d) (e.g., IP-VP$_{think}$-CP).

5.2.2. A neural network model with many free parameters

We chose a long-short term memory (**LSTM**) model (Hochreiter & Schmidhuber, 1997) as the comparison neural network model because LSTMs do well when learning information about sequences (Eisenstein, 2018), such as syntactic paths. The specific implementation we use involves tens of thousands of free parameters that enable the model to fit the input data well – see Appendix B.3 for the formal definition and relevant hyperparameter settings.

The LSTM implementation we use learns to accurately predict the next token in a sequence, and so its input is a "flattened" version of the syntactic path the modeled children previously discussed use. For example, the syntactic path from (10) is START-IP$_{present}$-VP$_{think}$-CP$_{null}$-IP$_{present}$-VP$_{be}$-PP$_{for}$-END, and becomes START, IP, present, VP, think, CP, null, IP, present, VP, be, PP, for, END. The LSTM model can then generate a probability for any *wh*-dependency's syntactic path, represented in this flattened, sequential form.

6. Results

Figures 6, 7, and 8 show the results for the FG-based modeled child, as well as the comparison modeled children and LSTM model. One key broad observation is that the FG-based modeled child's performance most closely aligns with target behavior patterns, outperforming all comparison modeled children and the LSTM model (summarized in Table 5). More specifically, the FG-based modeled child can generate both the superadditive pattern from Sprouse et al. (2012) and the positive correlation pattern from Liu et al. (2022); it also has the highest correlation (R^2=0.879) with the child preference patterns from De Villiers et al. (2008). We discuss each target behavior pattern in turn.

	Chunk-based modeled children							
	Chunks of varying size			**Trigram chunks**				
	FG	AG	PCFG	Phrasal	Fully Lex	Lex CP	Lex MV	LSTM
Sprouse et al. (2012)	✓	✓	✓	✗	✗	✓	✗	✗
Liu et al. (2022)	✓	✗	✓	✗	✓	✗	✓	✓
De Villiers et al. (2008)	0.879	0.045	0.601	0.084	0.046	0.268	<0.001	0.532

Table 5. Summary of results for all chunk-based modeled children and the LSTM model across the three target behavioral patterns. A qualitative summary (✓/✗) is shown for the behavioral patterns from Sprouse et al. (2012) and Liu et al. (2022), while the correlation value between predicted vs. actual child behaviors are shown for the behavioral patterns from De Villiers et al. (2008).

6.1. *Target behavior: Superadditivity*

Recall that adults have acceptability judgments on stimuli related to different syntactic islands that pattern as a positive slope in the "islands difference score" of Figure 3. In particular, the stimuli from Sprouse et al. (2012) involved four island types: Subject, Complex NP, Whether, and Adjunct islands. A successful modeled learner (modeled child or model) would be able to generate that same pattern (a positive slope) for all four island types, using its internalized representations of *wh*-dependency knowledge. Figure 6 shows the results for all four island types. We discuss each island type in turn.

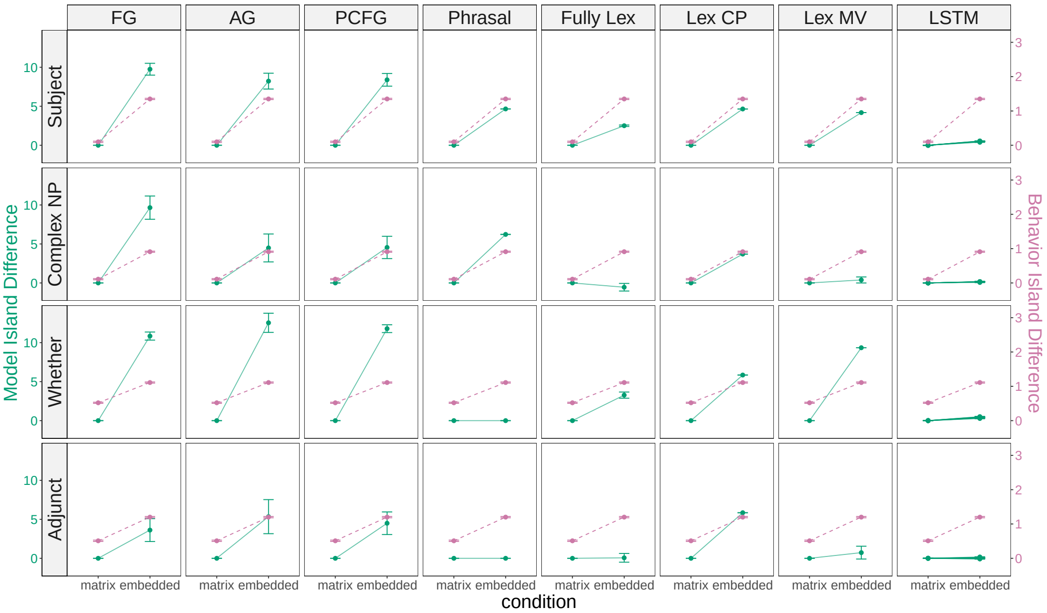


Figure 6. Results from all modeled children (variable-size chunks: FG, AG, PCFG; trigram chunks: Phrasal, Fully Lex, Lex CP, Lex MV) and the LSTM model for the four island types tested in Sprouse et al. (2012): Subject, Complex NP, Whether, and Adjunct. The dashed line represents the original human z-scored acceptability judgments from Sprouse et al. (2012), showing a positive increase in the island difference score from the main to the embedded condition. The modeled learner must generate this positive slope to qualitatively match the human behavior pattern.

6.1.1. Subject and Complex NP islands

To replicate the desired positive slope, the modeled learner must assign a higher score to the embedded non-island-crossing stimuli like (11a)-(12a) than to the embedded island-crossing stimuli like (11b)-(12b).

(11) Subject island embedded stimuli

a. EMBEDDED | NON-ISLAND
What does [$_{\text{IP}}$ Jack [$_{\text{VP}}$ think [$_{\text{CP}}$[$_{\text{IP}}$ __$_{what}$ is expensive]]]]?

b. EMBEDDED | ISLAND
*Who does [$_{\text{IP}}$ Jack [$_{\text{VP}}$ think [$_{\text{CP}}$[$_{\text{IP}}$[$_{\text{NP}}$ the necklace [$_{\text{PP}}$ for __$_{who}$]] is expensive]]]]?

(12) Complex NP island embedded stimuli

a. EMBEDDED | NON-ISLAND
What did [$_{\text{IP}}$ the chef [$_{\text{VP}}$ hear [$_{\text{CP}}$ that [$_{\text{IP}}$ Jeff [$_{\text{IP}}$ baked __$_{what}$]]]]]?

b. EMBEDDED | ISLAND
*What did [$_{\text{IP}}$ the chef [$_{\text{VP}}$ hear [$_{\text{NP}}$ the statement [$_{\text{CP}}$ that [$_{\text{IP}}$ Jeff [$_{\text{VP}}$ baked __$_{what}$]]]]]]?

For the Subject island condition, we see in the first row of Figure 6 that all the modeled learners successfully reproduce the human judgment patterns. One plausible explanation

https://doi.org/10.3390/1010000

for this overwhelming success is that the island-crossing *wh*-dependency (11b) has additional nodes (NP and PP) that are infrequent in the *wh*-dependencies that the modeled learners see. (In fact, most of the tested lexical items in the NPs were unseen, as well.) All the modeled learners were capable of tracking chunks that involved – or notably, didn't involve – NPs or PPs all that often. So, NP and PP nodes in the syntactic path reduce the probability of the syntactic path, and all the modeled learners were capable of encoding these (in)frequencies.

For the Complex NP island condition, we see in the second row of Figure 6 that most modeled learners succeed (the two exceptions are the Fully Lexicalized (**Fully Lex**) and Lexicalized Main Verb (**Lex MV)** trigram learners). Similar to the Subject island condition, an infrequent NP node seems to be a relevant difference between (12a) and (12b). Why then do we see two modeled learners fail, if they're capable of encoding the NP node infrequency? One answer is that these modeled learners' sensitivity to lexical items undermines their ability to view the NP-containing syntactic path as less probable. In particular, the Fully Lex trigram modeled learner includes all lexical information in its representation – that is, its trigrams include the individual lexical items, such as $\text{IP}_{present}$-VP_{hear}-CP_{that}. So, the learner's trigrams are often very infrequent, no matter which trigram it is. For example, $\text{IP}_{present}$-VP_{hear}-CP_{that}, which is used in the non-island (more acceptable) syntactic path of (12a), never actually appeared in the learner's input. More specifically, the trigrams for both the non-island and island Complex NP stimuli are all infrequent. This situation contrasts with the Subject island stimuli, which involved some more-frequent trigrams (e.g., START-$\text{IP}_{present}$-VP_{think} and $\text{IP}_{present}$-VP_{think}-CP_{null}). There, the NP-involving and PP-involving trigrams in the island-crossing syntactic path were in fact less probable than these other more-frequent trigrams, and so the modeled learner could correctly view the island-crossing path as less probable.

A related explanation can account for the failure of the Lex MV trigram learner on the Complex NP island judgments. In particular, because this modeled learner includes main verb lexical items in its trigrams, it's sensitive to the frequency of those verbs. The Complex NP stimuli involve some more-frequent trigrams with the main verbs "hear" and "make" that appeared in the learner's input. More specifically, the non-island path of (12a) involves only one of these more-frequent main verbs while the island-crossing path of (12b) involves two. This means that the Lex MV trigram learner actually gives the island-crossing *wh*-dependency higher probability – that is, it finds the *wh*-dependency in (12b) more acceptable than the non-island-crossing on in (12a). So, even though the modeled learner views the NP-node trigrams in (12b) as low probability, they're not low enough probability to counteract the effect of the higher-frequency trigrams involving the main verbs "hear" and "make".

Overall, most of the modeled learners succeeded at replicating adult acceptability judgment patterns for both Subject and Complex NP islands. Two of the trigram modeled learners that encoded lexical item information in their internalized chunks (Fully Lex, Lex MV) succeeded on the Subject island pattern but failed on the Complex NP island pattern, specifically due to their sensitivity to lexical items.

6.1.2. Whether and Adjunct islands

To replicate the desired positive slope, the modeled learner must assign a higher score to the embedded non-island-crossing stimuli like (13a)-(14a) than to the embedded island-crossing stimuli like (13b)-(14b).

(13) Whether island embedded stimuli

a. EMBEDDED | NON-ISLAND
What does [$_{\text{IP}}$ the detective [$_{\text{VP}}$ think [$_{\text{CP}}$ that [$_{\text{IP}}$ Paul [$_{\text{VP}}$ took __$_{what}$]]]]]?

b. EMBEDDED | ISLAND
*What does [$_{\text{IP}}$ the detective [$_{\text{VP}}$ wonder [$_{\text{CP}}$ whether [$_{\text{IP}}$ Paul [$_{\text{IP}}$ took __$_{what}$]]]]]?

(14) Adjunct island embedded stimuli

a. EMBEDDED | NON-ISLAND
What do [$_{\text{IP}}$ you [$_{\text{VP}}$ suspect [$_{\text{CP}}$ that [$_{\text{IP}}$ the boss [$_{\text{VP}}$ left __$_{what}$ in the car]]]]]?

b. EMBEDDED | ISLAND
*What do [$_{\text{IP}}$ you [$_{\text{VP}}$ worry [$_{\text{CP}}$ if [$_{\text{IP}}$ the boss [$_{\text{VP}}$ leaves __$_{what}$ in the car]]]]]?

For the Whether island condition, we see in the third row of Figure 6 that most modeled learners successfully reproduce the human judgment pattern (the exception is the Phrasal trigram learner). One plausible explanation for these modeling results again involves the modeled learners' sensitivity (or insensitivity) to the lexical items involved in the syntactic paths. In particular, the island-crossing syntactic path of (13b) includes a lower probability complementizer "whether" and a lower probability main verb "wonder"; in contrast, the non-island-crossing syntactic path of (13a) includes a higher-probability complementizer "that" and a higher-probability main verb "think". All modeled learners capable of tracking these lexical items in their chunks – that is, all the learners except the Phrasal trigram learner – were capable of encoding these relative (in)frequencies.

For the Adjunct island condition, we see in the fourth row of Figure 6 that several modeled learners succeeded, but four failed: the Phrasal, Fully Lex, and Lex MV trigram learners, as well as the LSTM learner. Lexical item frequency again can help explain these results, as salient differences between the non-island-crossing stimuli like (14a) and island-crossing stimuli like (14b) involve the complementizer and the main verbs. First, the Phrasal trigram learner is incapable of encoding these lexical items in its chunks, and so fails to distinguish them, just as in the Whether islands.

Interestingly, the Fully Lex and Lex MV trigram learners failed *because* of their sensitivity to the main verbs. More specifically, the main verbs in the non-island-crossing syntactic paths are not much more frequent than the main verbs in the island-crossing syntactic paths. So, these learners wouldn't assign a higher probability to the non-island-crossing syntactic paths just because of the main verbs. Importantly, the Lex MV learner only encodes the main verb lexical items and so fails to prefer the non-island-crossing syntactic path. The Fully Lex trigram learner is capable of encoding the relative frequencies of the complementizer, but fails to disprefer the island-crossing syntactic path because of how infrequent all its trigrams tend to be. More specifically, while trigrams involving complementizer "if" are less frequent than trigrams involving complementizer "that" (which distinguish island-crossing (14b) from non-island-crossing (14a)), they're not less frequent enough counteract the low probability of the other trigrams involved in non-island-crossing (14a).

The LSTM learner has a different issue: it generally assigns scores that are very similar for non-island-crossing and island-crossing *wh*-dependency paths. While the LSTM learner generally captures the correct qualitative pattern (i.e., a positive island difference for three island types), it actually does fail to do so for the Adjunct island type (see Appendix C.1 for more details of its performance). Although LSTM internal representations are difficult to decode, we posit that the LSTM learner fails for a similar reason that the trigram learners above do: improperly dealing with the complementizer lexical information. The LSTM runs that failed assign very similar scores to island-crossing and non-island-crossing sequences, despite the island structure having an unseen lexical item "if".

Overall, many modeled learners succeed at replicating adult acceptability judgment patterns for both Whether and Adjunct islands. Only modeled learners capable of encoding lexical item information were able to succeed, as lexical items distinguish the island-crossing from the non-island-crossing stimuli. However, sensitivity to the wrong lexical items can cause failure as well, just as with the Subject and Complex NP islands.

6.1.3. Summary for superadditivity

Four of the modeled learners succeeded at generating the observed behavior patterns associated with syntactic island knowledge: the FG-based modeled child, the other modeled children relying on variable-sized chunks (AG-based, PCFG-based), and one modeled child relying on trigram chunks (Lex CP). One key reason they succeeded was because they encoded relevant frequency distinctions from the input, involving specific structural elements (e.g., the rarity of NP and PP nodes in *wh*-dependencies) or specific lexical items (e.g., main verbs and complementizers).

6.2. Target behavior: Positive correlation

To replicate human behavior, the modeled learner's predictions must generate a positive correlation between verb-frame frequency and acceptability (the pattern from Liu et al. (2022)). Figure 7 shows that many modeled learners can indeed replicate this pattern. For example, the Lex MV baseline trigram model only includes the main verb lexical item in the dependency path representation and is thus well suited for this task by directly tracking the frequency of the main verb (R^2=0.434).

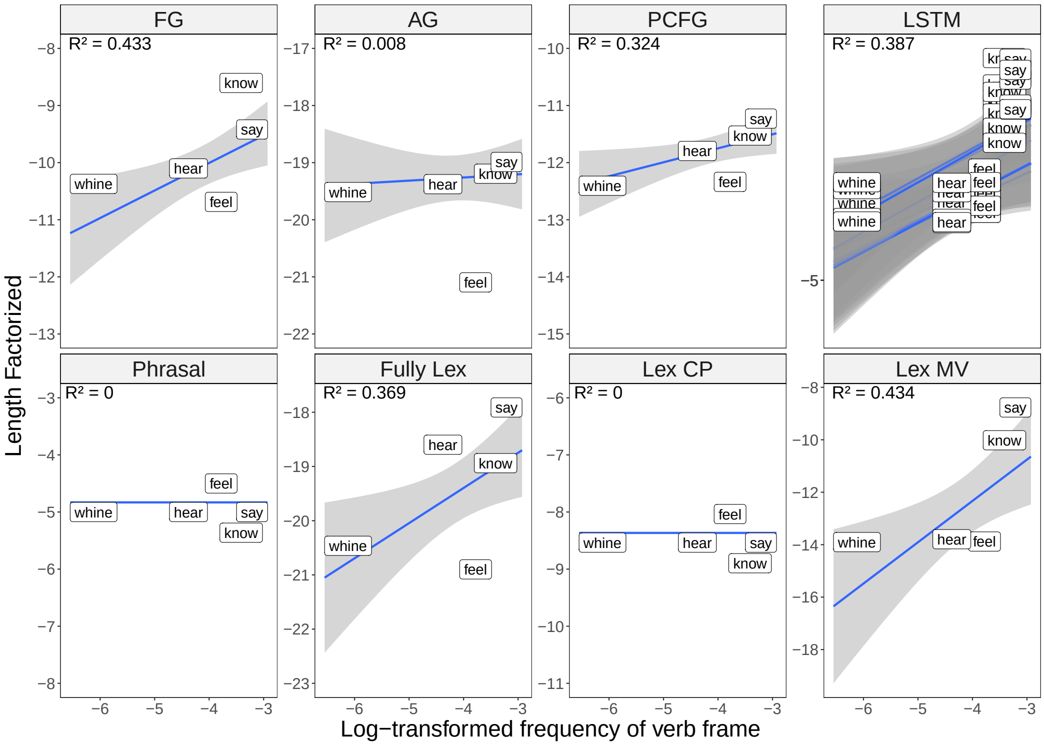
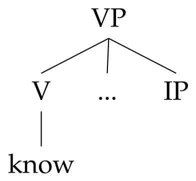


Figure 7. Results from all modeled children (variable-size chunks: FG, AG, PCFG; trigram chunks: Phrasal, Fully Lex, Lex CP, Lex MV) and the LSTM model for the stimuli from Liu et al. (2022), where humans showed a positive correlation between the frequency of the verb-frame in the utterance and the utterance's acceptability. The modeled learner must generate this positive correlation to qualitatively match the human behavior pattern.

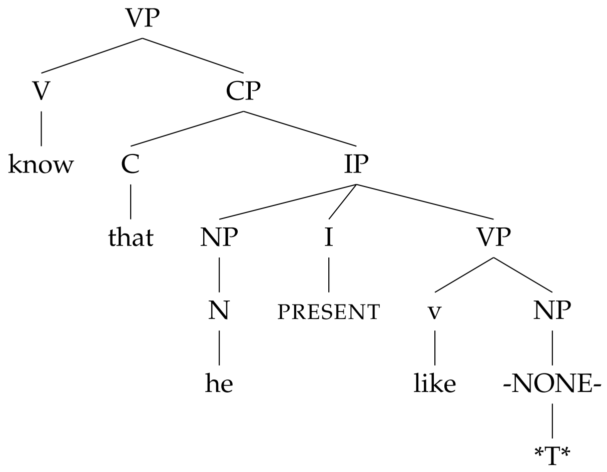
Notably, the FG-based modeled child has equivalent performance (R^2=0.433), even though it distributes probability differences for lexical items across many different chunks. This performance contrasts with the AG-based modeled child, who failed to show a positive correlation (R^2=0.008). Importantly, the FG learner's chunks are more general than the AG learner's chunks, and thus can be used more often when parsing new *wh*-dependencies. For example, the FG learner learns a "know" chunk like (15) that can be abbreviated with the rule VP → *know* ... IP. This chunk can be used for any *wh*-dependency with "know" followed by an IP (e.g., “What does she know that he likes?”, “What does she know he liked?”, “What does he know that she hates?”, “What does he know she hated?”, etc.)

(15) FG-based chunk



In contrast, the AG-based modeled child can’t learn partially-unexpanded chunks like (15). Instead, it can only learn fully-expanded chunks like (16) that can be represented with the rule “VP → know that he PRESENT like". This chunk can be used only for a narrower range of *wh*-dependencies that have “know” followed by complementizer “that”, embedded subject “he”, and an IP with present tense “like” (e.g., “What does she know that he likes?”, but not any of the others mentioned above that the FG learner could handle with its chunk).

(16) AG-based chunk



In other words, the FG learner – but not the AG learner – is able to form helpful reusable chunks (here: involving main verb “know” followed by an embedded clause.) So, the FG learner is sensitive to relevant probability differences that target certain lexical items within a structural context, such as the main verb when followed by a tensed embedded clause.

More specifically, the AG learner creates many chunks involving the same verb lexical items and so distributes the probability differences between individual verbs across many different chunks. This distribution of probability across many chunks can cause relative frequency differences between individual verbs to be hidden. In contrast, the PCFG learner is more like the FG learner in creating useful chunks for this scenario (R^2=0.324). More specifically, the PCFG learner concentrates the probability of a verb into a single rule (e.g., V → *know*), and so is able to capture a (main) verb distinction. Notably, the PCFG learner doesn’t distinguish if the verb is in the main clause or embedded clause; instead, the PCFG learner just so happens to have created verb-based chunks that serve to distinguish main verb frequency.

More generally, the learners that fail to replicate this pattern are of two types: (i) learners incapable of tracking verb lexical items (the Lex CP and Phrasal trigram learners), and (ii) a learner that learns unhelpful chunks (the AG learner). For the modeled learners incapable of tracking verb lexical items, it’s unsurprising that those learners fail to capture a relationship involving verb lexical items – by definition, their representations (in the form of the chunks they can learn) don't include the relevant items. In contrast, the AG modeled learner is capable of including verb lexical items in its chunks, but seems to have included them in unhelpful ways.

When we look at the LSTM model, we see that it also succeeds (across 10 runs, R^2=0.238-0.387; median=0.341). This performance suggests that a model with many free parameters – but which doesn't transparently use chunks – is able to generate the target acquisition output, given the acquisition input.

6.3. Target behavior: Child preferences

Figure 8 shows the correlation between the modeled learner's predictions and the child preference for embedded-clause *wh*-dependencies for all stimuli from De Villiers et al. (2008), reviewed in Table 4. More specifically, we used a linear regression predicting behavioral scores from model output and report R^2 values as a measure of explained variance. Given the small sample size of 9 test items, we focus primarily on comparing R^2 values across models.

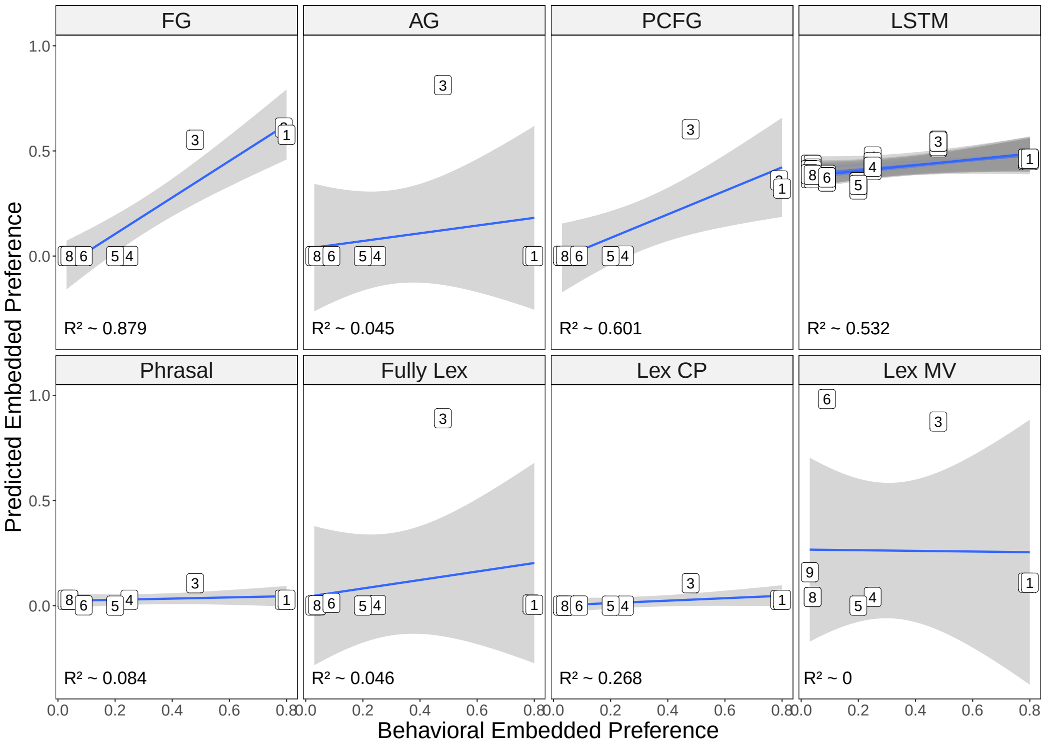


Figure 8. Results from all modeled children (variable-size chunks: FG, AG, PCFG; trigram chunks: Phrasal, Fully Lex, Lex CP, Lex MV) and the LSTM model) for the stimuli from De Villiers et al. (2008). We show the correlation between the modeled learner's predictions and the child preference for embedded-clause *wh*-dependencies for all stimuli from Table 4.

Two of modeled children using variable-size chunks, the FG learner and the PCFG learner, showed the strongest alignment with child behavior (FG: R^2=0.879, $p < 0.01$; PCFG: R^2=0.601, $p < 0.05$). No other modeled child relying on chunks fared as well (R^2= 0 - 0.268). In contrast, the LSTM model also aligned fairly well with child behavior for 3 of its 10 runs (R^2=0.228-0.532; median=0.410) – see Figure A3 in Appendix C.2 for details. These high-level results suggest certain types of chunks can be useful for acquisition – and in particular, the FG-based efficient chunks. However, a complex model with many free parameters (the LSTM model) can also succeed sometimes.

https://doi.org/10.3390/1010000

Looking more closely, we can also see a few notable results within this broader pattern of results. First, the FG learner has particular success at matching children's observed embedded-clause preference for items 1 and 2, repeated below as (17)-(18).

(17) Item 1: child embedded-clause preference = 0.80

a. How did [$_{\text{IP}}$ the boy [$_{\text{VP}}$ say $\_\_\textit{how}_{main}$ [$_{\text{CP}}$ he hurt himself]]]?
b. How did [$_{\text{IP}}$ the boy [$_{\text{VP}}$ say [$_{\text{CP}}$ [$_{\text{IP}}$ he [$_{\text{VP}}$ hurt himself $\_\_\textit{how}_{emb}$]]]]]?

(18) Item 2: child embedded-clause preference = 0.79

a. What did [$_{\text{IP}}$ the mother [$_{\text{VP}}$ say $\_\_\textit{what}_{main}$ [$_{\text{CP}}$ she bought]]]?
b. What did [$_{\text{IP}}$ the mother [$_{\text{VP}}$ say [$_{\text{CP}}$ [$_{\text{IP}}$ she [$_{\text{VP}}$ bought $\_\_\textit{what}_{emb}$]]]]]?

One key factor for these stimuli is the main verb "say", which is part of a high-frequency chunk in the FG learner's inventory: "IP → PAST (VP → (V → *say*)) CP" (abbreviated as "IP → PAST *say* CP"). We posit that, by deploying this chunk for the two stimuli in (17b),(18b) the FG learner was able to keep the embedded-clause *wh*-dependency score higher than the main-clause *wh*-dependency score.

To evaluate this possibility, we altered the stimuli to use the present tense of "say" instead of the past tense in the main clause, as in (19)-(20). The FG-based chunk using "say" in the past tense no longer can be used to generate these syntactic paths. With this manipulation, we find that the FG-based learner no longer prefers the embedded-clause *wh*-dependencies in (19b)-(20b). Thus, it seems likely that the FG learner relied on the high-frequency "say"-chunk in order to match child preference behavior. Importantly, this chunk isn't possible for the other chunk-based learners to include in their inventories, highlighting the utility of this kind of flexible chunk.

(19) Item 1 using present "does"

a. How does [$_{\text{IP}}$ the boy [$_{\text{VP}}$ say $\_\_\textit{how}_{main}$ [$_{\text{CP}}$ he hurt himself]]]?
b. How does [$_{\text{IP}}$ the boy [$_{\text{VP}}$ say [$_{\text{CP}}$ [$_{\text{IP}}$ he [$_{\text{VP}}$ hurt himself $\_\_\textit{how}_{emb}$]]]]]?

(20) Item 2 using present "does"

a. What does [$_{\text{IP}}$ the mother [$_{\text{VP}}$ say $\_\_\textit{what}_{main}$ [$_{\text{CP}}$ she bought]]]?
b. What does [$_{\text{IP}}$ the mother [$_{\text{VP}}$ say [$_{\text{CP}}$ [$_{\text{IP}}$ she [$_{\text{VP}}$ bought $\_\_\textit{what}_{emb}$]]]]]?

Another notable learner result we see in Figure 8 is a prediction from all modeled learners for a stronger embedded-clause preference for item 3, in contrast to the more-neutral child preference of 0.48. When we look more closely at item 3 from De Villiers et al. (2008), repeated as (21) below, we can see that the embedded clause doesn't involve a CP phrase (instead, only a non-finite IP appears). For most modeled learners, this non-finite IP structure is still fairly frequent in their experience, and so doesn't lower the embedded-clause *wh*-dependency score much at all.

(21) Item 3: child embedded-clause preference = 0.48

a. Who did [$_{\text{IP}}$ the police woman [$_{\text{VP}}$ help $\_\_\textit{who}_{main}$ [$_{\text{IP}}$ to call]]]?
b. Who did [$_{\text{IP}}$ the police woman [$_{\text{VP}}$ help [$_{\text{IP}}$ to [$_{\text{VP}}$ call $\_\_\textit{who}_{emb}$]]]]?

A third common behavior from the chunk-based modeled learners is an overall preference for main-clause *wh*-dependencies, with most embedded-clause stimuli receiving a score <0.5. This dispreference for embedded-clause *wh*-dependencies is likely due to embedded-clause syntactic paths involving lower-frequency chunks (e.g., chunks using CP) that therefore reduce the average probability of the syntactic path, compared to main-clause dependencies. The LSTM model behaves somewhat differently, in that it tends to be more neutral for all its preferences (i.e., preference for all items around 0.5). However, for the runs where it does generate preferences that better align with child behavior, it also displays a general main-clause preference (i.e., all items but item 3 have a predicted preference <0.5). See Appendix C.2 for more detailed discussion.

7. Discussion

Here we have explored the potential for a child looking for efficient syntactic chunks to acquire knowledge about syntactic islands. In particular, a modeled child relying on Fragment Grammars (**FGs**) to define its hypothesis space of possible chunks is able to generate predicted behavior patterns that align with three sets of human behavior patterns signaling knowledge of syntactic islands. No other chunk-based modeled learner succeeded as well, with most failing on at least one of the target behavior patterns. Our results also suggest that simply having more free parameters to encode the input (which the FG-based learner does, compared to the other chunk-based learners) isn't sufficient for acquisition success, as a comparison model relying on an LSTM with tens of thousands of free parameters wasn't able to match all target patterns. We interpret our results as support for a learning theory for syntactic islands where the child's goal is to find efficient syntactic chunks on the basis of the input.

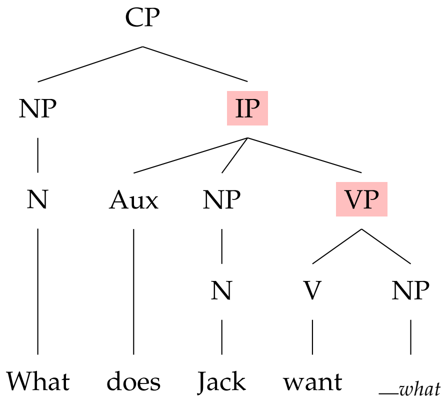
We now discuss assumptions of the current FG-based implementation of the efficient-chunks learning theory, and how they might be investigated in future work. In particular, we consider the modeled child's intake, chunking preferences, and children's cognitive limitations. We also discuss alternative acquisition targets that include a wider range of empirical data, incorporate the incremental nature of acquisition, and consider the impact of naturally-occurring linguistic variation in children's input. We conclude with how this approach to acquiring syntactic islands relates to the current theoretical landscape.

7.1. Assumptions of the current implementation

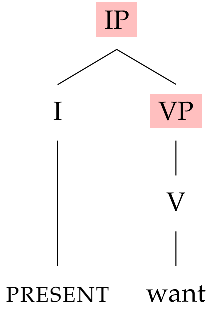
7.1.1. Intake

The current implementation of the learner looking for efficient syntactic chunks relied on a particular perception of the input – namely, phrase structure for the utterance, with the syntactic path between the *wh*-word and its gap highlighted, as in (22). More specifically, the intake the modeled child learned from included only the syntactic path information (as in (22b)). This intake reflects a learning assumption that children know to ignore other information available when learning about *wh*-dependencies.

(22) a. Phrase structure and highlighted syntactic path for "What does Jack want?"



b. Syntactic path only



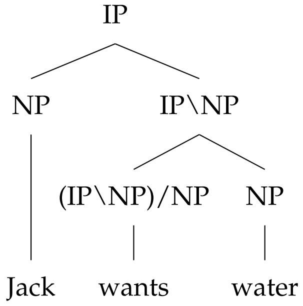
However, do children in fact need to ignore the other information available? That is, could they still succeed at learning the appropriate *wh*-dependency patterns we looked at here even without this intake filtering? Or instead, is this intake filtering necessary for acquisition success? There are several cases in the acquisition literature that suggest intake filtering is a key component for acquisition success for other phenomena (e.g., basic word order: L. Pearl and Weinberg 2007; English metrical stress: L. Pearl et al. 2017; the English passive: Nguyen and Pearl 2019).

One way to investigate the impact of intake filtering is to remove the intake filtering implemented here, and allow children to learn from the entire syntactic structure information available for the utterance. We discuss two possibilities that implement this option. First, children could simply learn from the entire syntactic structure available, with no special status given to the syntactic path (i.e., (22) without the highlighting). Dickson (2025) uses an FG-based modeled learner to explore this option for acquiring other types of syntactic knowledge, including some *wh*-dependency knowledge, with moderate success. However, a more thorough investigation using the target *wh*-dependency patterns we used here remains to be done. One notable issue Dickson (2025) found was data sparsity – because much more information is available in each data point, the FG-based modeled child likely requires more language experience to successfully sift through the possible syntactic chunks.

A second way to relax the intake filtering assumption is to include all the syntactic structure for the utterance, but keep the special status of the syntactic path, as in (22). That is, the modeled child has access to the entire structure of the utterance, but knows there's something important about the syntactic path. One way to implement this idea is by using a grammar formalism that indicates the syntactic path, such as "slash passing" in CCG (Steedman & Baldridge, 2011), HPSG (Borsley & Crysmann, 2021), and GPSG (Gazdar, 1985)). In these formalisms, syntactic categories can be functions that specify how the other units are combined with the current linguistic unit.

For example, in the CCG formalism from Steedman and Baldridge (2011), the category for the verb "wants" would be a function like (IP\NP)/NP as in (23), specifying how to derive a sentence-level IP.

(23)



In particular, the derivation is something like the following:

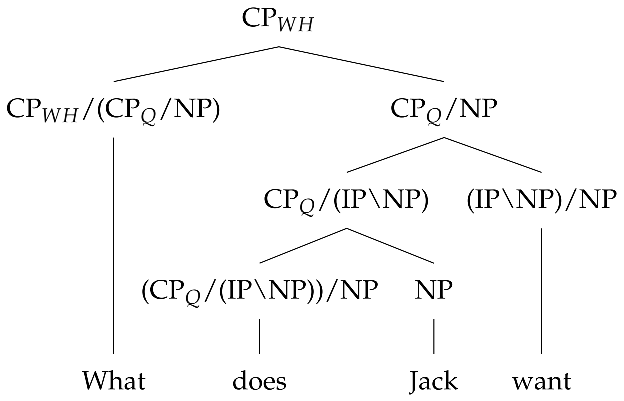
(i) Combine the unit "wants" with an NP on the right (e.g., "water") in order to generate the IP\NP unit

(ii) Combine this unit with an NP on the left (e.g., "Jack") to generate the IP

The slashes encode direction of the expected unit to be combined, with "/" specifying a combination on the right and "\" specifying a combination on the left.

To analyze (22) under the CCG formalism, many of the phrase labels would be replaced by these function labels, as in (24). For instance, the label for the auxiliary "does" is now (CP$_Q$/(IP\NP))/NP, which signals that this unit is first looking for an NP on the right (here, the inverted subject "Jack"). A new unit is formed once this combination occurs that is again looking for another unit type (IP\NP), which itself is looking for yet another unit type (NP), and so on until the *wh*-question is formed (Hockenmaier & Steedman, 2007).

(24)



Notably in (24) the "/NP" that percolates up the right edge marks the dependency path of the kind used as the modeled child's intake in our implementation (i.e., CP$_Q$/NP-(IP \NP)/NP is equivalent to IP-VP, with IP≈CP$_Q$/NP and VP≈(IP \NP)/NP). So, allowing the modeled child's intake to include full trees that also naturally highlight the syntactic path would be another way of relaxing the current intake restriction to learn only from the syntactic path of the *wh*-dependency. We do anticipate there may be a similar data sparsity issue to what Dickson (2025) found, as the syntactic category units available to form efficient chunks would be far larger than what we used in the current FG-based implementation. For instance, instead of only VP, a modeled child would potentially need to consider VP-based chunks such as (IP \NP)/NP (a transitive verb), (IP \NP) (an intransitive verb), ((IP \NP)/NP)/NP (a ditransitive verb with an indirect object), and ((IP \NP)/NP)/PP (a ditransitive verb with a prepositional object), among others.

7.1.2. Chunking preferences

The current FG-based implementation of the efficient-chunks learner has some flexibility about the preferred size of chunks. Recall from Section 3 that a key distinction between the FG-based modeled child and other modeled children allowing variable-sized chunks is how often they prefer to create larger chunks: an AG-based modeled child always prefers to expand potential chunks ($p_{expand} = 1.00$) while a PCFG-based modeled child never does ($p_{expand} = 0.00$)). The FG-based modeled child can therefore have a preference anywhere in between these extreme points (i.e., $0.00 < p_{expand} < 1.00$). Here, we used hyperparameter settings that led to $p_{expand} = 0.50$, based on prior successful FG implementations aimed at learning other linguistic phenomena (T. O'Donnell et al., 2011).

While this FG-based modeled child performed well, the PCFG-based modeled child performed almost as well at generating the target behavior patterns (recall Table 5). This result suggests that having a preference for smaller chunks (i.e., a p_{expand} closer to 0.0) may also be able to generate target behavior patterns as well as the FG-based implementation here. Future work can investigate more generally if FG-based modeled children with different chunk-size preferences (i.e., different p_{expand} values) are able to generate the target behaviors as well as (or perhaps better than) the FG-based implementation using a $p_{expand} = 0.50$. These future findings would allow us to better understand the necessary chunking preferences a child would need to have in order to succeed at the acquisition task investigated here.

7.1.3. Cognitive limitations

The FG-based modeled child here incorporated one major limitation in child language acquisition: the (limited) amount of data children encounter before they achieve acquisition success. However, another major limitation for children relates to their cognitive resources, which impacts how children extract information from the data they encounter and update their internal hypotheses (among other things). For instance, limited memory resources might cause children to either miss some of the available information in the moment (Forsythe & Pearl, 2020; Gagliardi et al., 2017; L. Pearl & Forsythe, 2025) or misperceive information that's there (Gulrajani & Lidz, 2024). The FG-based modeled child implemented here was idealized in this respect – it could perfectly extract the desired intake (i.e., the syntactic path), with no loss or skewing of information. Moreover, the modeled child implemented here received all the data at once that children would encounter during their learning period, rather than only encountering it incrementally as children do.

As another example, limited cognitive resources might cause children to imperfectly search their hypothesis space of possible chunk inventories. The FG-based modeled child implemented here was also idealized in this respect – it used computational-level inference techniques to identify a high-probability chunk inventory. In contrast, children would likely be approximating this inference as best they can with their limited cognitive resources, and may not in fact succeed as easily at identifying a high-probability chunk inventory.

Future work could implement modeled children that incorporate child-like limitations like those outlined above: forgetting or skewing information in the data, encountering data incrementally, and approximating inference (e.g., Sanborn et al. 2010). If modeled children looking for efficient chunks continue to succeed under these conditions, then we have additional support for the robustness of this learning theory. In contrast, if the efficient-chunks modeled children don't perform as well, we can better understand the necessary conditions that this acquisition theory depends on. Initial investigations of this type by Dickson and colleagues (Dickson, 2025; Dickson et al., 2024) have found that the FG-based modeled child implemented here can still succeed even in the face of fairly severe memory limitations that cause the modeled child to miss available information.

7.2. The target of acquisition

Here, we set the target of acquisition to be a set of behavioral patterns signaling knowledge about English *wh*-dependencies, specifically adult judgment patterns and child interpretation preferences. However, the ideal target of acquisition could (and, in our opinion, should) be broader. We discuss several concrete options for usefully expanding the target state.

7.2.1. Additional empirical data about *wh*-dependencies

***Wh*-dependencies with multiple gaps.**

A related set of empirical patterns we might wish to account for involves multiple-gap *wh*-dependencies, such as parasitic gaps (25), purpose clauses (26), and across-the-board extraction (27) (Engdahl, 1983; Grosu, 1973; Ross, 1967).

(25) Parasitic gaps

a. Acceptable: [Which book] did you judge $\_\_{main}$ before reading $\_\_{parasitic}$?
b. Acceptable: [Which book] did you judge $\_\_{main}$ before reading the review?
c. Unacceptable: [Which book] did you judge the cover before reading $\_\_{parasitic}$?

(26) Purpose clauses

a. Acceptable: [Which book] did you buy $\_\_{main}$ in order to give $\_\_{purpose}$ to Lindy?
b. Acceptable: [Which book] did you buy $\_\_{main}$ in order to give it to Lindy?
c. Unacceptable: [Which book] did you buy the movie in order to give $\_\_{purpose}$ to Lindy?

(27) Across-the-Board extraction

a. Acceptable: [Which book] did you read $\_\_{first}$ and review $\_\_{second}$?
b. Unacceptable: [Which book] did you read $\_\_{first}$ and review the movie?
c. Unacceptable: [Which book] did you read the summary and review $\_\_{second}$?

Notably, each of these constructions requires both gaps in order for either the second gap (25)-(26) or either gap (27) to be acceptable. No current acquisition theory accounts for how adults come to know these patterns. The efficient-chunks acquisition theory investigated here may offer an answer, but will need to be evaluated concretely in future work.

Other *wh*-dependency preferences.

Another type of behavior we might wish to account for is knowledge of *wh*-dependency preferences when there are multiple viable (grammatical) options, rather than simply recognizing when a *wh*-dependency is massively dispreferred (i.e., crossing a syntactic island). Omaki et al. (2014) describes child and adult interpretation preferences for the *wh*-dependencies like those in (28), where there are two possible gaps for *where*: one in the main clause and one in the embedded clause.

(28) Where did Lizzie {say | tell someone | say to someone} $\_\_{where_{main}}$
[that she was gonna catch butterflies $\_\_{where_{emb}}$]?

Notably, Omaki et al. (2014) manipulated the main verb phrase (*say/tell someone/say to someone*) and found that both child and adult preferences vary based on the lexical items. In particular, most children and adults prefer resolving the dependency in the embedded clause (answering where Lizzie will catch the butterflies) when the main verb is "say." However, when the main verb phrase is "tell someone" or "say to someone", the preference switches to main-clause resolution (answering where Lizzie told someone or said to someone).

Part of this preference is in fact captured by the current FG-based modeled child: when the main verb is "say", this modeled child prefers an embedded-clause interpretation, as opposed to its general main-clause preference. So, this modeled child could capture the difference between "say" and "tell (someone)". However, this modeled child can't capture the difference between "say" and "say to someone", as the "to someone" part isn't part of this modeled child's intake (i.e., "to someone" isn't part of the syntactic path). One concrete path for future work is to implement some of the suggestions from section 7.1.1 that allow more information into the modeled child's intake, while still preserving the overall approach of identifying efficient chunks. More generally, it seems reasonable that the target of acquisition for *wh*-dependency knowledge should include preferences like the ones described here, in addition to knowledge of syntactic islands.

7.2.2. Immature *wh*-dependency knowledge

One strength of previous chunking implementations of child language learning is their ability to capture incremental development in children's production performance (Freudenthal et al., 2015; McCauley & Christiansen, 2019). That is, adult-like knowledge is the eventual acquisition target, but there are stages along the way that are reasonable to consider as target states. In the spirit of these prior approaches that weren't specifically targeting *wh*-dependency knowledge, we might consider whether *wh*-dependency production data can be used to specify an immature target state that a modeled child could aim to produce. If so, then we would have a much richer target state to evaluate future acquisition theories against.

7.2.3. Linguistic variation

Another important expansion involves considering linguistic variation, both across languages and within languages. Ideally, a learning theory for syntactic islands (and *wh*-dependencies more generally) would work universally. For the modeled child trying to identify efficient syntactic chunks, this means that acquisition success should occur no matter what the specific syntactic islands are for a given language or dialect. Prior work examining other chunking approaches to syntactic islands found some success (e.g., English from lower socioeconomic status households: L. Pearl and Bates 2022a) but not complete success (e.g., Norwegian: Kobzeva and Kush 2025). It remains to be seen if the efficient-chunks theory can succeed more completely.

7.3. Theoretical implications

The efficient-chunks acquisition theory implemented here draws from two different traditions within language acquisition. Similar to generativist approaches, this acquisition theory assumes children have prior syntactic knowledge that allows them to impose certain syntactic structure on their input (Chomsky et al., 1973; Pinker, 1999) when transforming it into their intake. This assumption contrasts with many constructionist approaches to syntactic learning, particularly those involving chunking, which assume the child is operating over unstructured word sequences (Freudenthal et al., 2015; McCauley & Christiansen, 2019).

However, similar to constructionist approaches, this acquisition theory assumes that sophisticated syntactic knowledge (here, about syntactic islands) doesn't require specific knowledge *a priori*, but instead emerges during learning (A. E. Goldberg, 2006; McCauley & Christiansen, 2019; Tomasello, 2001). That is, in contrast to generative approaches to syntactic islands (Chomsky et al., 1973), no island-specific structural knowledge is built in.

Notably, the success of this acquisition theory has implications for current proposals about why there appears to be constrained variation over island constraints cross-linguistically. More specifically, one proposal is that this constrained variation is a result of constraints that are in place during acquisition (Chomsky et al., 1973; Pinker, 1999). Without these built-in constraints (i.e., built-in prior knowledge pertaining to syntactic islands), children could not learn the syntactic islands of their language. Therefore, the reason languages have constraints is because these constraints were built into the child mind in order to make acquisition possible. That is, human-internal constraints active during acquisition – in order to make acquisition possible – are why languages are shaped the way they are with respect to syntactic islands.

Our results weaken this argument by demonstrating how acquisition is possible without building in constraints specific to syntactic islands. In particular, while some knowledge of syntax is required *a priori*, children don't need island-specific knowledge to succeed. So, if languages show constrained variation when it comes to syntactic islands, this constrained variation could originate from somewhere else.

8. Conclusions

Here we have implemented a language acquisition theory for *wh*-dependency knowledge (including syntactic islands) where the learner aims to identify an efficient representation of syntactic chunks for *wh*-dependencies; knowledge of syntactic islands emerges from this high-efficiency chunk representation, rather than being represented separately. When implemented concretely in a modeled child who learns from realistic child input, this acquisition theory can explain a variety of language behavior patterns that signal knowledge of syntactic islands. In short, children could acquire sophisticated syntactic knowledge even with less-sophisticated innate linguistic machinery as long as they have the right learning objective in mind.

Author Contributions: "Conceptualization, N.D., L.P., and R.F.; methodology, N.D., L.P., and R.F; software, N.D., R.F; validation, N.D., R.F; formal analysis, N.D., R.F.; investigation, N.D., L.P., and R.F.; resources, N.D., L.P., and R.F.; data curation, N.D., L.P., and R.F.; writing—original draft preparation, N.D., L.P.; writing—review and editing, N.D., L.P., and R.F.; visualization, N.D., R.F.; supervision, L.P., R.F.; project administration, N.D., L.P., and R.F.; funding acquisition, N.D., L.P., and R.F. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Science Foundation DDRI grant number 2438734.

Data Availability Statement: Code for the modeled learner implementations and data visualizations are available at https://github.com/nielswd23/AcquiringIslandConstraints. Additional code for the Fragment Grammar model provided by Tim O'Donnell through personal communication.

Acknowledgments: We are grateful to the comments and feedback from the members of the Quantitative Approaches to Language Science group at UC Irvine, and from the audiences at the 46th and 48th Boston University Conference on Language Development, and the 5th conference of the Society for Computation in Linguistics.

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Appendix A. Implementation of Fragment Grammar

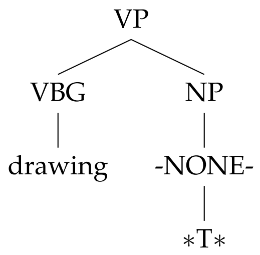
Appendix A.1. PCFG definition

A PCFG includes rules like those in (29), and probabilities associated with the rules.

		Rule	**Probability**
(29)	r_1	VP → VBG NP	0.40
	r_2	VBG → *drawing*	0.30
	r_3	NP → -NONE-	0.25
	r_4	-NONE- → $*$T$*$	1.0

So, if r_2 has probability 0.30 (i.e., $p(r_2)$=0.3), when the grammar generates structure, VBG will expand to "drawing" 30% of the time, and expand to something else (like "running" or "sleeping") the other 70% of the time.. The probability of a structure like (2), repeated below as (30), under this grammar is calculated by multiplying the probabilities of the rules that compose the tree i.e., $\prod_{r_x \in tree} r_x$).

(30)



During initialization of the FG-based modeled child, the probabilities for all rules are drawn from a multinomial distribution with a Dirichlet prior in order to form the base PCFG for modeled child. (T. O'Donnell et al., 2011).

Note that an FG also allows rules that represent larger chunks, such as VP → VBG (NP → (-NONE- → *T*)). The implementation we use compresses this rule by substituting the leaves (here: *T*) for the structure that branches into those leaves (here: (NP → (-NONE- → *T*))), so that the rule above would be represented as VP → VBG *T*.

Appendix A.2. Pitman-Yor Process

Formally, the Pitman-Yor Process (**PYP**) is a non-parametric distribution used to cluster tokens (Harmon et al., 2021; Pitman & Yor, 1997). Here, the modeled child uses the PYP to sample existing rules (chunks) from the base PCFG representation in order to learn new, bigger chunks, given the input. For instance, using a PYP, the modeled child may consider and ultimately learn a new chunk VP → (VBG → *drawing*) (NP → (-NONE- → *T*)) – represented as VP → *drawing* *T* – by clustering together rules r_1, r_2, r_3, and r_4 in (29).

Given frequent observations of recurring structures like (30), the modeled child learns to associate this expansion (i.e., VP → *drawing* *T*) with a single memorized derivation instead of expanding each nonterminal (VBG, NP, -NONE-) independently. Chunks observed more frequently are more likely to be reused in the future, enabling the modeled child to generalize over recurring patterns in the data.

Appendix A.3. Walkthrough of FG Pitman-Yor process

The second key feature in the FG modeled child is a "lazy evaluation scheme" adaptation to the Pitman-Yor process (T. O'Donnell et al., 2011; T. J. O'Donnell et al., 2009): chunks are allowed to leave some non-terminals unexpanded – that is, these non-terminals can be evaluated later. This feature allows the FG modeled child to consider chunks like VP → VBG *T*, with the non-terminal VBG unexpanded. In particular, the modeled child has a probability for continuing non-terminal expansion (p_{expand}) (T. J. O'Donnell et al., 2009), which can be learned from the input.

To illustrate this idea, Figure A1 shows how different treelets ("Computations") would be generated using the adapted Pitman-Yor process and a visual metaphor of customers sitting at tables in different restaurants. Here, the probability of choosing a specific option (restaurant table) depends on how much probability (how many customers) is already associated with that option (how many customers are already at the table).

Let's begin with the leftmost table of the VP → VBG NP restaurant. We follow the solid red lines to expand the non-terminals VBG (to *drawing*) and NP (to *T*). This generates a chunk with all non-terminals expanded (i.e., VP → *drawing pictures*), and is a chunk an AG or FG could consider. This chunk also generates the first treelet in the Computations row (VP → *drawing pictures*).

Moving to the next table in VP → VBG NP restaurant, we can follow the solid red line to expand the non-terminal VBG (to *framing*). We can also leave the NP unexpanded in this chunk, and follow the dotted gray line to create a separate chunk NP → *T*. The first chunk has one non-terminal unexpanded (i.e., VP → *framing* NP), and is a chunk only

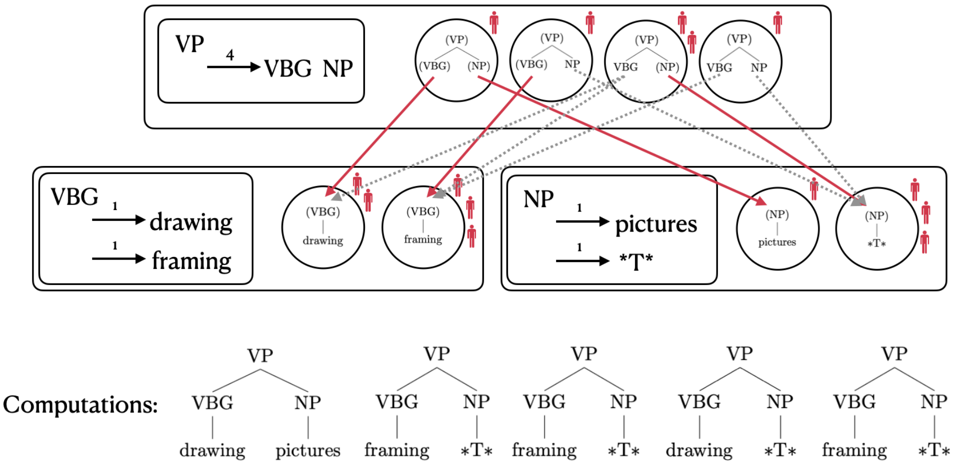


Figure A1. Example FG state adapted from T. J. O'Donnell et al. (2009), using a visual metaphor of customers sitting at table in different restaurants. Solid red lines indicate a path through the restaurants as a chunked structure. Dotted gray lines correspond to the lazy evaluation where the expansion of a given non-terminal is left unevaluated.

the FG could consider. The second chunk (NP → *T*) could be considered by a PCFG, AG, or FG. Notably, this chunk has a higher probability than many others (i.e., it has three "people" as its "table"). Combining these chunks together allows the second treelet in the Computations row (VP → *framing* *T*).

Moving to the third table in VP → VBG NP restaurant, we can leave the VBG unexpanded in this chunk, and follow the dotted gray line to create a separate chunk VBG → *drawing* or VBG → *framing*. We can also follow the solid red line to expand the non-terminal NP to *T*. The first chunk has one non-terminal unexpanded (i.e., VP → VBG*T*), and is a chunk only the FG could consider. The second chunk (VBG → *drawing/framing*) could be considered by a PCFG, AG, or FG. Notably, this chunk also has a higher probability than many others (i.e., it has three "people" as its "table"). Combining these chunks together allows the third and fourth treelets in the Computations row (VP → *framing/drawing* *T*).

Moving to the rightmost table in VP → VBG NP restaurant, we can leave the VBG unexpanded in this chunk, and follow the dotted gray line to create a separate chunk VBG → *framing*. We can also leave the NP unexpanded in this chunk, and follow the dotted gray line to make a separate chunk NP → *T*. These are "minimal chunks" that can be considered by a PCFG, AG, or FG. Combining these chunks together allows the fifth treelet in the Computations row (VP → *framing* *T*).

Notably, the third table option seems to offer the highest probability chunk options for generating these treelets, since its chunks involve more probability ("people"): VP → VBG *T* (2 people) and VBG → *drawing/framing* (2 or 3 people). Only the FG can consider the first chunk of this option, which is larger than a minimal PCFG chunk but still includes unexpanded non-terminals.

Appendix A.4. Implementation of the FG modeled child

We follow T. J. O'Donnell (2015) for the FG Pitman-Yor parameter settings: a=0, b=1. We set the Dirichlet hyperparameter π=1, capturing a weak uniform prior over possible chunks.

The probability that a potential chunk expands a non-terminal to make a larger chunk (p_{expand} in the main text) is sampled from a beta distribution (T. J. O'Donnell et al., 2009), with a mean ("sticky concentration parameter") $\nu = 1$ and a sample size ("sticky distribution parameter") μ = 0.5. These settings correspond to $Beta(0.5{,}0.5)$ in the traditional α and β parameterization of this distribution. With α and β <1, this distribution places much of the probability mass on the extremes of 0 and 1. So, the modeled child will often learn strategies that expand a particular non-terminal with probability close to 1 (almost always) or close to 0 (almost never).

We performed 1,000 sweeps of the Metropolis-Hastings sampling algorithm to identify potential fragment grammars (chunk inventories). We used the highest-probability grammar.

To assign probabilities to particular data points (syntactic paths), we calculate the maximum *a posteriori* score under the grammar (Eisenstein, 2018), corresponding to the the highest-probability parse for the item.

Appendix B. Comparison learners

Appendix B.1. Implementation of the PCFG and AG modeled children

For the PCFG modeled child, p_{expand} should = 0, which is parameterized via a beta distribution with a mean ("sticky concentration parameter") ν = 1 and a sample size ("sticky distribution parameter") μ = 0. This corresponds to $Beta(0{,}1)$ in the traditional α and β parameterization, and places all the probability mass on 0. So, the modeled child will only learn strategies that never expand a non-terminal.

For the AG modeled child, p_{expand} should = 1, which is parameterized via a beta distribution with a mean ("sticky concentration parameter") ν = 1 and a sample size ("sticky distribution parameter") μ = 1. This corresponds to $Beta(1{,}0)$ in the traditional α and β parameterization, and places all the probability mass on 1. So, the modeled child will only learn strategies that always expand a non-terminal.

The remaining implementation is the same as the FG modeled child (i.e., Pitman-Yor parameter settings, Dirichlet settings, Metroplis-Hastings sweeps, grammar selection, and calculating highest-probability parses for items).

Appendix B.2. Implementation of trigram-based modeled children

Each trigram $t \in Trigrams$ is comprised of three units: u_1-u_2-u_3. We calculate the probability of t from the input, by observing the trigram's frequency and using Laplace smoothing to account for unseen trigrams, as in (A1).

$$P_t(u_1\text{-}u_2\text{-}u_3) = \frac{\text{count}(u_1\text{-}u_2\text{-}u_3) + 1}{\text{count}(u_1\text{-}u_2) + |Trigrams|} \tag{A1}$$

We score a data point as follows: For each element in the data point sequence S, we calculate the joint log probability of S by summing over the log probabilities of each trigram that comprises the sequence $t_s \in S$, as in (A2).

$$\log P(S) = \sum_{t_s=1}^{S} \log(P_{t_s}(u_1\text{-}u_2\text{-}u_3)) \tag{A2}$$

Appendix B.3. LSTM implementation

The long-short term memory (**LSTM**) model (Hochreiter & Schmidhuber, 1997) is a type of Recurrent Neural Network with an additional memory gating mechanism, making it better at learning that requires information to propagate across long sequences (Eisenstein, 2018).

The input sequences to the LSTM we used are the length of the maximum sequence (14) plus a start symbol and an end symbol. A padding character was added for shorter sequences. This padding character was masked when calculating loss. We trained a single-layer LSTM on each sample of the training data with the objective of minimizing cross-entropy loss. We selected the size of the hidden state to be the size of the vocabulary: 344. We held out 20% of the data for each sample of the training data and performed a hyperparameter grid search to determine which values resulted in a minimum of the held-out loss. We found that the following parameter setting resulted in the lowest held-out loss across samples: embedding dimension = 300, batch size = 300, number of epochs = 500, learning rate = 0.0001.

We score a data point as follows: For each element in the data point sequence S, we perform a logsoftmax over the LSTM hidden state to get log probabilities for the following unit u_{t+1}. This log probability distribution corresponds to the model's expectation about the next unit. Extracting the correct next unit from this distribution gives us a list of log probabilities corresponding to each observed unit in the sequence. Summing these probabilities gives us the joint log probability of the sequence, as summarized in (A3).

$$\log P(S) = \sum_{t=1}^{T-1} \log P(u_{t+1}|u_{1:t}) \tag{A3}$$

Appendix C. LSTM results

Appendix C.1. Sprouse et al. (2012) island pattern

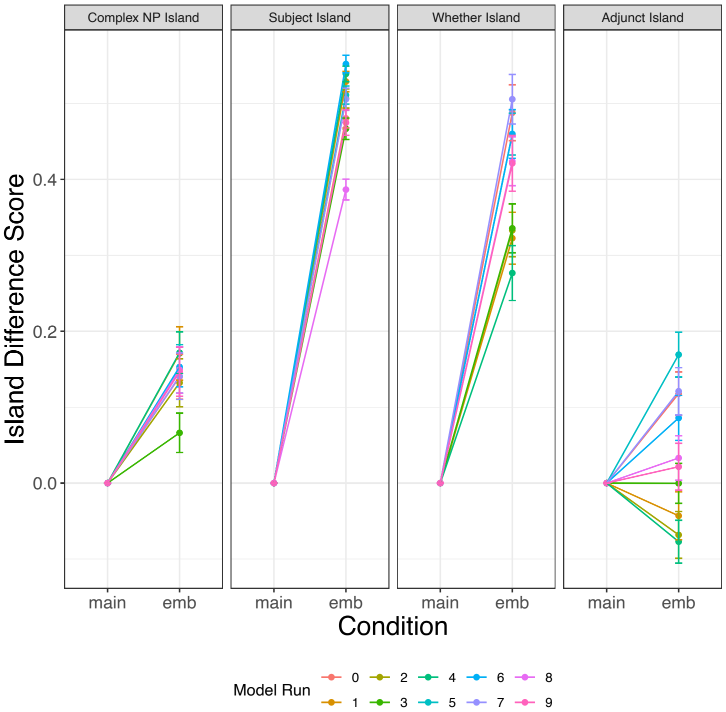


Figure A2. Zoomed-in results of the LSTM model for the behavioral patterns from Sprouse et al. (2012).

As mentioned in the main text, the LSTM model is capturing behavioral patterns for all island types except the Adjunct Island (see Figure A2, where 4 of 10 runs show an island difference with a non-positive slope). The model runs that struggle seem to fail due to improperly dealing with the complementizer lexical information. In particular, the failing runs strongly expect the high-probability “null" lexical item following the “CP", and aren’t sensitive to the low-probability “that" vs. unseen “if" distinction.

Appendix C.2. De Villiers et al. (2008) child preferences

Recall that the chunk-based modeled children showed a general dispreference for the embedded-clause *wh*-dependency. In contrast, the LSTM model predicts preferences close to 50% for all items (see Figure A3).

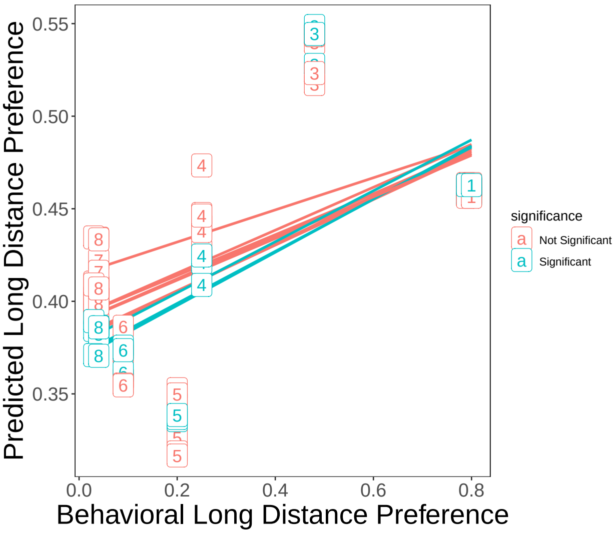


Figure A3. LSTM model predictions (across 10 runs) vs. the behavioral results from De Villiers et al. (2008). The runs are grouped according to whether the linear regression–predicting behavioral scores from model output–yielded a statistically significant result ($p < 0.05$).

Interestingly, the best and worst model runs for the Adjunct island behavioral patterns were also the best and worst runs for these child preference patterns. As with the Adjunct behavioral patterns, the complementizer lexical items seem to offer one explanation for the predicted child preference patterns. In particular, the failing runs fail to detect that unseen complementizers (e.g., “what”, “how”, “where”) are much worse than other lexical item options.

References

Arnon, I. (2021). The Starting Big approach to language learning. *Journal of Child Language*, *48*(5), 937–958.

Arnon, I., & Clark, E. V. (2011). Why brush your teeth is better than teeth–Children’s word production is facilitated in familiar sentence-frames. *Language Learning and Development*, *7*(2), 107–129.

Behm, L., Turk-Browne, N. B., & Kibbe, M. M. (2025). The ubiquity of episodic-like memory during infancy. *Trends in Cognitive Sciences*.

Boeckx, C. (2012). *Syntactic islands*. Cambridge University Press.

Borsley, R. D., & Crysmann, B. (2021). Unbounded dependencies. In *Head-driven phrase structure grammar: The handbook* (2nd ed., Vol. 9, pp. 571–634). Berlin: Language Science Press.

Chase, W. G., & Simon, H. A. (1973). Perception in chess. *Cognitive psychology*, *4*(1), 55–81.

Chater, N., & Vitányi, P. (2007). ‘Ideal learning’of natural language: Positive results about learning from positive evidence. *Journal of Mathematical psychology*, *51*(3), 135–163.

Chomsky, N. (1981). Lectures on Government and Binding.

Chomsky, N., Anderson, S., & Kiparsky, P. (1973). Conditions on transformations. *1973*, 232–286.

Cuneo, N., & Goldberg, A. E. (2023). The discourse functions of grammatical constructions explain an enduring syntactic puzzle. *Cognition*, *240*, 105563.

Davis, K. F., Parker, K. P., & Montgomery, G. L. (2004). Sleep in infants and young children: Part one: normal sleep. *Journal of Pediatric Health Care*, *18*(2), 65–71.

Derrick, D., Mayer, C., & Gick, B. (2024). Uniformity in speech: The economy of reuse and adaptation across contexts. *Glossa: a journal of general linguistics*, *9*(1).

De Villiers, J., Roeper, T., Bland-Stewart, L., & Pearson, B. (2008). Answering hard questions: Wh-movement across dialects and disorder. *Applied Psycholinguistics*, *29*(1), 67–103.

Dickson, N. (2025). *Acquiring syntax by chunking trees: A computational account of child syntactic learning* (Unpublished doctoral dissertation). University of California, Irvine.

Dickson, N., Futrell, R., & Pearl, L. (2024). I Forgot but It's Okay: Learning about Island Constraints under Child-Like Memory Constraints. In *The proceedings of the 48th annual boston university conference on language development.*

Dickson, N., Pearl, L., & Futrell, R. (2022). Learning constraints on wh-dependencies by learning how to efficiently represent wh-dependencies: A developmental modeling investigation with Fragment Grammars. *Proceedings of the Society for Computation in Linguistics*, *5*(1), 220–224.

Ding, N. (2025). Sequence chunking through neural encoding of ordinal positions. *Trends in Cognitive Sciences*.

Dowman, M. (2000). Addressing the Learnability of Verb Subcategorization with Bayesian Inference. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 22).

Eisenstein, J. (2018). Natural language processing. *Jacob Eisenstein*, *507*.

Engdahl, E. (1983). Parasitic gaps. *Linguistics and philosophy*, 5–34.

Fandakova, Y., Sander, M. C., Werkle-Bergner, M., & Shing, Y. L. (2014). Age differences in short-term memory binding are related to working memory performance across the lifespan. *Psychology and Aging*, *29*(1), 140.

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.

Foraker, S., Regier, T., Khetarpal, N., Perfors, A., & Tenenbaum, J. (2009). Indirect evidence and the poverty of the stimulus: The case of anaphoric one. *Cognitive Science*, *33*(2), 287–300.

Forsythe, H., & Pearl, L. (2020). Immature representation or immature deployment? Modeling child pronoun resolution. *Society for Computation in Linguistics*, *3*(1).

Frank, M. C., Goldwater, S., Griffiths, T. L., & Tenenbaum, J. B. (2010). Modeling human performance in statistical word segmentation. *Cognition*, *117*(2), 107–125.

Frank, S., Goldwater, S., & Keller, F. (2013). Adding sentence types to a model of syntactic category acquisition. *Topics in Cognitive Science*, *5*(3), 495–521.

Freudenthal, D., Gobet, F., & Pine, J. M. (2024). MOSAIC+: A Crosslinguistic Model of Verb-Marking Errors in Typically Developing Children and Children With Developmental Language Disorder. *Language Learning*, *74*(1), 111–145.

Freudenthal, D., Pine, J. M., & Gobet, F. (2006). Modeling the development of children's use of optional infinitives in Dutch and English using MOSAIC. *Cognitive Science*, *30*(2), 277–310.

Freudenthal, D., Pine, J. M., Jones, G., & Gobet, F. (2015). Simulating the cross-linguistic pattern of Optional Infinitive errors in children's declaratives and Wh-questions. *Cognition*, *143*, 61–76.

Futrell, R., & Mahowald, K. (2025). How linguistics learned to stop worrying and love the language models. *arXiv preprint arXiv:2501.17047*.

Gagliardi, A., Feldman, N. H., & Lidz, J. (2017). Modeling statistical insensitivity: Sources of suboptimal behavior. *Cognitive Science*, *41*(1), 188–217.

Gathercole, S. E., Pickering, S. J., Ambridge, B., & Wearing, H. (2004). The structure of working memory from 4 to 15 years of age. *Developmental psychology*, *40*(2), 177.

Gazdar, G. (1985). *Generalized phrase structure grammar*. Harvard University Press.

Goldberg, A., Cuneo, N., & Fergus, A. (2024). Addressing a challenge to the Backgroundedness account of islands. *DOI: https://doi.org/10.31234/osf. io/hmc9n*.

Goldberg, A. E. (1995). Constructions: A construction grammar approach to argument structure. *University of Chicago*.

Goldberg, A. E. (2006). *Constructions at work: The nature of generalization in language.* Oxford University Press.

Goldwater, S., & Griffiths, T. (2007). A fully Bayesian approach to unsupervised part-of-speech tagging. In *Annual meeting-Association for Computational Linguistics* (Vol. 45, p. 744).

Griffiths, T. L., Chater, N., & Tenenbaum, J. B. (2024). *Bayesian models of cognition: Reverse engineering the mind*. MIT Press.

Grosu, A. (1973). On the nonunitary nature of the coordinate structure constraint. *Linguistic Inquiry*, *4*(1), 88–92.

Gulrajani, A., & Lidz, J. (2024). Reassessing a model of syntactic island acquisition. In *Proceedings of the society for computation in linguistics 2024* (pp. 43–51).

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

Harmon, Z., Barak, L., Shafto, P., Edwards, J., & Feldman, N. H. (2021). Making heads or tails of it: a competition–compensation account of morphological deficits in language impairment. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 43).

Hirzel, M. R. (2022). *Island constraints: What is there for children to learn?* (Unpublished doctoral dissertation). University of Maryland, College Park.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, *9*(8), 1735–1780.

Hockenmaier, J., & Steedman, M. (2007). CCGbank: a corpus of CCG derivations and dependency structures extracted from the Penn Treebank. *Computational Linguistics*, *33*(3), 355–396.

Huang, J.-T. (1982). Logical relations in Chinese and the theory of grammar. *Doctoral dissertation, MIT*.

Jessop, A., Pine, J., & Gobet, F. (2025). Chunk-based incremental processing and learning: An integrated theory of word discovery, implicit statistical learning, and speed of lexical processing. *Psychological Review*.

Johnson, M., Griffiths, T., & Goldwater, S. (2007). Adaptor grammars: A framework for specifying compositional nonparametric Bayesian models. *Advances in neural information processing systems*, *19*.

Kobzeva, A., & Kush, D. (2025). Acquiring constraints on filler-gap dependencies from structural collocations: Assessing a computational learning model of island-insensitivity in Norwegian. *Language Acquisition*, 1–44.

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).

Lau, J. H., Clark, A., & Lappin, S. (2015). Unsupervised prediction of acceptability judgements. In *Proceedings of the 53rd annual meeting of the association for computational linguistics and the 7th international joint conference on natural language processing (volume 1: Long papers)* (pp. 1618–1628).

Lau, J. H., Clark, A., & Lappin, S. (2017). Grammaticality, acceptability, and probability: A probabilistic view of linguistic knowledge. *Cognitive science*, *41*(5), 1202–1241.

Lieder, F., & Griffiths, T. L. (2020). Resource-rational analysis: Understanding human cognition as the optimal use of limited computational resources. *Behavioral and brain sciences*, *43*, e1.

Liu, Y., Ryskin, R., Futrell, R., & Gibson, E. (2022). A verb-frame frequency account of constraints on long-distance dependencies in English. *Cognition*, *222*, 104902.

Matchin, W., Almeida, D., Hickok, G., & Sprouse, J. (2025). A Functional Magnetic Resonance Imaging Study of Phrase Structure and Subject Island Violations. *Journal of Cognitive Neuroscience*, *37*(2), 414–442.

Mayer, C. (2021). Capturing gradience in long-distance phonology using probabilistic tier-based strictly local grammars. In *Proceedings of the society for computation in linguistics 2021* (pp. 39–50).

McCauley, S. M., & Christiansen, M. H. (2019). Language learning as language use: A cross-linguistic model of child language development. *Psychological review*, *126*(1), 1.

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological review*, *63*(2), 81.

Momma, S., & Dillon, B. (2023). Discourse factors do not explain islands. *Available at SSRN 4635713*.

Nguyen, E., & Pearl, L. (2019). Using developmental modeling to specify learning and representation of the passive in English children. In *Proceedings of the boston university conference on language development* (Vol. 43, pp. 469–482).

O'Donnell, T., Snedeker, J., Tenenbaum, J., & Goodman, N. (2011). Productivity and reuse in language. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 33).

O'Donnell, T. J. (2015). *Productivity and reuse in language: A theory of linguistic computation and storage*. MIT Press.

O'Donnell, T. J., Tenenbaum, J. B., & Goodman, N. D. (2009). Fragment grammars: Exploring computation and reuse in language.

Omaki, A., Davidson White, I., Goro, T., Lidz, J., & Phillips, C. (2014). No fear of commitment: Children's incremental interpretation in English and Japanese wh-questions. *Language Learning and Development*, *10*(3), 206–233.

Paris, S. G. (1978). The development of inference and transformation as memory operations. In *Memory development in children* (pp. 129–156). Psychology Press.

Pearl, L. (2022). Poverty of the stimulus without tears. *Language Learning and Development*, *18*(4), 415–454.

Pearl, L. (2023a). Computational cognitive modeling for syntactic acquisition: Approaches that integrate information from multiple places. *Journal of Child Language*, *50*(6), 1353–1373.

Pearl, L. (2023b). Modeling syntactic acquisition. In J. Sprouse (Ed.), *The Oxford Handbook of Experimental Syntax* (pp. 209–270).

Pearl, L., & Bates, A. (2022a). A new way to identify if variation in children's input could be developmentally meaningful: Using computational cognitive modeling to assess input across socio-economic status for syntactic islands. *Journal of Child Language*, *51*(4), 800–833.

Pearl, L., & Bates, A. (2022b). A new way to identify if variation in children's input could be developmentally meaningful: Using computational cognitive modeling to assess input across socio-economic status for syntactic islands. *Journal of Child Language*, *51*(4), 800–833.

Pearl, L., & Forsythe, H. (2025). *Learning to be inaccurate like an adult: Using computational cognitive modeling to investigate the acquisition of pronoun interpretation in spanish.* University of California, Irvine. Available online: https://ling.auf.net/lingbuzz/006141 (accessed on).

Pearl, L., Goldwater, S., & Steyvers, M. (2010). Online learning mechanisms for Bayesian models of word segmentation. *Research on Language and Computation*, *8*, 107–132.

Pearl, L., Ho, T., & Detrano, Z. (2017). An argument from acquisition: Comparing English metrical stress representations by how learnable they are from child-directed speech. *Language Acquisition*, *24*(4), 307–342.

Pearl, L., & Sprouse, J. (2013). Syntactic islands and learning biases: Combining experimental syntax and computational modeling to investigate the language acquisition problem. *Language Acquisition*, *20*(1), 23–68.

Pearl, L., & Weinberg, A. (2007). Input filtering in syntactic acquisition: Answers from language change modeling. *Language learning and development*, *3*(1), 43–72.

Pearl, L. S. (2021). *How statistical learning can play well with Universal Grammar.* Wiley Online Library.

Pearl, L. S., & Mis, B. (2016). The role of indirect positive evidence in syntactic acquisition: A look at anaphoric one. *Language*, *92*(1), 1–30.

Pearl, L. S., & Sprouse, J. (2019). Comparing solutions to the linking problem using an integrated quantitative framework of language acquisition: Supplementary material. *Language*, *95*(4).

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

Perkins, L., Feldman, N., & Lidz, J. (2017). Learning an input filter for argument structure acquisition. In *Proceedings of the 7th workshop on cognitive modeling and computational linguistics (cmcl 2017)* (pp. 11–19).

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

Perruchet, P., Poulin-Charronnat, B., Tillmann, B., & Peereman, R. (2014). New evidence for chunk-based models in word segmentation. *Acta psychologica*, *149*, 1–8.

Perruchet, P., & Vinter, A. (1998). PARSER: A model for word segmentation. *Journal of memory and language*, *39*(2), 246–263.

Phillips, L., & Pearl, L. (2015). The utility of cognitive plausibility in language acquisition modeling: Evidence from word segmentation. *Cognitive science*, *39*(8), 1824–1854.

Pinker, S. (1999). *Words and rules: The ingredients of language*. New York, NY: Harper Collins.

Pitman, J., & Yor, M. (1997). The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator. *The Annals of Probability*, 855–900.

Pulido, M. F. (2021). Individual chunking ability predicts efficient or shallow L2 processing: Eye-tracking evidence from multiword units in relative clauses. *Frontiers in Psychology*, *11*, 607621.

Ramkumar, P., Acuna, D. E., Berniker, M., Grafton, S. T., Turner, R. S., & Kording, K. P. (2016). Chunking as the result of an efficiency computation trade-off. *Nature communications*, *7*(1), 12176.

Rosenbloom, P., & Newell, A. (1982). *Learning by chunking: A production system model of practice (tech. rep. no. 82-135).* Carnegie-Mellon University Computer Science Department.

Ross, J. R. (1967). Constraints on variables in syntax.

Rowe, M. L. (2012). A longitudinal investigation of the role of quantity and quality of child-directed speech in vocabulary development. *Child development*, *83*(5), 1762–1774.

Sanborn, A. N., Griffiths, T. L., & Navarro, D. J. (2010). Rational approximations to rational models: alternative algorithms for category learning. *Psychological review*, *117*(4), 1144.

Sprouse, J., Villata, S., & Goodall, G. (2021). Island effects. *The Cambridge handbook of experimental syntax*, 227–257.

Sprouse, J., Wagers, M., & Phillips, C. (2012). A test of the relation between working-memory capacity and syntactic island effects. *Language*, 82–123.

Steedman, M., & Baldridge, J. (2011). Combinatory categorial grammar. *Non-Transformational Syntax: Formal and Explicit Models of Grammar*, 181–224.

Thalmann, M., Souza, A. S., & Oberauer, K. (2019). How does chunking help working memory? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *45*(1), 37.

Tomasello, M. (2001). First steps toward a usage-based theory of language acquisition.

Warstadt, A., Choshen, L., Mueller, A., Williams, A., Wilcox, E., & Zhuang, C. (2023). Call for Papers–The BabyLM Challenge: Sample-efficient pretraining on a developmentally plausible corpus. *arXiv preprint arXiv:2301.11796*.

Winckel, E., Abeillé, A., Hemforth, B., & Gibson, E. (2025). Discourse-based constraints on long-distance dependencies generalize across constructions in English and French. *Cognition*, *254*, 105950.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

https://doi.org/10.3390/1010000