Syntactic Islands and Universal Grammar:
A computational model of the acquisition of constraints on long-distance dependencies

Lisa Pearl & Jon Sprouse
Department of Cognitive Sciences
University of California, Irvine

Abstract

The induction problems facing language learners have played a central role in debates about the types of learning biases that exist in the human brain. Many linguists have argued that the necessary learning biases to solve these language induction problems must be both innate and language-specific (i.e., the Universal Grammar (UG) hypothesis). Though there have been several recent high-profile investigations of the necessary types of learning biases, the UG hypothesis is still the dominant assumption for a large segment of linguists due to the lack of studies addressing central phenomena in generative linguistics. To address this, we focus on how to learn constraints on long-distance dependencies, also known as syntactic island constraints. We use formal acceptability judgment data to identify the target state of learning for syntactic island constraints, and conduct a corpus analysis of child-directed data to affirm that there does appear to be an induction problem when learning these constraints. We then create a computational learning model that successfully learns the pattern of acceptability judgments observed in formal experiments, based on realistic input data. We then discuss the learning biases required by this model to determine if any must clearly be innate and domain-specific. We find that only one of the proposed biases could potentially be innate and domain-specific, though it could also plausibly be learned. We discuss questions raised by the nature of the linguistic knowledge that is required by this learner, as well as the consequences of this learner for the learning bias debates.

1. Introduction

Although nearly all forms of human learning face induction problems, and therefore nearly all forms of human learning are aided by various types of learning biases, the induction problems facing language learners have played a central role in the debates about the types of learning biases that exist in the human brain. Many linguists have argued that the data available to young children during the language learning process are compatible with multiple hypotheses about linguistic knowledge, resulting in an induction problem that has been given a number of different labels in the linguistics literature: the "Poverty of the Stimulus" (e.g., Chomsky, 1980; Crain, 1991; Lightfoot, 1989), the "Logical Problem of Language Acquisition" (e.g., Baker, 1981; Hornstein & Lightfoot, 1981), and "Plato's Problem" (e.g., Chomsky, 1988; Dresher, 2003). This induction problem, whatever its name, then requires one or more learning biases in order to be resolved, and the central question is simply what form those learning biases take.

Many linguists have argued that the necessary learning biases must take the form of innately specified, language-specific constraints, often corresponding to specific linguistic phenomena (e.g., anaphoric *one*: Baker, 1978; Lidz, Waxman, & Freedman, 2003; interpretation of disjunctives: Crain & Pietroski, 2002; structure dependence: Chomsky, 1965). This hypothesis is known as the *Universal Grammar (UG) hypothesis* (Chomsky, 1965). The UG hypothesis is perhaps one of the most controversial claims in the entire cognitive science of language; as such, it is perhaps unsurprising that several other types of learning biases have been proposed to explain how children solve the induction problem, such as:

(i) a sensitivity to the distributional data in the available input (e.g., Foraker, Regier, Kheterpal, Perfors, & Tenenbaum, 2009; Legate & Yang, 2007; McMurray & Hollich, 2009; Mitchener & Becker, 2011; Pearl, 2011; Pearl & Lidz, 2009; Pearl & Mis, 2011; Pearl & Weinberg, 2007; Perfors, Tenenbaum, & Regier, 2011; Pullum & Scholz, 2002; Regier & Gahl, 2004; Sakas & Fodor, 2001; Scholz & Pullum, 2002; Yang, 2002; Yang, 2004)

(ii) a preference for simpler/smaller/narrower hypotheses (e.g., Foraker et al., 2009; Mitchener & Becker, 2011; Pearl & Lidz, 2009; Pearl & Mis, 2011; Perfors et al., 2011; Regier & Gahl, 2004)

(iii) a preference for highly informative data (Fodor, 1998b; Pearl & Weinberg, 2007; Pearl, 2008)

(iv) a preference for learning in cases of local uncertainty (Pearl & Lidz, 2009)

(v) a preference for data with multiple correlated cues (Soderstrom, Conwell, Feldman, & Morgan, 2009)

Because the space of possible learning biases is so large, and because a learning model with even one UG-based bias is still a UG-theory, it is critical to be as explicit as possible about what makes a learning bias part of UG, and what makes it non-UG. We suggest that learning biases may be categorized along (at least) three dimensions:

a) Are they *domain-specific* or *domain-general* ?
b) Are they *innate* or *derived* from prior experience?
c) Are they a constraint on the *hypothesis space*, or a constraint on the *learning mechanism*?

Under this system, the UG hypothesis simply holds that there is at least one innate, domain-specific learning bias (either on the hypothesis space or on the learning mechanism). Similarly a non-UG approach would be one that contains no innate, domain-specific biases: only innate, domain-general biases, derived, domain-general biases, and derived, domain-specific biases are allowed. For example, all of the learning biases listed in (i-v) above are either innate and domain-general, or derived and domain-general, therefore would not qualify as UG-biases. However, a sensitivity to linguistic representations that are innately specified (and their distributions in the input) would be an innate and domain-specific bias, and therefore qualify as a UG-bias (e.g., Legate & Yang, 2007; Mitchener & Becker, 2011; Pearl, 2011; Pearl & Lidz, 2009; Pearl & Mis, 2011; Sakas & Fodor, 2001; Yang, 2002; Yang 2004).

There have been several recent high-profile investigations of the types of learning biases required to learn various aspects of human language. For example, Perfors et al. (2011) have shown that an ideal learner using Bayesian inference will choose hierarchical representations over other kinds of possible representations, given child-directed speech data. This then shows that children do not necessarily need to know beforehand that language uses hierarchical representations; instead, this knowledge can be derived from a domain-general sensitivity to the distributional properties of the data. Importantly, children must still know that hierarchical representations are possible – but they do not need to have competing representations ruled out a priori.[1]

As another example, a number of researchers have recently conducted computational investigations of the acquisition of English anaphoric *one* (e.g., "Look, a red bottle! Oh look, another *one*."). Regier & Gahl (2004) demonstrated that a learner using online Bayesian inference can learn the correct syntactic representation and semantic interpretation of *one* from child-directed speech, provided that the child expands the range of informative data beyond the traditional data set of unambiguous data. Their model highlights the utility of a bias to use statistical distribution information in the data and a bias to prefer simpler/smaller/narrower hypotheses when encountering ambiguous data. Pearl & Lidz (2009) discovered this was an effective strategy only as long as the child knew to ignore certain kinds of ambiguous data; therefore, they proposed a learning preference for learning in cases of local uncertainty, which would rule out the troublesome ambiguous data. Pearl & Mis (2011, submitted) discovered that expanding the range of informative data even further negated the need for the local uncertainty bias; instead, a modeled learner could reproduce the observed behavior of children as long as it recognized the distributional similarities between *one* and other referential pronouns like *it*. Notably, however, this learner did not achieve the adult knowledge state, even though it reproduced child behavior. Pearl & Mis (2011) suggest that an additional strategy is still needed to reach the adult knowledge state. One possibility is the learning strategy investigated by Foraker at el. (2009), in which an ideal Bayesian learner with detailed linguistic knowledge about the link between semantic interpretation and certain syntactic structures (syntactic complements and syntactic modifiers) was able to use the difference in distribution for *one* with these structures to converge on the correct knowledge for *one*. In the Foraker et al. (2009) model, the learning mechanism is domain-general; however, it is still unclear whether the detailed

[1] Notably, however, this does not address the induction problem traditionally associated with structure dependence, which concerns hypothesizing structure-dependent *rules* that utilize these hierarchical representations (Berwick et al., 2011).

linguistic knowledge that is assumed can be derived through domain-general means or would instead be innate and domain-specific.

These previous studies have made at least two contributions to the language learning debates. First, they have demonstrated a concrete set of methodologies for investigating the types of learning biases that are required by language learning. Specifically, by combining child-directed speech corpora with computationally explicit learning models, it is possible to parametrically test the necessity of different types of learning biases. Second, they have demonstrated that at least some basic syntactic phenomena (e.g., hierarchical representations and anaphoric *one*) could in principle be learned without innate, domain-specific biases. Notably, however, there are some lingering questions such as whether all of the assumptions of the models could be learned without innate, domain-specific biases, and whether the end-states of the models are identical to the end-states hypothesized for adult speakers.

Although these findings have substantially advanced our understanding of the acquisition of some aspects of syntax, there are at least two ways that the computational approach to the investigation of acquisition (and the UG hypothesis) can be significantly advanced. First, the phenomena that have been investigated so far are generally not considered central to the syntactic theories of UG proponents. This likely means that the theoretical consequences of the previous studies have been limited due to the (relatively) peripheral nature of the phenomena. In order to truly test the UG hypothesis, and in order for the resulting acquisition models to have a real impact on existing syntactic theories (Chomsky 1965), we need to choose a set of syntactic phenomena that are central to (UG-based) syntactic theories. Second, while the methodology for testing learning biases is relatively clear, the data required to actually perform those tests are still relatively scarce. Realistic syntactic learning models require child-directed speech corpora annotated with specific syntactic structural information, such as phrase structure trees. Unfortunately, many of the freely available corpora do not yet have this kind of syntactic annotation (though there are other types of syntactic annotation available for some corpora, such as dependency tree annotations in CHILDES (Sagae et al., 2010)). Our goal in this paper is to address these two concerns by (i) constructing a corpus of child-directed speech with the syntactic annotations that we need to test syntactic learning models, and (ii) investigating the learning biases required to learn a set of phenomena that is undeniably central to (UG-based) syntactic theories – namely, syntactic island constraints.

We began our investigation by using formal acceptability judgment experiments to identify the target state (i.e., the adult state) of learning for syntactic island constraints (based on the experiments in Sprouse, Wagers, and Phillips (2012)). Next, we syntactically annotated several corpora of child-directed speech from the CHILDES database (MacWhinney, 2000), and searched those corpora for the structures used in the experimental definition of syntactic island constraints. This step identified the data from which syntactic islands must be learned, and also served to formalize the apparent induction problem that has been claimed by linguists (a concern raised by MacWhinney, 2004; Pullum & Scholz, 2002; Sampson, 1989; 1999; and Tomasello, 2004; among others). Finally, we created a computational learning model that successfully learned the pattern of acceptability judgments observed in the formal experiments from both the child-directed speech corpora and also from syntactically annotated adult-directed speech and text corpora. The question then is how to categorize the biases required by this learner. Anticipating the discussion slightly, only two of the biases could potentially be argued to be innate and domain-specific: (i) the knowledge of a shallow distinction between types of Complementizer Phrases (e.g., CPs headed by *that* versus CPs headed by *if*), and (ii) the ability

to track sequences of phrase structure nodes in long-distance dependencies. While we are reluctant to label these as clearly UG-biases, they do raise questions as to how the fine-grained linguistic knowledge of CP types is learned, and how it is that the system "knows" to attend to sequences of nodes during the parsing of long-distance dependencies. We will suggest that these sophisticated biases may arise based on the interaction of the other independently motivated biases. To the extent that those interactions are plausible, this model would suggest that syntactic island constraints can be learned without the need for any innate, domain-specific (i.e., UG) biases.

With this basic methodology in place, the rest of this article is organized as follows: Section 2 provides both a brief introduction to syntactic island constraints, and a discussion of the formal acceptability judgment experiments (from Sprouse et al. (2012)) that we used as the target state of learning. Section 3 provides a discussion of the syntactic annotation process and the results of the structural search of the child-directed speech corpora. Section 4 reports the details of the statistical learner that we propose, and the results of training this learner on the child-directed speech corpora and also on adult-directed speech and text corpora. Section 5 provides a general discussion of the nature of the linguistic knowledge that is required by this learner, as well as the consequences of this learner for the learning bias debates. Section 6 concludes.

2. A brief introduction to syntactic island effects

One of the most interesting aspects of the syntax of human languages is the fact that dependencies can exist between two non-adjacent items in a sentence. For example, in English, Noun Phrases (NPs) typically appear adjacent (or nearly adjacent) to the verbs that select them as semantic arguments (e.g., "Jack likes Lily."). However, in English *wh*-questions, *wh*-words do not appear near the verb that selects them as semantic arguments. Instead, *wh*-words appear at the front of the sentence (1a), resulting in a long-distance dependency between the *wh*-word and the verb that selects it (we will mark the canonical position of the *wh*-word, which is often called the *gap position*, with an underscore). One of the interesting aspects of these long-distance *wh*-dependencies is that they appear to be unconstrained by length (Chomsky, 1965; Ross, 1967): The distance between the *wh*-word and the verb that selects it can be increased by any number of words and/or clauses (1b-d). Though there is clearly an upper bound on the number of words and/or clauses that an English speaker can keep track of during sentence processing, this restriction appears to be based on the limited nature of human working memory capacity rather than an explicit grammatical restriction on the length of *wh*-dependencies in English. In this way, linguists often describe *wh*-dependencies as *unbounded* or *long-distance* dependencies.

(1) a. What does Jack think __?
b. What does Jack think that Lily said __?
c. What does Jack think that Lily said that Sarah heard __?
d. What does Jack think that Lily said that Sarah heard that David stole __?

Though it is true that *wh*-dependencies are unconstrained by length, they are not entirely unconstrained. Linguists have observed that if the gap position of a *wh*-dependency appears within certain syntactic structures, the resulting sentence will be unacceptable (Chomsky, 1965; Ross, 1967; Chomsky, 1973; Huang, 1982; and many others):

(2) a. *What did you make [the claim that Jack bought __]?
b. *What do you think [the joke about __] offended Jack?
c. *What do you wonder [whether Jack bought __]?
d. *What do you worry [if Jack buys __]?
e. *What did you meet [the scientist who invented __]?
f. *What did [that Jack wrote __] offend the editor?
g. *What did Jack buy [a book and __]?
h. *Which did Jack borrow [__ book]?

Drawing on the metaphor that the relevant syntactic structures are *islands* that prevent the *wh*-word from *moving* to the front of the sentence, Ross (1967) called the unacceptability that arises in these constructions *island effects*, and the syntactic constraints that he proposed to capture them *island constraints*. Though island effects are typically exemplified by *wh*-dependencies, it should be noted that island effects arise with several different types of long-distance dependencies in human languages, such as relative-clause formation (3), topicalization (4), and adjective-*though* constructions (5):

(3) a. I like the car that you think [that John bought __].
b. *I like the car that you wonder [whether John bought __].

(4) a. I don't know who bought most of these cars, but that car, I think [that John bought __].
b. *I know who bought most of these cars, but that car, I wonder [whether John bought __]?

(5) a. Smart though I think [that John is __], I don't trust him to do simple math.
c. *Smart though I wonder [whether John is __], I trust him to do simple math.

In the 45 years since island effects were first investigated (Chomsky, 1965; Ross, 1967), there have been literally hundreds of articles in dozens of languages devoted to the investigation of island effects, resulting in various proposals regarding the nature of island constraints (e.g., Abrusan, 2011; Chomsky, 2001; Deane, 1991; Erteschik-Shir, 1973; Goldberg, 2007; Hagstrom, 1998; Kluender & Kutas, 1993; Nishigauchi, 1990; Reinhart, 1997; Szabolcsi & Zwarts, 1993; Trueswell, 2007; Tsai, 1994; and many others), the cross-linguistic variability of island effects (e.g., Engdahl, 1980; Hagstrom, 1998; Huang, 1982; Lasnik & Saito, 1984; Rizzi, 1982; Torrego, 1984), and even the real-time processing of dependencies that contain island effects (e.g., Kluender & Kutas, 1993; Mckinnon & Osterhout, 1996; Phillips, 2006; Stowe, 1986; Traxler & Pickering, 1996; and many others). Though most of this literature is beyond the scope of the present article, it does serve to underscore the central role that syntactic island effects have played in the development of (generative) syntactic theory. Furthermore, the predominant analysis of syntactic island effects in generative syntactic theory is well known to rely on innate, domain-specific learning biases. For example, in the Government and Binding framework of the 1980s, syntacticians proposed a syntactic constraint called the *Subjacency Condition*, which basically holds that the dependency between a displaced element (e.g., a *wh*-word) and the gap position cannot cross two or more *bounding nodes* (Chomsky, 1973; Huang, 1982; Lasnik &

Saito, 1984; and many others). The definition of *bounding nodes* can vary from language to language in order to account for the various patterns of island effects that have been observed cross-linguistically. For example, the bounding nodes in English are argued to be NP (Noun Phrase) and IP (Inflection Phrase) (Chomsky, 1973), and bounding nodes in Italian and Spanish are argued to be NP and CP (Complementizer Phrase) (Rizzi, 1980; Torrego, 1984). Crucially, this framework assumes that the Subjacency Condition itself is part of UG, as are the possible options for bounding nodes (NP, IP, or CP). The language learner then simply needs to determine which bounding nodes are relevant for her specific language in order to learn syntactic island constraints. Although recent evolutions of syntactic theory have terminologically abandoned Subjacency and bounding nodes, it has been argued that modern incarnations of syntactic constraints (such as *phase impenetrability*) are essentially formal variants of the original Subjacency analysis (Boeckx & Grohmann, 2007).

Between the centrality of syntactic island effects as a topic of research in (generative) syntactic theory, and the reliance on a UG-based mechanism for their acquisition, it seems clear that syntactic island effects are an ideal case study in the role of innate, domain-specific learning biases in language acquisition. However, investigating the learning of syntactic island effects requires a formally explicit definition of the target state beyond the diacritics that are typically used to delineate unacceptable sentences in syntactic articles. To that end, we decided to explicitly construct the target state from data from Sprouse et al. (2012), who collected formal acceptability judgments for four island types using the magnitude estimation task: Complex NP islands (2a), (simple) Subject islands (2b), Whether islands (2c), and (conditional) Adjunct islands (2d). These four islands were selected by Sprouse et al. (2012) for several reasons. First, they have been argued to be captured by syntactic constraints (e.g., Subjacency or the Condition on Extraction Domains), as opposed to the island types that have historically been captured with semantic constraints (e.g., factive islands, negative islands). Second, dependencies spanning these islands are still somewhat intelligible, and so can provide a more nuanced assessment of unacceptability, rather than being complete "word salad". This is because these islands are the more acceptable incarnations of their particular types: Complex NP islands are more acceptable than Relative Clause islands, simple Subject islands are more acceptable than sentential Subject islands, Whether islands are more acceptable than Wh-islands with full *wh*-words in embedded spec-CP, and conditional Adjunct islands are more acceptable than causal Adjunct islands. Thus, a successful learner must accomplish a harder task than if these islands were the less acceptable varieties: The learner must realize that dependencies spanning these more acceptable islands are still ungrammatical when compared to grammatical dependencies, even though these island-spanning dependencies are still relatively intelligible.

The Sprouse et al. (2012) results are particularly useful for two reasons. First, the magnitude estimation task employs a continuous scale (the positive number line) for acceptability judgments, which results in gradient responses that are comparable to the probabilistic outputs of statistical learning models. Second, Sprouse et al. used a (2x2) factorial definition of each island effect, which controls for the two salient syntactic properties of island-violating sentences: (i) they contain a long-distance dependency, and (ii) they contain an island structure. By translating each of these properties into separate factors, each with two levels (dependency GAP POSITION: matrix, embedded; STRUCTURE present in question: non-island, island), Sprouse et al. were able to define island effects as a superadditive interaction of the two factors (in other words, an island effect is the additional unacceptability that arises when the two factors are combined, above and beyond the independent contribution of each factor). That is, a

syntactic island occurs when there is more unacceptability than what the EMBEDDED dependency and the presence of an ISLAND structure in the question contribute by themselves.

(6) Complex NP islands

a.	Who __ claimed that Lily forgot the necklace?	MATRIX | NON-ISLAND
b.	What did the teacher claim that Lily forgot __?	EMBEDDED | NON-ISLAND
c.	Who __ made the claim that Lily forgot the necklace?	MATRIX | ISLAND
d.	*What did the teacher make the claim that Lily forgot __?	EMBEDDED | ISLAND

(7) Subject islands

a.	Who __ thinks the necklace is expensive?	MATRIX | NON-ISLAND
b.	What does Jack think __ is expensive?	EMBEDDED | NON-ISLAND
c.	Who __ thinks the necklace for Lily is expensive?	MATRIX | ISLAND
d.	*Who does Jack think the necklace for __ is expensive?	EMBEDDED | ISLAND

(8) Whether islands

a.	Who __ thinks that Jack stole the necklace?	MATRIX | NON-ISLAND
b.	What does the teacher think that Jack stole __ ?	EMBEDDED | NON-ISLAND
c.	Who __ wonders whether Jack stole the necklace?	MATRIX | ISLAND
d.	*What does the teacher wonder whether Jack stole __ ?	EMBEDDED | ISLAND

(9) Adjunct islands

a.	Who __ thinks that Lily forgot the necklace?	MATRIX | NON-ISLAND
b.	What does the teacher think that Lily forgot __ ?	EMBEDDED | NON-ISLAND
c.	Who __ worries if Lily forgot the necklace?	MATRIX | ISLAND
d.	*What does the teacher worry if Lily forgot __ ?	EMBEDDED | ISLAND

Because the factorial definition treats island effects as a superadditive interaction of two factors, the presence of a syntactic island is also visually salient: If the acceptability of the four question types (as indicated by their z-scores) is plotted in an interaction plot, the presence of a syntactic island appears as two non-parallel lines (the left panel of Figure 1), and results in a significant statistical interaction; the absence of a syntactic island appears as two parallel lines (the right panel of Figure 1), and results in no significant statistical interaction.

Figure 1. Example graphs showing the presence (left panel) and absence (right panel) of a syntactic island using the factorial definition from Sprouse et al. (2012).

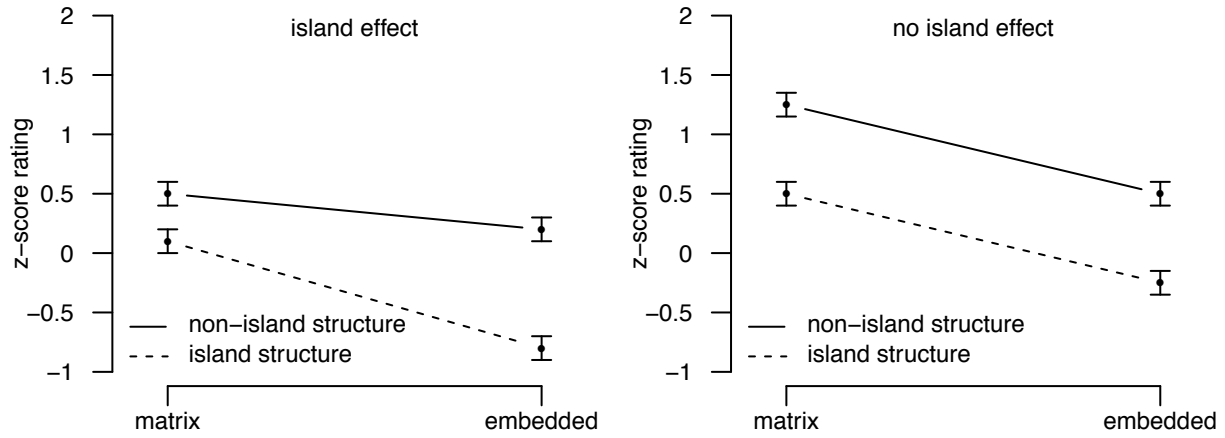
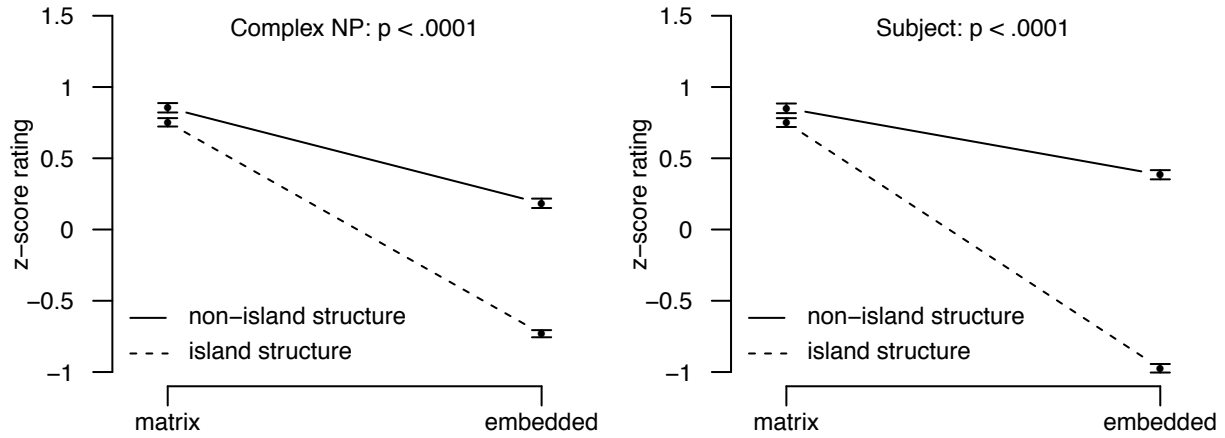
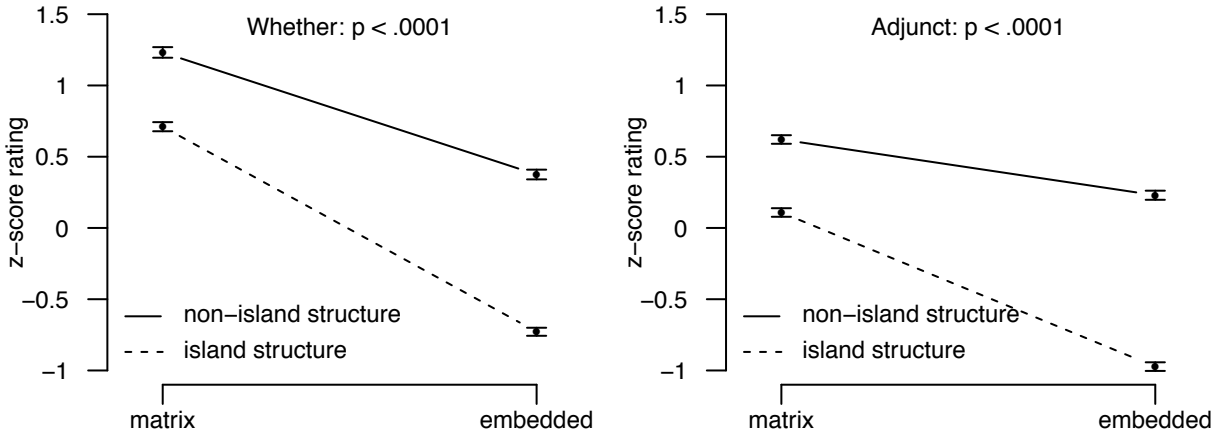


Figure 2 plots the experimentally obtained judgments for the island types investigated in Sprouse et al. (2012), which shows that adult speakers appear to have implicit knowledge of these four syntactic islands. We can thus use the superadditive interactions for the four island types in Figure 2 as an explicit target state for our statistical learner.

Figure 2. Experimentally derived acceptability judgments for the four island types from Sprouse et al. (2012) (N=173).





3. Identifying the induction problem using syntactically annotated corpora

The next step in identifying an induction problem is determining the data available to children, since this is the input they would use to reach the target state knowledge. To assess a child's input for constraints on *wh*-dependencies, we examined child-directed speech samples to determine the frequency of the structures used as experimental stimuli in Sprouse et al. (2012). While the CHILDES database has many corpora that are annotated with syntactic dependency information (Sagae, Davis, Lavie, MacWhinney, & Wintner, 2010), it is difficult to automatically extract the kind of *wh*-dependency information we needed to identify. For this reason, we selected five well-known corpora of child-directed speech from the CHILDES database (MacWhinney, 2000) to annotate with phrase structure tree information: the Adam, Eve, and Sarah corpora from the Brown data set (Brown, 1973), the Valian dataset (Valian, 1991), and the Suppes dataset (Suppes 1974). We first automatically parsed the child-directed speech utterances using a freely available syntactic parser (the Charniak parser[2]), yielding the basic phrase structure trees. However, due to the conversational nature of the data, there were many errors. We subsequently had the parser's output hand-checked by two separate annotators from a group of UC Irvine undergraduates who had syntax training, with the idea that errors that slipped past the first annotator would be caught by the second.[3] We additionally hand-checked the output of our automatic extraction scripts when identifying the frequency of *wh*-dependencies used as experimental stimuli in Sprouse et al. (2012) in order to provide a third level of error detection.

The data from these five corpora comprise child-directed speech to 25 children between the ages of one and five years old, with 813,036 word tokens total. Of all the utterances, 31,247 contained *wh*-words and verbs, and so were likely to contain syntactic dependencies. Table 1 shows the number of utterances found containing the structures and dependencies examined in Sprouse et al. (2012).

[2] Available at ftp://ftp.cs.brown.edu/pub/nlparser/.

[3] This work was conducted as part of NSF grant BCS-0843896, and the parsed corpora are available at http://www.socsci.uci.edu/~lpearl/CoLaLab/TestingUG/index.html.

Table 1. The corpus analysis of the child-directed speech samples from CHILDES, given the experimental stimuli used in Sprouse et al. (2012) for the four island types examined. The syntactic island condition (which is ungrammatical) is italicized.[4]

	MATRIX | NON-ISLAND	EMBEDDED | NON-ISLAND	MATRIX | ISLAND	*EMBEDDED | ISLAND*
Complex NP	7	295	0	0
Subject	7	29	0	0
Whether	7	295	0	0
Adjunct	7	295	15	0

From Table 1, we can see that these utterance types are fairly rare in general, with the most frequent type (EMBEDDED | NON-ISLAND) appearing 0.009% of the time (295 of 31,247). Secondly, we see that being grammatical doesn't necessarily mean an utterance type will occur in the input. Specifically, while both the MATRIX | NON-ISLAND and MATRIX | ISLAND utterance types are grammatical, they rarely occur in the input (7 for MATRIX | NON-ISLAND, either 0 or 15 for MATRIX | ISLAND). This is problematic from a learning standpoint if a learner is keying grammaticality directly to input frequency. Unless the child is very sensitive to small frequency differences (even 15 out of 31,247 is less than 0.0005% of the relevant input), the difference between the frequency of grammatical MATRIX | ISLAND or MATRIX | NON-ISLAND utterances and that of ungrammatical EMBEDDED | ISLAND utterances is very small for Adjunct island effects. It's even worse for Complex NP, Subject, and Whether island effects, since the difference between grammatical MATRIX | ISLAND utterances and ungrammatical EMBEDDED | ISLAND structures is nonexistent. Thus, it appears that child-directed speech input presents an induction problem to a learner attempting to acquire an adult grammar for dependencies crossing syntactic islands.

The existence of an induction problem then requires some sort of learning bias in order for children to end up with the correct adult grammar. We note that this induction problem arises when we assume that children are limiting their attention to direct evidence of the language knowledge of interest (something Pearl & Mis (submitted) call the *direct evidence assumption*) – in this case, utterances containing *wh*-dependencies and certain linguistic structures. One useful bias may involve children expanding their view of which data are relevant (Foraker et al., 2009; Pearl & Mis, 2011; Perfors, Tenenbaum, & Regier, 2011), and thus including *indirect positive evidence* (Pearl & Mis, submitted) for syntactic islands in their input.[5] We explore this option in the learning algorithm we describe in the next section.

---

[4] Note that the number of MATRIX | NON-ISLAND data are identical for all four island types since that control structure was identical for each island type (a *wh*-dependency linked to the subject position in the main clause, with the main clause verb (e.g., *thinks*) taking a tensed subordinate clause (e.g., *Lily forgot the necklace*)). Similarly, the number of EMBEDDED | NON-ISLAND data are identical for Complex NP, Whether, and Adjunct islands since that control structure was identical for those island types (a *wh*-dependency linked to the object position in the embedded clause, with the main clause verb taking a tensed subordinate clause).

[5] Interestingly, the idea of indirect positive evidence is similar in spirit to what linguistic parameters are meant to do in generative linguistic theory - if multiple linguistic phenomena are controlled by the same parameter, data for any of these phenomena can be treated as an

4. A statistical learning algorithm for syntactic islands

Though there appears to be an induction problem for syntactic islands, children clearly must utilize some learning procedure in order for them to become adults who have the acceptability judgments observed in Sprouse et al. (2012).[6] We first describe some necessary components for any learning algorithm, and then propose an online learning algorithm that is likely to be psychologically plausible and useful for learning about syntactic islands, paying particular attention to the learning biases that algorithm requires.

4.1. The learning algorithm in general

The essence of the acquisition process involves applying learning procedures to the available input in order to produce knowledge about language (Niyogi & Berwick, 1996; Yang, 2002; among many others). Pearl & Lidz (2009) suggest that the process can be further specified by considering the following components:

(i) children's representations of the hypothesis space
(ii) the set of input children learn from (the data *intake* (Fodor, 1998b)), and how that input set is identified and represented
(iii) the updating procedure, and how it uses the intake

Learning biases may then operate over these different components. For example, with respect to learning intuitions about syntactic islands, children could have a bias to represent their hypotheses about linguistic structures as something more abstract than licit strings of grammatical categories or licit phrase structure trees (e.g., grammatical sequences of bounding nodes: Chomsky (1973)); they could have a bias to learn from many different kinds of syntactic dependencies (indirect positive evidence: Pearl & Mis, submitted); they could have a bias to use probabilistic reasoning to update their beliefs about which structures are grammatical (Denison, Reed, & Xu, 2011; Dewar & Xu, 2010; Gerken, 2006; Griffiths & Tenenbaum, 2005; Tenenbaum & Griffiths, 2001; Xu & Tenenbaum, 2007). In a modeled learner, we can (and must) precisely specify each component of the acquisition process, including whether a bias is present and what the bias does to the hypothesis space, the input, and/or the update procedure.

Recall that the debate about the UG hypothesis revolves around one type of learning bias: innate, domain-specific biases. However, as noted in section 1, learning biases can involve any logically possible combination of the three dimensions over which biases vary. For example, a more abstract representation of linguistic structure could be derived from phrase structure trees, which themselves may be derived from distributional properties of the linguistic input by using

---

equivalence class, where learning about some linguistic phenomena yields information about others (Chomsky, 1981; Pearl & Lidz, in press; Viau & Lidz, 2011).

[6] We follow the field of syntax in assuming that well-controlled acceptability judgments can be used to infer grammaticality (see Chomsky, 1965; Schütze, 1996; Schütze & Sprouse, 2011; Sprouse & Almeida, 2012). We also follow the conclusion in Sprouse et al. (2012) that the acceptability judgment pattern observed for syntactic islands is due to grammatical constraints, and likely cannot be explained as an epiphenomenon of sentence processing.

probabilistic learning. This might then be classified as a *derived*, *domain-specific* bias about the representation of *the hypothesis space*. Probabilistic learning, in contrast, might be classified as an *innate*, *domain-general* bias about *the learning mechanism*. Crucially, only learning biases that are necessarily both *innate* and *domain-specific* are candidates for UG. A learning bias fitting this description, for example, could be an explicit innate constraint on the hypothesis space that specifically disallows dependencies that cross syntactic islands. Such a bias is *innate* by definition and *domain-specific* since it applies only to language structures. In addition, we could likely classify it as a bias about *the hypothesis space*, since it explicitly constrains the hypothesis space of the learner to exclude dependencies that cross syntactic islands. In the next section, we describe an acquisition process that does not rely on this kind of bias.

4.2. A learning process for syntactic island constraints

Turning first to the input representation, we suggest that children may be tracking the occurrence of structures that can be derived from phrase structure trees. To illustrate, the phrase structure tree for "Who did she like?" can be represented with the bracket notation in (10a), which depicts the phrasal constituents of the tree. We also assume that the learner can extract one crucial piece of information from this phrase structure tree: all of the phrasal nodes that dominate (or "contain") the gap location but not the *wh*-element associated with the gap, which we will metaphorically call the *container nodes* for the gap. A simple way to identify the container nodes is simply those phrasal constituents currently unclosed (opened with a left bracket), given the position of the gap. In (10b), the container nodes for the gap in "Who did she like?" are shown: the gap is contained by the VP "like __", which in turn is contained by the IP "she like __". The *wh*-element *who* associated with the gap is inside the CP, so the CP contains both the gap and the *wh*-element, and is therefore not a container node for the gap. We can represent this dominance information as a sequence of container nodes, as in (10c). Another example is shown in (11a-c), with the utterance "Who did she think the gift was from?" Here, the gap position associated with the *wh*-element *who* is dominated by several nodes (11b), which can be represented by the container node sequence in (11c).

Since container nodes play an integral role in all syntactic formulations of island constraints (Ross, 1967; Chomsky, 1973; etc), they seem like a necessary starting point for constructing such constraints. Furthermore, the sentence-processing literature has repeatedly established that the search for the gap location is an active process (Crain & Fodor, 1985; Stowe, 1986; Frazier & Flores d'Arcais, 1989) that tracks the container nodes of the gap location (for a more recent review, see Phillips (2006) for a list of real-time studies that have demonstrated the parser's sensitivity to island boundaries). In this way, our assumption that the learner could in principle have access to this information from the phrase structures is a well-established fact of the behavior of the human sentence parser (though it is still an open question whether paying attention to this specific information is a UG-bias or not).

(10) a. $[_{CP}$ Who did $[_{IP}$ she $[_{VP}$ like __]]]?
b. IP VP
c. IP-VP

(11) a. [$_{CP}$ Who did [$_{IP}$ she [$_{VP}$ think [$_{CP}$ [$_{IP}$ [$_{NP}$ the gift] [$_{VP}$ was [$_{PP}$ from __]]]]]]]]?
b. IP VP CP IP VP PP
c. IP-VP-CP-IP-VP-PP

In order to represent the input this way, children need the ability to parse and track dependencies in a given utterance. Work by Fodor and Sakas (Fodor, 1998a; Fodor, 1998b; Sakas & Fodor, 2001; Fodor, 2009) suggests that this ability may be useful for learning many different kinds of syntactic structures. We would likely consider this ability to be a learning bias that is *domain-specific* since it applies to language data, and a bias about the *hypothesis space* since it involves the learner representing the input in a particular way that determines the basic elements in the hypothesis space. It is likely that the process of chunking data into cohesive units is *domain-general* and *innate* (e.g., parsing visual scenes into cohesive units), though it is possible that the particular units that are being chunked (i.e., phrasal constituents) can be *derived* from distributional properties of the input.[7]

Turning to the hypothesis space, given this input representation, we propose that the hypotheses concern which container node sequences are grammatical and which are not. That is, one hypothesis might be something like "The container node sequence IP-VP is grammatical". Children's acquisition then consists of assigning some probability to each hypothesis, explicitly or implicitly. We propose a learning algorithm below that implicitly assigns a probability to each hypothesis like this, based on the form of the container node sequence. In order to represent the hypothesis space this way, children need only to represent the input in terms of these container node sequences, which comes from being able to parse and track dependencies in a given utterance. So, this again requires a learning bias that is *domain-specific* and about the *hypothesis space* (parsing into container node sequences), though the units over which this process operates are likely *derived* and the general process itself may be *domain-general*.

The learning algorithm we propose involves the learner tracking the frequency of smaller sub-sequences of container node sequences, as encountered in the input. In particular, we suggest that a learner could track the frequency of container node trigrams (i.e., a continually updated sequence of three container nodes) in the input utterances.[8] For example, the container node sequences from (10c) would be represented as a sequence of trigrams as in (12c), and the container node sequences from (11c) would be represented as a sequence of trigrams as in (13c):

[7] For example, Klein & Manning (2002) describe an unsupervised algorithm for inferring hierarchical structure, given grammatical categories as input. Work by Mintz (2003, 2006) describes a psychologically plausible algorithm for inferring grammatical categories from child-directed language input. Putting these two together would be one way of deriving the phrasal constituents used in parsing.

[8] Note that this means the learner is learning from data containing dependencies besides the one of interest, treating the other dependencies as indirect positive evidence (Pearl & Mis, submitted). For example, a learner deciding about the sequence IP-VP-CP-IP-VP would learn from IP-VP dependencies that the trigram *start-IP-VP* appears. This is a learning bias that expands the relevant intake set of the learner – all dependencies are informative, not just the ones being judged as grammatical or ungrammatical.

(12) a. [$_{CP}$ Who did [$_{IP}$ she [$_{VP}$ like __]]]?
b. IP VP
c. start-IP-VP-end =
start-IP-VP
IP-VP-end

(13) a. [$_{CP}$ Who did [$_{IP}$ she [$_{VP}$ think [$_{CP}$ [$_{IP}$ [$_{NP}$ the gift] [$_{VP}$ was [$_{PP}$ from __]]]]]]]]?
b. IP VP CP IP VP PP
c. start-IP-VP-CP-IP-VP-PP-end =
start-IP-VP
IP-VP-CP
VP-CP-IP
CP-IP-VP
IP-VP-PP
VP-PP-end

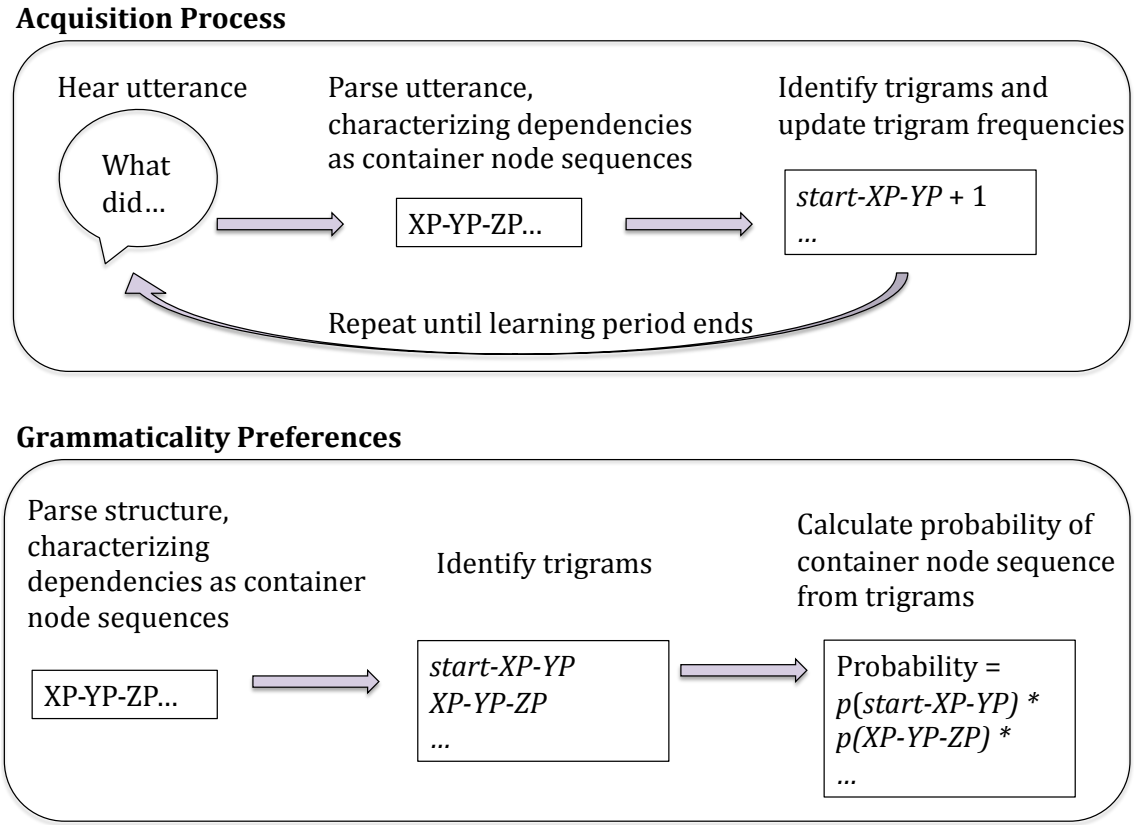
The learner generates the probability of a given container node trigram based on the observed data. Then, to gauge the grammaticality of any given container node chain (such as an island), the learner calculates the probability of observing that sequence of container node trigrams, which is simply the product of the trigram probabilities.[9] For example, in (12), the sequence IP-VP would have a probability equal to the product of the trigram *start-IP-VP* and the trigram *IP-VP-end*.

All other things being equal, this automatically makes longer dependencies less probable than shorter dependencies since more probabilities are multiplied together for longer dependencies, and those probabilities are always less than 1. Note, however, that the frequency of the individual trigrams comprising those dependencies still has a large effect. In particular, a shorter dependency that includes a sequence of very infrequent trigrams will still be less probable than a longer dependency that contains very frequent trigrams. Thus, the frequencies observed in the input temper the detrimental effect of dependency length. The learning algorithm and calculation of grammaticality preferences[10] are schematized in Figure 3, and two examples of grammaticality preferences are shown in (14) and (15).

[9] We note that the learner we implement in section 4.4 uses smoothed trigram probabilities (using Lidstone's Law (Manning & Schütze, 1999) with smoothing constant $\alpha = 0.5$), so unobserved trigrams have a frequency slightly above 0. Specifically, the learner imagines that unobserved trigrams have been observed α times, rather than 0 times, and all other trigrams have been observed α + their actual observed occurrences. We note also that the overall trend of results we observe later on does not critically depend on the value of α, which effectively serves to distinguish trigrams that rarely occur from trigrams that never occur. The smaller α is, the more these are distinguished.

[10] Here and throughout we will use the term *grammaticality preference* to refer to the result of the learning algorithm (a probability), and *acceptability judgments* to refer to the actual observed behavior of adults in an experimental setting (e.g., Sprouse et al., 2012). As discussed at the end of section 4, an acceptability judgment is the result of several factors, of which the

Figure 3. Steps in the acquisition process and calculation of grammaticality preferences.



(14) "Where does the reporter think Jack stole from?"
[$_{CP}$ Where does [$_{IP}$ [$_{NP}$ the reporter] [$_{VP}$ think [$_{CP}$ [$_{IP}$ [$_{NP}$ Jack] [$_{VP}$ stole [$_{PP}$ from __]]]]]]]?"
IP VP CP IP VP PP
Sequence: start-IP-VP-CP-IP-VP-PP-end
Trigrams: start-IP-VP
IP-VP-CP
VP-CP-IP
CP-IP-VP
IP-VP-PP
VP-PP-end
Probability(IP-VP-CP-IP-VP-PP) =
p(start-IP-VP)*p(IP-VP-CP)*p(VP-CP-IP)*p(CP-IP-VP)*p(IP-VP-PP)*p(VP-PP-end)

grammaticality preferences generated by our learner are just one. Other factors affecting acceptability judgments include semantic plausibility, lexical properties, and parsing difficultly.

(15) *"Who does Jack think the necklace for is expensive?"
[$_{CP}$ Who does [$_{IP}$ [$_{NP}$ Jack] [$_{VP}$ think [$_{CP}$ [$_{IP}$ [$_{NP}$ the necklace [$_{PP}$ for __]] [$_{VP}$ is expensive]]]]]]?

IP VP CP IP NP PP

Sequence: start-IP-VP-CP-IP-NP-PP-end

Trigrams: start-IP-VP
IP-VP-CP
VP-CP-IP
CP-IP-NP
IP-NP-PP-
NP-PP-end

Probability(IP-VP-CP-IP-NP-PP) =
p(start-IP-VP)*p(IP-VP-CP)*p(VP-CP-IP)*p(CP-IP-NP)*p(IP-NP-PP)*p(NP-PP-end)

To implement this learning algorithm, a child would need sufficient memory to hold an utterance's parse and dependencies in mind in order to extract the container node trigram sequences. This likely involves *domain-general*, *innate* memory capacities. The child also needs sufficient memory to hold three units in mind in order to track the trigram frequencies. Studies in statistical learning suggest that young children have sufficient memory capacity to track frames consisting of three units (Mintz, 2006; Wang & Mintz, 2008) and to compare three transitional probabilities (Saffran et al., 1996; Aslin et al., 1998; Saffran et al., 1999; Graf Estes et al., 2007; Saffran et al., 2008; Pelucchi et al., 2009a; 2009b). This again likely involves *domain-general*, *innate* memory capacities. We note that one concern with using trigrams in natural language processing is that the sheer number of trigrams can lead to a sparse data problem, so that the learner could not possibly hope to have enough input to observe examples of all legal trigrams.[11] However, that is not likely to be as much of a problem for the learner we propose, since we are constructing trigrams over units much more abstract than individual vocabulary items. If we have fewer than 10 container nodes (as we might if we only use IP, VP, CP, NP, PP, and AdjP as the relevant phrasal constituents), then the number of trigrams children must track is less than 10^3 (1000). We believe that this is less than the number of vocabulary items children know by the time they would be learning grammaticality preferences about dependency structures[12], and so this doesn't seem particularly taxing for children to track. The learning bias to track trigrams is likely to be *domain-general* (since trigrams can be tracked outside of language), *innate*, and about the *learning mechanism*.

Identifying which units are potential container nodes is critical to the psychological plausibility of this learning model. One possibility is that container nodes are specified by UG, similar to the original conception of bounding nodes. Another possibility is that learners may adopt an initial strategy of using the basic-level phrasal constituents noted above (derived from parsing), which is minimally taxing memory-wise. Later, if they find that their intuitions do not match the observed data, they may adopt finer-grained distinctions. The one we explore later on

[11] Additionally, tracking a huge number of trigrams may strain a learner's memory.

[12] For example, Hart & Risley (1995) suggest that a three-year-old has a lexicon of around 1000 items, and diary data from Braunwald (1978) suggests that even children as young as two may already have this number of lexicon items. All of the acquisition studies investigating islands that we are aware of do not examine children younger than three.

involves noting the complementizer used for a CP (e.g., *that*, *whether, if,* null, etc.) and subcategorizing CP container nodes based on the specific lexical item in complementizer position (e.g., CP_{that} vs. $CP_{whether}$ vs. CP_{if} vs. CP_{null}, etc.). This is, in effect, a very simplistic item-based strategy for the subcategorization of CPs. Depending on the number of fine-grained distinctions required, this may be more or less taxing on a child's memory. In terms of learning biases, this process may involve a type of simplicity strategy, where only as much detail is used as is necessary. This could then be classified as a *domain-general*, *innate* bias about the *learning mechanism.* A third possibility is that learners could subcategorize CP container nodes from the outset, perhaps because children's linguistic experience has already highlighted that different complementizers have different semantic and pragmatic implications for the clauses that follow them, and this is known by the time that long-distance dependencies are learned. This could then be classified as a *domain-specific, derived* bias about the representation of the *hypothesis space*. There are clearly several logical possibilities concerning both the time-course of the use of subcategorized CP container nodes and the reason that the learner decides to use them. We will not attempt to test each of these possibilities here; instead we will simply compare learning models that use basic-level CP container nodes to models that use subcategorized CP container nodes to establish the empirical necessity of subcategorized CP container nodes (see section 4.5.1 and 4.5.2 for the comparison, and section 5.8 for a discussion of the relationship between computational learning models and hypotheses about the time-course of acquisition). Notably, however, the subcategorization we propose is very surface-level (tied to the specific lexical item in complementizer position), as opposed to something more structurally sophisticated.

Given this learning algorithm, a child can generate a grammaticality preference for a given dependency at any point during learning, based on the input previously observed, by calculating its probability from the frequency of the trigrams that comprise it (see Figure 3). Similarly, a relative grammaticality preference can be calculated by comparing the probabilities of two dependencies' container node sequences. This will allow us, for example, to compare the inferred grammaticality of dependencies spanning island structures vs. dependencies spanning non-island structures. This ability to generate a probability for a larger structure based on its trigrams is likely to be a *domain-general*, *innate* ability about the *learning mechanism*.

Table 2 summarizes the learning biases required for the proposed learning procedure, characterizing them along the two dimensions relevant for the UG hypothesis: domain-specific vs. domain-general, and innate vs. derived. Note that none of the learning biases (or their components) appear to be both *necessarily* domain-specific and innate simultaneously. Only one bias could potentially be part of a UG-based approach to the acquisition of island constraints: the bias to attend to container nodes. In other words, whether the model is based on the UG hypothesis or not hinges on whether this bias must be innately specified or could instead arise through other means.

Table 2. Classification of the learning biases required by the proposed acquisition process. The critical bias types (domain-specific and innate) are shaded to help illustrate the fact that no process in this learning model necessarily requires a bias that is both domain-specific and innate simultaneously.

Description of process	**Domain-specific**	**Domain-general**	**Innate**	**Derived**
Parse utterance & identify dependencies	*			*
Attend to container nodes	*		?	?
Extract trigrams of container nodes		*	*	
Update probability of each trigram		*	*	
Calculate probability of utterance's dependency		*	*	

4.3 Empirically grounding the learner

Looking first to the learner's input, we should consider whose grammaticality preferences we are attempting to match. If we are modeling how children acquire their grammaticality preferences, we should look at child-directed speech. If we are instead interested in how adults acquire their preferences (perhaps because we have empirical data from adults), then we may be interested in a mix of adult-directed speech and adult-directed text. Tables 3 and 4 describe the composition of six corpora across three corpus types: child-directed speech from the Adam and Eve corpora from Brown (1973), the Valian corpus (Valian, 1991), and the Suppes corpus (Suppes, 1974) of CHILDES (MacWhinney, 2000), adult-directed speech from the Switchboard section of the Treebank-3 corpus (Marcus et al., 1999) and adult-directed text from the Brown section of the Treebank-3 corpus (Marcus et al., 1999).

Table 3: Basic composition of the child-directed and adult-directed input corpora.

	Child-directed: speech	Adult-directed: speech	Adult-directed: text
total utterances	101838	74576	24243
total *wh*-dependencies	20923	8508	4230

Table 4. Description of child-directed and adult-directed input corpora. Percentages are shown for container node sequences, based on the total *wh*-dependencies in each corpus, with the quantity observed in the corpus on the line below. An example of each container node sequence is given below the sequence.

Container node sequence and example utterance	Child-directed: speech	Adult-directed: speech	Adult-directed: text
IP Who saw it?	12.8% 2680	17.2% 1464	33.0% 1396
IP-VP What did she see?	76.7% 16039	73.0% 6215	63.3% 2677
IP-VP-AdjP-IP-VP What are you willing to see?	0.0% 0	<0.1% 1	0.1% 5
IP-VP-AdjP-IP-VP-PP What are you willing to go to?	0.0% 0	<0.1% 1	0.0% 0
IP-VP-AdjP-PP What are they good for?	0.0% 0	<0.1% 1	<0.1% 1
IP-VP-CP_{for}-IP-VP-PP What did she put on for you to dance to?	<0.1% 1	0.0% 0	0.0% 0
IP-VP-CP_{null}-IP Who did he think stole it?	0.1% 24	0.6% 52	0.3% 12
IP-VP-CP_{null}-IP-VP What did he think she stole?	1.1% 236	0.4% 30	0.2% 8
IP-VP-CP_{null}-IP-VP-IP-VP What did he think she wanted to steal?	0.1% 28	<0.1% 3	0.0% 0
IP-VP-CP_{null}-IP-VP-IP-VP-IP-VP What did he think she wanted to pretend to steal?	<0.1% 2	0.0% 0	0.0% 0
IP-VP-CP_{null}-IP-VP-IP-VP-IP-VP-PP Who did he think she wanted to pretend to steal from?	0.0% 0	<0.1% 1	0.0% 0
IP-VP-CP_{null}-IP-VP-IP-VP-PP Who did he think she wanted to steal from?	<0.1% 1	0.0% 0	0.0% 0
IP-VP-CP_{null}-IP-VP-NP What did he think she said about it?	<0.1% 1	<0.1% 5	<0.1% 1
IP-VP-CP_{null}-IP-VP-PP What did he think she wanted it for?	0.1% 28	<0.1% 5	<0.1% 1

IP-VP-CP_{null}-IP-VP-PP-PP What did he think she wanted out of?	<0.1% 1	0.0% 0	0.0% 0
IP-VP-CP_{that}-IP-VP What did he think that she stole?	<0.1% 2	<0.1% 5	<0.1% 2
IP-VP-CP_{that}-IP-VP-IP-VP What did he think that she wanted to steal?	0.0% 0	<0.1% 1	0.0% 0
IP-VP-CP_{that}-IP-VP-PP Who did he think that she wanted to steal from?	0.0% 0	<0.1% 1	0.0% 0
IP-VP-IP Who did he want to steal the necklace?	<0.1% 9	<0.1% 2	0.0% 0
IP-VP-IP-VP What did he want her to steal?	5.6% 1167	3.4% 287	1.3% 57
IP-VP-IP-VP-IP-VP What did he want her to pretend to steal?	<0.1% 11	<0.1% 6	<0.1% 1
IP-VP-IP-VP-IP-VP-PP Who did he want her to pretend to steal from?	0.2% 43	<0.1% 6	0.0% 0
IP-VP-IP-VP-NP What did he want to say about it?	<0.1% 6	0.0% 0	0.0% 0
IP-VP-IP-VP-NP-IP-VP What did he have to give her the opportunity to steal?	0.0% 0	0.0% 0	<0.1% 1
IP-VP-IP-VP-NP-PP What did she want to steal more of?	<0.1% 1	<0.1% 1	0.0% 0
IP-VP-IP-VP-PP What did she want to steal from?	0.4% 74	0.4% 33	<0.1% 4
IP-VP-IP-VP-PP-PP What did she want to get out from under?	0.0% 0	0.0% 0	<0.1% 1
IP-VP-NP What did she say about the necklace?	0.2% 52	0.1% 10	0.1% 5
IP-VP-NP-IP-VP What did he give her the opportunity to steal?	0.0% 0	<0.1% 1	<0.1% 2
IP-VP-NP-PP What was she a member of?	<0.1% 7	<0.1% 6	0.0% 0

IP-VP-PP Who did she steal from?	2.5% 524	4.3% 369	1.3% 57
IP-VP-PP-CP_{null}-IP What did she feel like was a very good place?	0.0% 0	<0.1% 1	0.0% 0
IP-VP-PP-CP_{null}-IP-VP What did she feel like he saw?	<0.1% 1	0.0% 0	0.0% 0
IP-VP-PP-IP-VP What did she think about buying?	0.0% 0	<0.1% 3	0.0% 0
IP-VP-PP-NP Where was she at in the building?	0.0% 0	<0.1% 2	0.0% 0
IP-VP-PP-NP-PP What do you put it on top of?	<0.1% 2	0.0% 0	0.0% 0
IP-VP-PP-NP-PP-IP-VP What is she in the habit of doing?	0.0% 0	<0.1% 1	0.0% 0
IP-VP-PP-PP What does he eat out of?	0.1% 22	0.0% 0	0.0% 0
IP-VP-PP-VP What did he think about stealing?	<0.1% 1	0.0% 0	0.0% 0

Notably, two sequences dominate the input, no matter what the corpus: IP-VP and IP, corresponding to main clause object and main clause subject dependencies, respectively. Interestingly, child-directed speech seems similar to adult-directed speech in terms of the proportion of *wh*-dependencies, with IP-VP dominating IP (child-directed speech: 76.7%/12.8%, adult-directed speech: 73.0%/17.2%). This suggests that, at this level of abstraction, child-directed speech and adult-directed speech are fairly equivalent, which is not necessarily the case if we look at less abstract representations such as complete phrase structure trees, grammatical category sequences, or vocabulary items. In contrast, adult-directed written text tends to be biased slightly more towards main clause subject dependencies (IP), though main clause object dependencies (IP-VP) are still far more prevalent (IP-VP: 63.3% to IP: 33.0%). Also, we note that overt complementizers (such as *that*, indicated with CP_{that} in Table 4) are rare in general. This will become relevant when we examine the learned grammaticality preferences for dependencies involving the complementizer *that*.

Turning to the learning period for our modeled learners, we can draw on empirical data from Hart & Risley (1995) and assume children hear approximately 1 million utterances between birth and 3 years of age. If we assume our learners' learning period is approximately 3 years (perhaps between the ages of 2 and 5 years old, if we're modeling children's acquisition), we can estimate the number of *wh*-dependencies they hear out of those one million utterances. Given child-directed speech samples from Adam and Eve (Brown 1973), Valian (Valian 1991), and Suppes (Suppes, 1974), we estimate the proportion of *wh*-dependencies (20,923) to total utterances (101,823) as approximately 0.2. We thus set the learning period to 200,000 *wh*-

dependency data points. So, our learners will encounter 200,000 data points containing *wh*-dependencies, drawn randomly from a distribution characterized by the corpora in table 4.

4.4 Success metrics and learner implementation

We can test our modeled learners by comparing their learned grammaticality preferences to empirical data on adult acceptability judgments from Sprouse et al. (2012). The container node sequence that arises for the sentence types in (6-9) above is given in (16-19). As we can see from (16-19), our modeled learners will compare the dependencies spanning island structures to only three container node sequences, despite the different sentence types involved: IP, IP-VP-CP/CP_{that}-IP-VP, and IP-VP-CP/CP_{null}-IP.[13]

(16) Complex NP islands

a.	IP	MATRIX | NON-ISLAND
b.	IP-VP-CP/CP_{that}-IP-VP	EMBEDDED | NON-ISLAND
c.	IP	MATRIX | ISLAND
d.	*IP-VP-NP-CP/CP_{that}-IP-VP	EMBEDDED | ISLAND

(17) Subject islands

a.	IP	MATRIX | NON-ISLAND
b.	IP-VP-CP/CP_{null}-IP	EMBEDDED | NON-ISLAND
c.	IP	MATRIX | ISLAND
d.	*IP-VP-CP/CP_{null}-IP-NP-PP	EMBEDDED | ISLAND

(18) Whether islands

a.	IP	MATRIX | NON-ISLAND
b.	IP-VP-CP/CP_{that}-IP-VP	EMBEDDED | NON-ISLAND
c.	IP	MATRIX | ISLAND
d.	*IP-VP-CP/$CP_{whether}$-IP-VP	EMBEDDED | ISLAND

[13] This shows that generating an acceptability judgment is likely more nuanced than how our modeled learners implement it here, since the portion of the utterance beyond the gap position influences human judgments. For example, *Who saw it?* is not judged equivalent to *Who thought that Jack said that Lily saw it*?, even though both are IP dependencies. This is why experimental studies have to balance the structures involved in the utterances, as Sprouse et al. (2012) did. In contrast, a learner using the container node sequence representation judges all utterances with equivalent dependencies as equally grammatical, which is why several control structures have the same container node sequence (see also the discussion in section 5).

(19) Adjunct islands

a.	IP	MATRIX | NON-ISLAND
b.	IP-VP-CP/CP_{that}-IP-VP	EMBEDDED | NON-ISLAND
c.	IP	MATRIX | ISLAND
d.	*IP-VP-CP/CP_{if}-IP-VP	EMBEDDED | ISLAND

Recall that this factorial definition of island effects makes the presence of island effects visually salient. If the acceptability of the four utterance types is plotted in an interaction plot, the presence of an island effect shows up as two non-parallel lines (e.g., the left panel of Figure 1), while the absence of an island effect shows up as two parallel lines (e.g., the right panel of Figure 1). Sprouse et al. (2012) found an island effect pattern for all four island types.

To evaluate the success of our modeled learners, we can plot the predicted grammaticality preferences in a similar interaction plot: If the lines are non-parallel, then the learner has acquired the knowledge required to implement island constraints; if the lines are parallel, then the learner did not acquire the knowledge required to implement island constraints. All our modeled learners will follow the learning algorithm and grammaticality preference calculation outlined in Figure 3. In particular, they will receive data incrementally, identify the container node sequence and trigrams contained in that sequence, and update their corresponding trigram frequencies. They will then use these trigram frequencies to infer a probability for a given *wh*-dependency, which can be equated to its judged acceptability – more probable dependencies are more acceptable, while less probable dependencies are less acceptable. Though the inferred acceptability can be generated at any point during learning (based on the trigram frequencies at that point), we will show results only from the end of the learning period.

4.5 Modeling results: When island constraints can be learned

Because the result of a grammaticality preference calculation is often a very small number (due to multiplying many probabilities together), we will instead report the log probability. This allows for easier comparison with acceptability judgments. All log probabilities are negative. The more positive numbers (i.e. closer to zero) represent "more acceptable" structures while more negative numbers (i.e., farther from zero) represent "less acceptable" structures.[14] We will first look at modeled learners who use only basic-level container nodes (e.g., CP), and then at learners who use finer-grained container nodes (e.g., $CP_{\textit{that}}$).

4.5.1 Basic-level container nodes

As a first learning model, we will only assume that basic-level container nodes are distinguished by the learner. This means that all CP nodes are represented as CP, irrespective of what complementizer is used (i.e., both $CP_{\textit{that}}$ and $CP_{\textit{whether}}$ are represented as a single node type: CP). As we will see, this assumption has detrimental consequences for the success of the learner.

[14] This measurement is similar to *surprisal*, which is traditionally defined as the negative log probability of occurrence (Tribus, 1961) and has been used recently within the sentence processing literature (Hale, 2001; Jaeger & Snider, 2008; Levy, 2008; Levy, 2011). Under this view, less acceptable dependencies are more surprising.

Figure 4 shows the learner's grammaticality preferences for the dependencies from Sprouse et al. (2012), based on child-directed input and represented with log probabilities. Figure 5 shows the learner's grammaticality preferences based on adult-directed input. Table 5 reports the log probabilities depicted in Figures 4 and 5.

Figure 4. Log probabilities derived from child-directed speech for a learner that does not discriminate CP node types. The apparent lack of dashed "island structure" line in the Whether and Adjunct island graphs indicates that the line is identical to the solid "non-island" structure line, as can be seen from the overlapping endpoints.

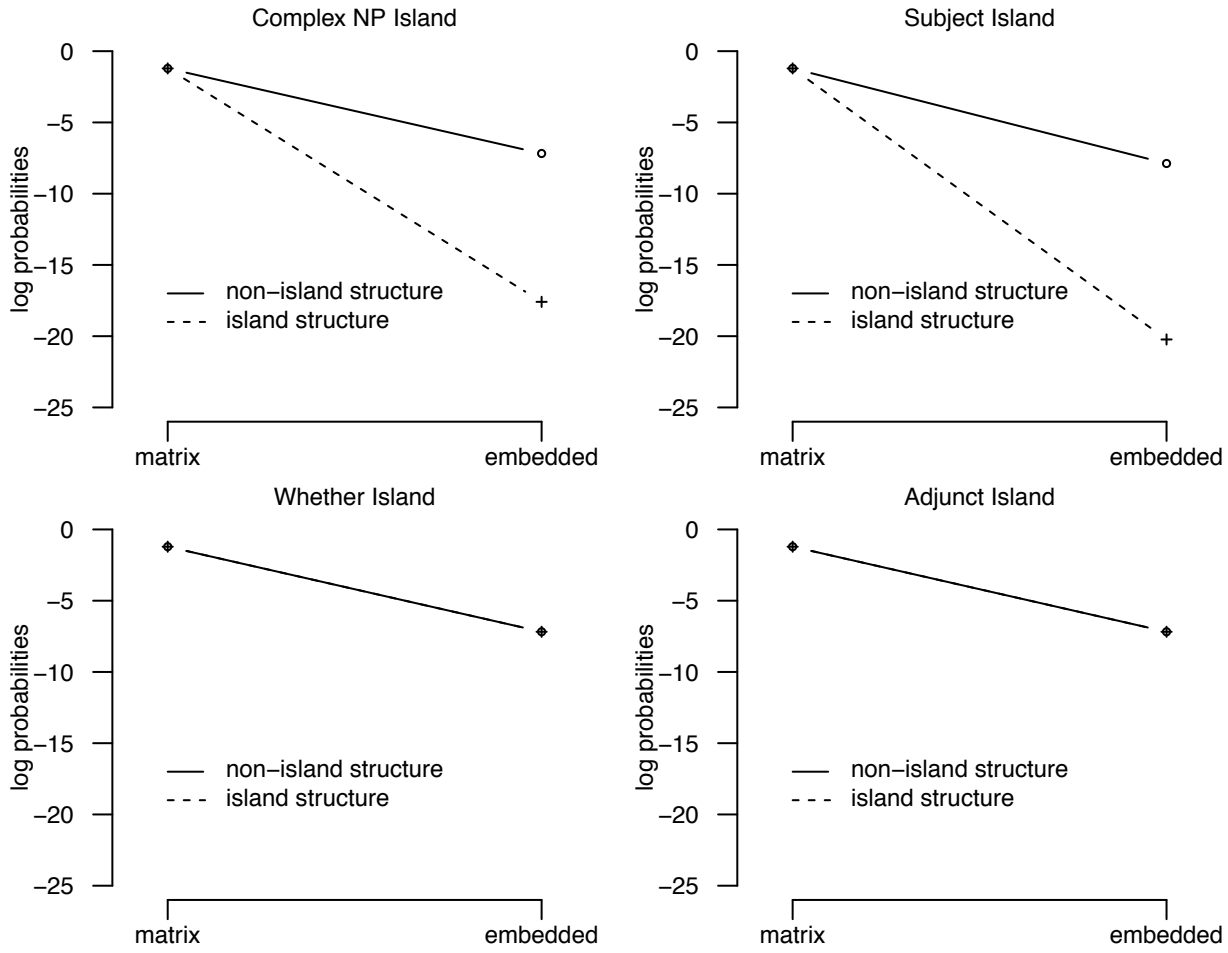


Figure 5. Log probabilities derived from adult-directed speech and text for a learner that does not discriminate CP node types. The apparent lack of dashed "island structure" line in the Whether and Adjunct island graphs indicates that the line is identical to the solid "non-island" structure line, as can be seen from the overlapping endpoints.

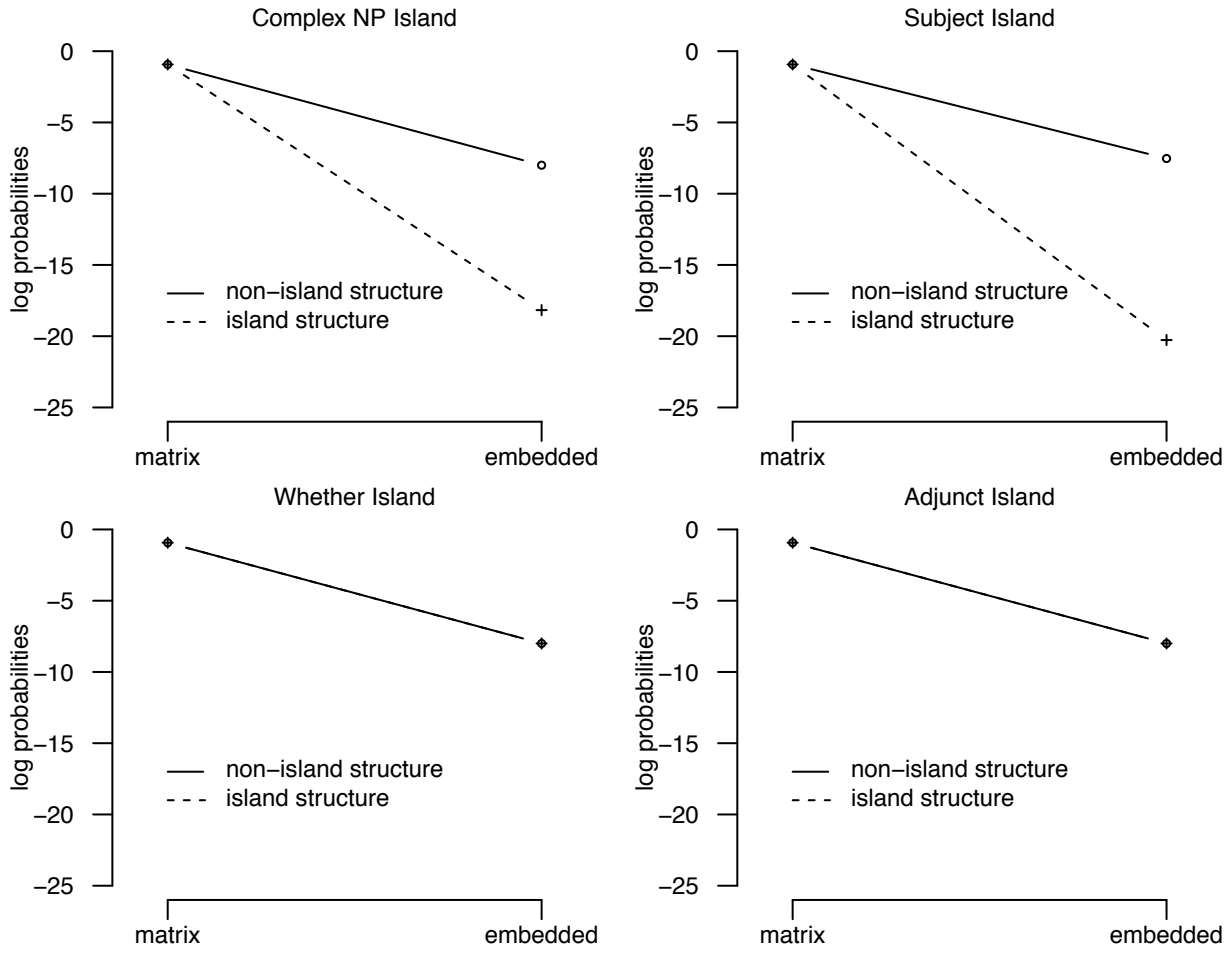


Table 5. Inferred acceptability of different *wh*-dependencies from Sprouse et al. (2012), represented with log probability.

		Child-directed speech	Adult-directed speech & text
Control dependencies			
matrix subject	IP	-1.21	-0.93
embedded subject	IP-VP-CP-IP	-7.88	-7.53
embedded object	IP-VP-CP-IP-VP	-7.18	-8.00
Island-spanning dependencies			
Complex NP	IP-VP-NP-CP-IP-VP	-17.59	-18.17
Subject	IP-VP-CP-IP-NP-PP	-20.23	-20.27
Whether	IP-VP-CP-IP-VP	-7.18	-8.00
Adjunct	IP-VP-CP-IP-VP	-7.18	-8.00

Table 6 reports the log odds comparison ($\log(prob_1/prob_2)$) between the control dependencies and the dependencies spanning island structures, given the structures used in Sprouse et al. (2012). This provides a direct way to compare the relative inferred grammaticality preferences of different dependencies, according to our modeled learners. Positive numbers mean the first structure (with $prob_1$) is more probable, while negative numbers mean that the second structure (with $prob_2$) is more probable. For example, a log odds of x would mean that the first structure is x times more probable (grammatical) than the second structure, while a log odds of $-x$ would mean the second structure is x times more probable (grammatical) than the first structure.

Table 6. Relative acceptability of different *wh*-dependencies, based on the log odds of the inferred probabilities. Numbers represent the comparison of the control dependency in the row (as $prob_1$) to the island-spanning dependency in the column (as $prob_2$).

		Island-spanning dependencies			
		Complex NP	Subject	Whether	Adjunct
Control dependencies	Child-directed speech				
	matrix subject	16.38	19.02	5.97	5.97
	embedded subject	--	12.35	--	--
	embedded object	10.41	--	0.00	0.00
	Adult-directed speech & text				
	matrix subject	17.24	19.34	7.08	7.08
	embedded subject	--	12.73	--	--
	embedded object	10.17	--	0.00	0.00

Figure 4, Figure 5, and Table 5 show that our modeled learners using child-directed speech (Figure 4) or adult-directed input (Figure 5), with no distinction between CP node types, can learn the correct grammaticality preferences for two of the four islands examined: Complex NP and Subject islands. Both of these island types show the non-parallel lines that indicate an interaction in Figures 5 and 6, and all control dependencies are significantly more grammatical (by a factor of at least 10) than the island spanning dependencies (Table 5, first two columns). However, these learners fail to distinguish Whether and Adjunct islands from the control structures. Not only are the lines parallel in figures 5 and 6, indicating no interaction, but also overlapping (resulting in graphs that appear to only contain one line). Table 6 shows that at least one control structure (embedded object, IP-VP-CP-IP-VP) is viewed as equally grammatical to the dependencies spanning Whether and Adjunct islands (Table 6, last two columns). Upon closer inspection, this is not surprising because the learner does not distinguish between structures with the sequence IP-VP-$CP_{null/that}$-IP-VP and structures with the sequence IP-VP-$CP_{whether/if}$-IP-VP , which means that Whether and Adjunct island violations, which contain specific types of CPs ($CP_{whether}$ and CP_{if}), are treated identically to grammatical utterances containing CP_{null} or CP_{that}, such as "What did he think (that) she saw?".

4.5.2 Finer-grained container nodes: CP-specification

We implemented a second learner that allowed for finer distinctions among the CP nodes. In particular, this learner distinguishes CP nodes by the complementizer that appears in the CP, such as *that, whether, if*, etc. For this learner, Whether islands will be represented as IP-VP-$CP_{whether}$-IP-VP and Adjunct islands as IP-VP-CP_{if}-IP-VP. It is widely assumed that children must keep track of the lexical content of complementizers, as the choice of complementizer has both syntactic and semantic consequences for sentences. In this case, we are further assuming that children include this information to distinguish different sequences of container nodes. As this is clearly a relatively sophisticated linguistic assumption, we will discuss it, and whether it could be considered part of the UG hypothesis, in more detail in section 5.

For this second model, acceptable dependencies will appear as IP-VP-CP_{null}-IP-VP or IP-VP-CP_{that}-IP-VP, which will allow our learners to distinguish these from the island-spanning dependencies. Figures 6 and 7 represent the results of this kind of learner, given child-directed and adult-directed data as input, respectively. Table 7 lists the log probabilities depicted in Figures 6 and 7, while Table 8 shows the log odds comparison between control dependencies and island-spanning dependencies.

Figure 6. Log probabilities derived from child-directed speech for a learner that discriminates CP types.

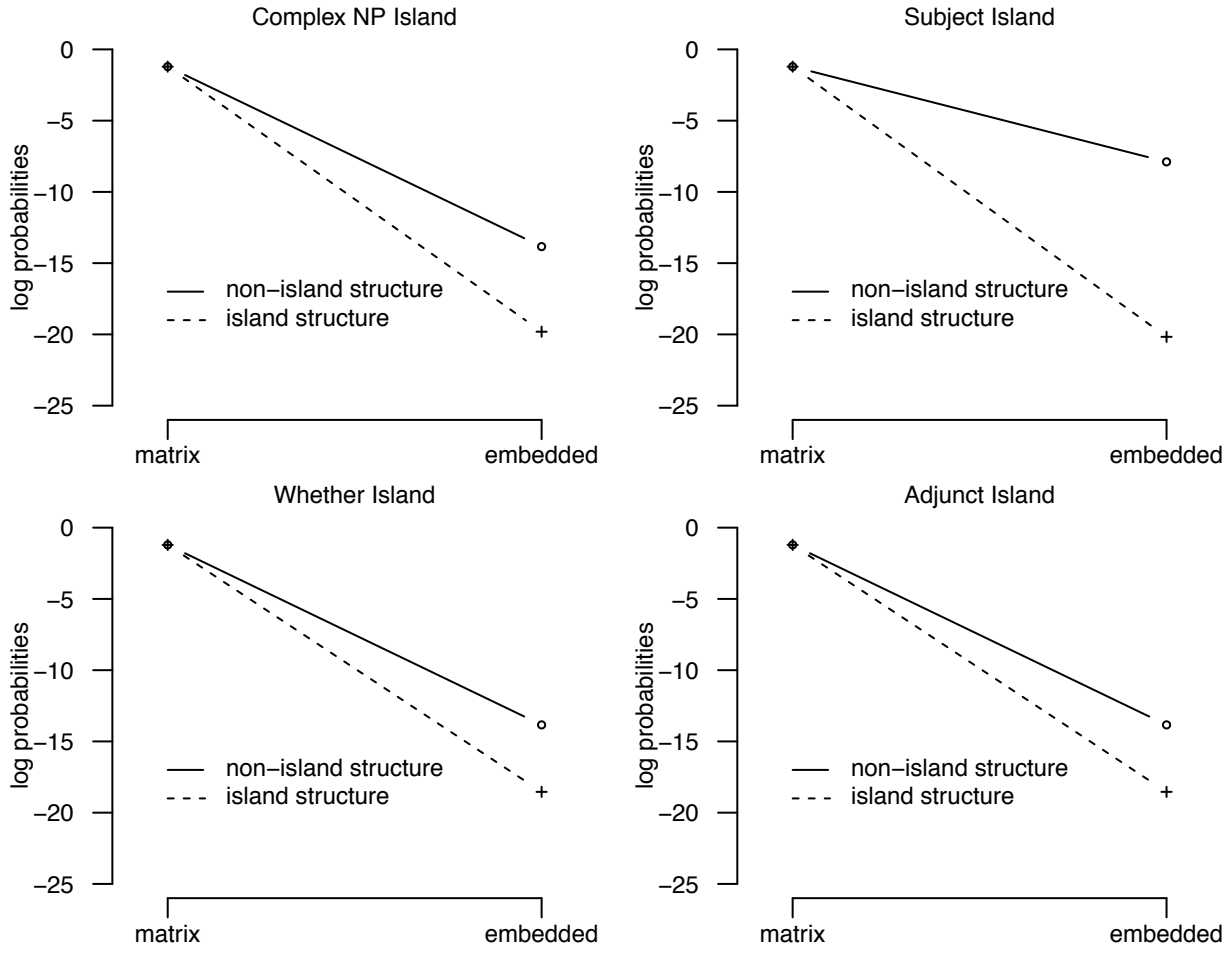


Figure 7: Log probabilities derived from adult-directed speech and text for a learner that discriminates CP types.

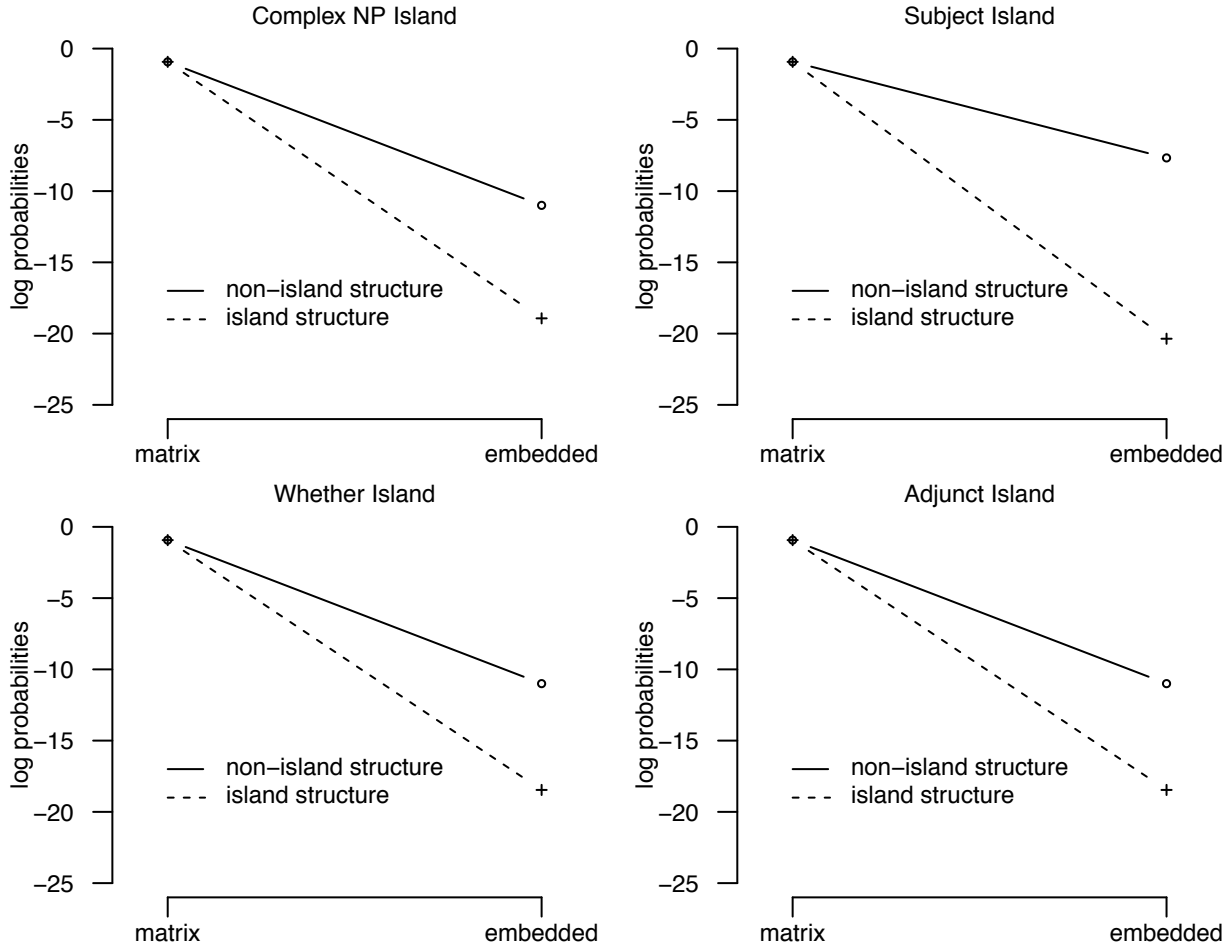


Table 7. Inferred grammaticality of different *wh*-dependencies from Sprouse et al. (2012), represented with log probability.

		Child-directed speech	Adult-directed speech & text
Control dependencies			
matrix subject	IP	-1.21	-0.93
embedded subject	IP-VP-CP_{null}-IP	-7.89	-7.67
embedded object	IP-VP-CP_{that}-IP-VP	-13.84	-11.00
Island-spanning dependencies			
Complex NP	IP-VP-NP-CP_{that}-IP-VP	-19.81	-18.93
Subject	IP-VP-CP_{null}-IP-NP-PP	-20.17	-20.36
Whether	IP-VP-$CP_{whether}$-IP-VP	-18.54	-18.46
Adjunct	IP-VP-CP_{if}-IP-VP	-18.54	-18.46

Table 8. Relative grammaticality of different *wh*-dependencies, based on the log odds of the inferred probabilities. Numbers represent the comparison of the control dependency in the row (as $prob_1$) to the island violation dependency in the column (as $prob_2$).

		Island-spanning dependencies			
		Complex NP	Subject	Whether	Adjunct
Control dependencies	Child-directed speech				
	matrix subject	18.60	18.95	17.33	17.33
	embedded subject	--	12.28	--	--
	embedded object	5.97	--	4.70	4.70
	Adult-directed speech & text				
	matrix subject	18.01	19.44	17.53	17.53
	embedded subject	--	12.69	--	--
	embedded object	7.93	--	7.46	7.46

Compared to our results from learners with undifferentiated CP container nodes, we see in Figures 6 and 7 that learners using either child-directed or adult-directed data would arrive at the correct pattern of grammaticality preferences for all four islands. Table 8 shows that all control dependencies are viewed as at least 4 times more grammatical than the island-spanning dependencies (and often more than 10 times more grammatical). In particular, the ability to distinguish CP container nodes allows the learners to have the right grammaticality preferences for the Whether and Adjunct islands, while still maintaining the right preferences for Complex NP and Subject islands. Even though complementizer *that* rarely appears in dependencies in the input (2 times in child-directed speech and 9 times in adult-directed data), it still appears more often than complementizers *whether* and *if*, which never appear. This allows the learners to view control dependencies involving complementizer *that* as more grammatical than island violation dependencies involving complementizer *whether* or complementizer *if*.

At this point it should be noted that while these results demonstrate that our modeled learner can acquire the general superadditive interaction pattern observed in the actual acceptability judgment experiments, there are still noticeable differences between the observed acceptability judgments and the inferred grammaticality preferences learned by this model. The reason for this is that actual acceptability judgments are based on dozens of factors that are not included in this model. For example, lexical items, semantic probability, and processing difficulty have all been demonstrated to impact acceptability judgments (Schütze, 1996; Cowart, 1997; Keller, 2000; Sprouse, 2009). The inferred grammaticality of this particular model would constitute only one (relatively large) factor among many that affect acceptability. Furthermore, the grammaticality preferences of this model are themselves limited to the dependency alone – they ignore all of the other syntactic properties of the sentence.

4.5.3 Why learning from container node trigrams works

We might reasonably wonder what it is about container node trigrams that leads to acquisition success. The answer becomes clear once we examine the container node trigram probabilities

involved in each island-spanning dependency, as shown in Table 9 below. Notably, for each of the island-spanning dependencies, there is at least one extremely low probability container node trigram. This occurs because these trigram sequences are never observed in the input – it is only the smoothing parameter that prevents these probabilities from being 0. Note that some trigrams are low probability due to being rarely encountered in the input (e.g., CP_{that}-IP-VP in child-directed speech) – but, crucially, this is still more than never. This is what causes the learner to view the embedded object control dependency (shown in Table 9), which is grammatical, as more probable: even though CP_{that} rarely appears, it *does* appear.

More strikingly in Table 9, we can see an explicit demonstration of how the detrimental effect of dependency length can be tempered by container node trigram frequency. Two grammatical dependencies are included for comparison, both triply embedded object dependencies. One does not involve CP nodes (IP-VP-IP-VP-IP-VP, e.g., "What does [$_{IP}$ Lily [$_{VP}$ want [$_{IP}$ to [$_{VP}$ pretend [$_{IP}$ to [$_{VP}$ steal __?]]]]]]"), while one does (IP-VP-CP_{null}-IP-VP-CP_{null}-IP-VP, e.g., "What does [$_{IP}$ Lily [$_{VP}$ think [$_{CP}$ [$_{IP}$ Jack [$_{VP}$ heard [$_{CP}$ [$_{IP}$ she [$_{VP}$ stole __]]]]]]]]?"). In both cases, these long grammatical dependencies are viewed by the learner as more probable than the island-spanning dependencies, precisely because the container node trigrams that comprise them have been seen with some frequency in the input. This shows us that container node trigrams are capturing the fact that there is some local part of the dependency that is "bad". This local "badness" spans at least two trigrams in the island-spanning dependencies we examined. Crucially, intuitions about what counts as bad are derived directly from the input, and do not correlate with dependency length.

One concern with this approach is that it might be seen to equate difficulty with ill-formedness (Phillips, 2012b).[15] In particular, one might worry that very long dependencies would start to resemble ungrammatical dependencies, even though there seems to be a qualitative difference between them. For example, a quadruply embedded object dependency involving CPs likely has a probability close to the island-spanning dependencies examined here. One way around this is to make the smoothing factor α much smaller (e.g., .00005 instead of 0.5) (Phillips, 2012b). This effectively further penalizes trigrams that have never been observed – their probability, though non-zero, is significantly smaller and thus lowers the probability of the dependency they are part of. Another way around this issue would be to back off from the notion of a combined probability for the entire dependency (Phillips, 2012b). Instead, a learner could simply note the presence of a very low probability trigram in any given dependency (this might arise naturally if that part of the dependency is difficult to process, because that container node trigram hasn't been encountered before). In this way, overall length becomes irrelevant – instead, it is simply about the presence (and perhaps quantity) of very low probability trigrams. This immediately separates the island-spanning dependencies examined here from the grammatical dependencies.

[15] We are especially grateful to Colin Phillips for his thoughts and suggestions concerning this.

Table 9. Container node trigram probabilities for each of the island-crossing dependencies after the learning period has finished, assuming finer-grained CP container nodes. Very low probability container node trigrams, which were never observed in the input, and their probabilities are in **bold**.

		Container Node Trigrams & Probabilities			
Island-spanning dependencies	Child-directed speech				
	Complex NP IP-VP-NP-CP_{that}-IP-VP *log(prob) = -19.81*	start-IP-VP 0.42 CP_{that}-IP-VP 0.000044	IP-VP-NP 0.0015 IP-VP-end 0.40	**VP-NP-CP_{that}** **0.0000012**	**NP-CP_{that}-IP** **0.0000012**
	Subject IP-VP-CP_{null}-IP-NP-PP *log(prob) = -20.17*	start-IP-VP 0.42 **IP-NP-PP** **0.0000012**	IP-VP-CP_{null} 0.0073 NP-PP-end 0.00021	VP-CP_{null}-IP 0.0073	**CP_{null}-IP-NP** **0.0000012**
	Whether IP-VP-$CP_{whether}$-IP-VP *log(prob) = -18.54*	start-IP-VP 0.42 IP-VP-end 0.40	**IP-VP-$CP_{whether}$** **0.0000012**	**VP-$CP_{whether}$-IP** **0.0000012**	**$CP_{whether}$-IP-VP** **0.0000012**
	Adjunct IP-VP-CP_{if}-IP-VP *log(prob) = -18.54*	start-IP-VP 0.42 IP-VP-end 0.40	**IP-VP-CP_{if}** **0.0000012**	**VP-CP_{if}-IP** **0.0000012**	**CP_{if}-IP-VP** **0.0000012**
	Adult-directed speech & text				
	Complex NP IP-VP-NP-CP_{that}-IP-VP *log(prob) = -18.93*	start-IP-VP 0.41 CP_{that}-IP-VP 0.000040	IP-VP-NP 0.0011 IP-VP-end 0.38	**VP-NP-CP_{that}** **0.0000013**	**NP-CP_{that}-IP** **0.0000013**
	Subject IP-VP-CP_{null}-IP-NP-PP *log(prob) = -20.36*	start-IP-VP 0.41 **IP-NP-PP** **0.0000012**	IP-VP-CP_{null} 0.0045 NP-PP-end 0.00030	VP-CP_{null}-IP 0.0045	**CP_{null}-IP-NP** **0.0000013**
	Whether IP-VP-$CP_{whether}$-IP-VP *log(prob) = -18.46*	start-IP-VP 0.41 IP-VP-end 0.38	**IP-VP-$CP_{whether}$** **0.0000013**	**VP-$CP_{whether}$-IP** **0.0000013**	**$CP_{whether}$-IP-VP** **0.0000013**
	Adjunct IP-VP-CP_{if}-IP-VP *log(prob) = -18.46*	start-IP-VP 0.41 IP-VP-end 0.38	**IP-VP-CP_{if}** **0.0000013**	**VP-CP_{if}-IP** **0.0000013**	**CP_{if}-IP-VP** **0.0000013**

Grammatical dependencies	Child-directed speech				
	Embedded object	start-IP-VP	IP-VP-CP_{that}	VP-CP_{that}-IP	CP_{that}-IP-VP
	IP-VP-CP_{that}-IP-VP	0.42	0.000044	0.000044	0.000044
	log(prob) = -13.84	IP-VP-end			
		0.40			
	Triply embedded obj	start-IP-VP	IP-VP-IP	VP-IP-VP	IP-VP-IP
	IP-VP-IP-VP-IP-VP	0.42	0.031	0.031	0.031
	log(prob) = -6.81	VP-IP-VP	IP-VP-end		
		0.031	0.40		
	Triply emb obj + CPs	start-IP-VP	IP-VP-CP_{null}	VP-CP_{null}-IP	CP_{null}-IP-VP
	IP-VP-CP_{null}-IP-VP-	0.42	0.0073	0.0073	0.0067
	CP_{null}-IP-VP	IP-VP-CP_{null}	VP-CP_{null}-IP	CP_{null}-IP-VP	IP-VP-end
	log(prob) = -13.67	0.0073	0.0073	0.0067	0.40
	Adult-directed speech & text				
	Embedded object	start-IP-VP	IP-VP-CP_{that}	VP-CP_{that}-IP	CP_{that}-IP-VP
	IP-VP-CP_{that}-IP-VP	0.41	0.000040	0.000040	0.000040
	log(prob) = -11.00				
		IP-VP-end			
		0.38			
	Triply embedded obj	start-IP-VP	IP-VP-IP	VP-IP-VP	IP-VP-IP
	IP-VP-IP-VP-IP-VP	0.41	0.017	0.017	0.017
	log(prob) = -7.89				
		VP-IP-VP	IP-VP-end		
		0.017	0.38		
	Triply emb obj + CPs	start-IP-VP	IP-VP-CP_{null}	VP-CP_{null}-IP	CP_{null}-IP-VP
	IP-VP-CP_{null}-IP-VP-	0.41	0.0045	0.0045	0.0020
	CP_{null}-IP-VP	IP-VP-CP_{null}	VP-CP_{null}-IP	CP_{null}-IP-VP	IP-VP-end
	log(prob) = -15.59	0.0045	0.0045	0.0020	0.38

5. General Discussion

In this study, we investigated an acquisition problem previously believed to strongly implicate UG: learning that dependencies cannot span certain syntactic structures known as syntactic islands. UG has been one solution offered to solve induction problems in language acquisition, so we first verified that learning about syntactic islands appears to present an induction problem, particularly if the child has a narrow view of what evidence is relevant. We then demonstrated that a simple statistical learning model that takes a broader view of relevant data[16] is able to reach the target knowledge state, where dependencies spanning syntactic islands are perceived as ungrammatical. The statistical learning model itself included one derived, domain-specific learning bias, three innate, domain-general learning biases, and one domain-specific learning bias that may or may not need to be innate. This result raises the possibility that syntactic island

[16] As mentioned above, this is in a similar spirit to recent computational models by several researchers (Foraker et al., 2009; Pearl & Mis, 2011; submitted; Perfors et al., 2011; and Regier & Gahl, 2004) as well as the idea behind linguistic parameters (Chomsky, 1981; Pearl & Lidz, in press; Viau & Lidz, 2011).

effects could in principle be learned without the UG hypothesis, assuming that the aforementioned bias does not need to be innate. In this section we discuss this bias in more detail, as well as some interesting questions about the role of sophisticated linguistic knowledge in the learning process (and relatedly, how that linguistic knowledge is learned) that are raised by this model. We also discuss how feasible this learner would be for the full range of constraints on *wh*-dependencies, both across constructions in a single language and across languages, and what ramifications this learner would have for syntactic theories.

5.1. Is tracking trigram sequences of container nodes (and container nodes in general) an example of UG?

As discussed in section 4.2, it is a fairly common assumption in the learning literature that children can track trigrams (of various types). We also take it to be uncontroversial that children must be able to identify the container nodes for a *wh*-dependency, as this must be part of the parsing process for (actively) identifying gap locations. At the very least this means that the information about the frequency of the trigram sequences of container nodes is in principle available to the language learner. In other words, the availability of the information itself is likely derived, domain-specific information. That being said, the availability of the information is logically distinct from the fact that the learner must actually attend to this particular information, to the exclusion of all other possible types of information that could be attended to (intermediate projections, intervening words, etc).[17] It is possible that the fact that the learner must attend to this particular information is itself an innate, domain-specific bias. We have no empirical evidence that can bear on this question. Therefore, instead of trying to come down one way or the other for this particular bias, we will discuss the ramifications of each possibility.

If this bias is indeed part of UG, then it at least suggests a recalibration of the content of UG, at least for island constraints (Chomsky, 1973; Huang 1982; Lasnik & Saito, 1984). Prior to this model, the most common UG approach to island constraints was to assume that the two components of the Subjacency condition were innately available to the learner: (i) a principle that precluded movement that crosses two bounding nodes, and (ii) a specification of the possible bounding nodes. The current model suggests that UG can instead be viewed as a set of less syntactically detailed biases that simply specify which information to attend to in the input. In other words, there is still an innate and domain-specific bias (attend to container nodes), but the syntactic theory itself (e.g., (i) above) does not need to be innately specified. The constraint previously accounted for by the theory instead falls out from the probabilities of the container node trigrams.

The other possibility is that this particular learning bias is not UG-based at all. It seems relatively unlikely that this bias is domain-general, as we know of no examples of island-like constraints on dependencies in other domains of cognition. Therefore the only possibility is that it is a derived, domain-specific bias. Exactly how this bias could be derived is currently unclear. As discussed in section 4.2, one idea is that there may be some default bias to use basic-level phrasal constituents derived from parsing, since those are already being tracked and the learner may have a preference to use already existing representations. If container nodes can be derived, then this would obviate the need for the attentional bias as part of UG.

[17] We are grateful to Bob Frank for bringing this to our attention.

However, there still remains the sophisticated bias of tracking trigrams of container nodes. The fact that the component biases are present (the ability to track trigrams, the ability to track container nodes) opens up the possibility that there is an innate mechanism for combining simpler biases into more complex biases. However, we currently have no theory of how such a bias combinatorics might work (i.e., a "grammar" of biases), let alone how to prevent it from creating combined biases that are unnecessary (or even deleterious to the learning process). As more computational models of syntactic acquisition are proposed, it may become more feasible to explore this possibility.

5.2. Is subcategorizing CPs an example of UG?

As demonstrated in section 4.5, the acquisition of Whether and Adjunct islands requires the learner to distinguish between different types of CPs when tracking the frequency of trigrams of container nodes. Once again, this is a relatively sophisticated learning bias that must be built from two independently motivated (and less sophisticated) learning biases. For example, it is uncontroversial to assume that children learn to distinguish different types of CPs: The lexical content of CPs has substantial consequences for the semantics of a sentence (e.g., declaratives versus interrogatives), and even within declarative sentences, it has been shown that speakers are sensitive to the distribution of *that* versus null complementizers (Jaeger, 2010). This is likely a derived, domain-specific learning bias. However, our model requires combining this uncontroversial assumption with our novel bias to track container node trigrams, such that different CPs lead to different trigram sequences. Once again, the result is a relatively sophisticated learning bias that superficially resembles an innate, domain-specific bias, but is likely built upon a series of independent (and potentially non-UG) biases.

5.3. Low probability trigrams that are relatively acceptable: Parasitic gaps

Though this statistical learning model demonstrates that syntactic islands can in principle be learned from child-directed input, this particular model cannot capture certain known exceptions to syntactic island constraints, such as *parasitic gap* constructions (Engdahl, 1983). Parasitic gap constructions are *wh*-questions in which the *wh*-word is associated with two gap positions: One gap position occurs in a licit gap location (i.e., not inside a syntactic island) while the other gap position occurs inside a syntactic island. Whereas a single gap within an island structure results in unacceptability (20a and 21a), the addition of another gap outside of the island seems to eliminate the unacceptability (20b and 21b) (see Phillips (2006) for experimentally collected acceptability judgments):

(20) a. *Which book did you laugh [before reading __]?
b. Which book did you judge $\_\__{true}$ [before reading $\_\__{parasitic}$]?

(21) a. *What did [the attempt to repair __] ultimately damage the car?
b. What did [the attempt to repair $\_\__{parasitic}$] ultimately damage $\_\__{true}$?

The two gaps in a parasitic gap construction are often described as the *true gap*, which occurs outside of the island, and the *parasitic gap*, which occurs inside of the island. The name is a metaphorical reference to the fact that the *parasitic gap* could not exist without the *true gap*,

much like a parasite cannot exist without a host. Though there are several structural restrictions on parasitic gap constructions (e.g., the true gap cannot c-command the parasitic gap), there is no constraint on the linear order of the two gaps, as illustrated by (20-21).

The acceptability (and presumably the grammaticality) of parasitic gap constructions poses a problem for our statistical learner. This is because the probability of the trigram sequence for the dependency between the *wh*-word and the parasitic gap will be the same as the probability of the trigram sequence for the structurally equivalent syntactic island violation. In other words, our learner would infer that parasitic gap constructions are ungrammatical because there is a low-probability trigram in the dependency. For example, the container node sequences for (20) would be as in (22). The sequence for both the ungrammatical gap in (20a) and the grammatical (parasitic) gap in (20b) are identical, and in fact would be as (un)acceptable as other adjunct islands, such as those using the complementizer *if*.

(22)

a. *Which book did [$_{IP}$ you [$_{VP}$ laugh [$_{CP}$ without [$_{IP}$ [$_{VP}$ reading __]]]]]?
Ungrammatical gap sequence: IP-VP-CP$_{without}$-IP-VP

b. Which book did [$_{IP}$ you [$_{VP}$ judge __$_{true}$ [$_{CP}$ without [$_{IP}$ [$_{VP}$ reading __$_{parasitic}$]]]]]]?
Parasitic gap sequence: IP-VP-CP$_{without}$-IP-VP

Given that this is not the desired target state, the learning algorithm proposed here is unlikely to be the one children use in practice. However, it may be possible to modify the learning model to account for these constructions. For example, recent studies demonstrate that the human parser continues to actively search for a second gap even after encountering a licit first gap (Wagers & Phillips, 2009). It could be that the learning algorithm assembles a grammaticality preference based on some kind of aggregation of all container node sequences for gaps in a given utterance. However, unless there is an innate, domain-specific bias to aggregate gap information (which would then make this a UG bias), this would need to be derived from linguistic experience somehow. One way is for children to have experience with multiple gaps associated with the same *wh*-element. In order for this to be true, child-directed input (or adult-directed, if acquisition is relatively late) must contain examples of *wh*-elements associated with multiple gaps, such as examples of parasitic gaps and across-the-board extractions (e.g., What did you read __ and then review __?). We are currently examining additional syntactically-annotated child-directed corpora to answer this and other related questions.

5.4. High probability trigrams that are relatively unacceptable: Cross-linguistic variation

Just as the current model would be forced to treat parasitic gaps as ungrammatical because any dependency that contains a very low-frequency trigram is ungrammatical, the model would similarly treat all dependencies that contain only higher-frequency trigrams as grammatical. This is not problematic in English, as all such dependencies are in fact grammatical. However, Rizzi (1982) reports an interesting paradigm in Italian in which it looks as though simply doubling a grammatical sequence of trigrams leads to ungrammaticality (Phillips, 2012b). Rizzi (1982) reports that Italian does not have *wh*-island effects the way that English does, as an extraction of an NP from a *wh*-island structure is grammatical ((23) = Rizzi's (6a)):

(23) Tuo fratello, a cui mi domando che storie abbiano raccontato, era molto preoccupato.
your brother, to whom$_1$ I wonder which stories$_2$ they have told __$_2$ __$_1$, was very worried.

…to whom$_1$ [$_{IP}$ I [$_{VP}$ wonder [$_{CP}$ which stories$_2$ [$_{IP}$ they [$_{VP}$ have told __$_2$ __$_1$]]]]]

Dependency for *to whom*: IP-VP-CP$_{wh}$-IP-VP

Rizzi analyzes this fact as evidence that the (Subjacency-based) bounding nodes in Italian are NP and CP, which correctly captures the fact that extraction from a CP is possible even when the specifier of CP is filled with a wh-phrase. This analysis makes an interesting prediction: If CP is a bounding node, extraction should not be able to cross two CPs with filled specifier positions. Rizzi reports that this prediction appears is borne out ((24)=Rizzi's (15b)):

(24) *Questo argomento, di cui mi sto domandando a chi potrei chiedere quando dovrò parlare, mi sembra sempre più complicato.
*this topic, of which$_1$ I am wondering to whom$_2$ I may ask __$_2$ when$_3$ I'll have to speak __$_1$ __$_3$, to me seems ever more complicated

…of which$_1$ [$_{IP}$ I [$_{VP}$ am wondering [$_{CP}$ to whom$_2$ [$_{IP}$ I [$_{VP}$ may ask __$_2$ [$_{CP}$ when$_3$ [$_{IP}$ I [$_{VP}$ 'll have [$_{IP}$ to [$_{VP}$ speak __$_1$ __$_3$]]]]]]]

Dependency for *of which*: IP-VP-CP$_{wh}$-IP-VP-CP$_{wh}$-IP-VP-IP-VP

The problem for our learner is that the container node sequence of the ungrammatical sentence in (24) (CP$_{wh}$-IP-VP-CP$_{wh}$-IP-VP) consists of the very same trigrams that are in the grammatical sentence in (23) (CP$_{wh}$-IP-VP, IP-VP-CP$_{wh}$, and VP-CP$_{wh}$-IP). Therefore our learner will treat it as grammatical. Whether sentences such as (24) are unacceptable or not is an empirical question; nonetheless, the example serves to illustrate one of the primary limitations of the current model: The grammaticality of each sentence is predicated solely upon the frequency of the individual "parts", where the parts are trigrams of container nodes. If any one trigram is low-frequency, as in parasitic gaps, the model will treat the sentence as ungrammatical; if all of the trigrams are higher-frequency, as in example (24), the model will treat the sentence as grammatical.

5.5. The problem of complementizer *that*

Another issue with this particular model concerns complementizer *that* – specifically, because of the rarity of complementizer *that* in the input data, a learner using this model will generally disprefer dependencies using complementizer *that* (Phillips, 2012b). In some cases, this may be desirable, such as *that*-trace effects, which occur when the gap immediately follows *that* (25a), but do not arise when *that* is omitted (25b) (see Cowart (1997) for experimentally collected acceptability judgments). The current model can capture the distinction between these, shown in (25), using either child-directed or adult-directed data (child-directed log-odds: 7.12, adult-directed log odds: 5.40).

(25)

a. *Who do [$_{IP}$ you [$_{VP}$ think [$_{CP}$ that [$_{IP}$ __ [$_{VP}$ read the book]]]]]?
b. Who do [$_{IP}$ you [$_{VP}$ think [$_{CP}$ [$_{IP}$ __ [$_{VP}$ read the book]]]]]?

Unfortunately, the current model will also generate a preference for object gaps when *that* is omitted (26b) compared to object gaps when *that* is present (26a):

(26)

a. What do [$_{IP}$ you [$_{VP}$ think [$_{CP}$ that [$_{IP}$ Jack [$_{VP}$ read __]]]]]?
b. What do [$_{IP}$ you [$_{VP}$ think [$_{CP}$ [$_{IP}$ Jack [$_{VP}$ read __]]]]]?

Interestingly, Cowart (1997) reports that there is a small preference in adult acceptability judgments for (26b) over (26a), but it is significantly smaller than the preference for (25b) over (25a). In other words, there is an object *that*-trace effect, but it is much smaller. The current model generates relatively equal dispreference for (25a) and (26a) when using the child-directed corpora, which contain relatively few instances of *that*: The log-odds of (25a) versus (25b) is 7.12, and the log-odds of (26a) versus (26b) is 6.61. However, the model generates an asymmetrical dispreference that is more in line with Cowart's (1997) data when using the adult-directed corpora, which contain more instances of *that*: The log-odds for (25) are 5.40, and the log-odds for (26) are 2.81. This could be taken to be a developmental prediction of the current model: Children will disprefer object gaps in embedded *that*-CP clauses more than adults, and the dispreference will weaken as they are exposed to additional tokens of *that*.

5.6 The implications of this model for syntactic theory

In some ways, the (implicit) output of the current learning model looks very similar to existing theories of syntactic islands: Island effects arise due to constraints on sequences of abstract units derivable from phrase structure trees. In our case, these units are container nodes; for the syntactic theory of Subjacency, these units are *bounding nodes* or *barriers* (Chomsky, 1973; Chomsky, 1986). This is to be expected given that the syntactic analysis of long-studied phenomena such as syntactic islands have substantial empirical support (e.g., Chomsky, 1973; 1986; Huang, 1982; Lasnik & Saito, 1984; Rizzi, 1980; Ross, 1967; Torrego, 1984; among many others). It is simply a case of describing a formal learning model that can yield the correct analysis based on child-directed input. However, there are some real differences between our current model and existing syntactic theories; to put it another way, our current model makes predictions about the representation of these constraints that differ from the representations assumed in current syntactic theories.

The first potential difference between our model and syntactic theory concerns the relatedness of A'-constructions. Syntactic theory treats *wh*-questions, relativization, topicalization, comparatives, and adjective-though constructions as a natural class (e.g., all derived by *movement*) because they all show island effects. As it currently stands, our model is agnostic about this relationship (Phillips, 2012b). It would be simple enough to code a relationship between all of these constructions, but that simply begs the question of how it is that the learner is biased to treat these constructions as equivalent. Like all learning biases, this one needs to be investigated to determine whether it is a UG-bias or not.

Another potential difference concerns the tendency in syntactic theories to capture several island types with a single constraint. For example, under the Subjacencey condition, both WH islands and Subject islands arise because of the choice of IP as a bounding node (as in English). They therefore disappear when IP is not a bounding node (as in Italian). Similarly, under the Condition on Extraction Domains (Huang, 1982), both Subject and Adjunct islands arise because subjects and adjuncts are ungoverned constituents. Our model can capture relationships between island types, but does so by treating each island as an amalgam of trigrams. In this way, islands violations are related if they share (low-probability) trigrams. As Table 9 indicates, Whether and Adjunct islands share three nearly identical low-probability trigrams (modulo the head of the CP), so this model may predict some relationship between them (depending on the role of the head of CP). Complex NP and Subject islands do not share any low-probability trigrams with each other, or with Whether and Adjunct islands, perhaps predicting that both will be independent island types from the others.

Finally, it should be noted that syntactic theories of island constraints have been constructed to account for cross-linguistic variation such as the potential Italian facts discussed in section 5.4 (see Sprouse and Hornstein (2012), and Phillips (2012a), for a brief overview of the cross-linguistic variation in islands). The current model makes the strong prediction that the pattern of island effects in any given language should be completely predictable based on the probability of the trigrams of container nodes in the child-directed input. This is an empirical question, albeit one that is difficult to test without access to structurally annotated corpora in the relevant languages.

5.7. The implications of these results for the theory of acquisition

First and foremost, the success of the current model suggests that syntactic island effects – a set of phenomena that are central to (UG-based) syntactic theories – may not in principle require UG to be learned, depending on how one views the bias to track container nodes. Scaling this learning mechanism up to handle the full range of cross-constructional and cross-linguistic facts is an interesting topic for future research, as is the question of how to integrate this mechanism with a larger theory of syntactic acquisition. As difficult as those tasks may seem, they are both more tractable now that we have a first working model, and the structurally-annotated child-directed corpora with which to test future models.

It is also interesting to note that we were able to successfully model the acquisition of a complex linguistic phenomenon (syntactic island constraints) without sophisticated probabilistic inference mechanisms, such as Bayesian inference (e.g., Feldman et al., 2009; Foraker et al., 2009; Frank et al., 2009; Goldwater et al., 2009; Pearl & Lidz, 2009; Pearl et al., 2011; Perfors et al., 2011; Regier & Gahl, 2004).[18] Instead, a fairly simple probabilistic learning component (tracking frequencies of particular linguistic representations) was sufficient to learn the pattern from child-directed input. Given the relative complexity of syntactic islands with respect to other phenomena in linguistic theory, this suggests that there may be other (complex) linguistic phenomena that can be modeled with similarly simple probabilistic mechanisms. This may

[18] Of course, our model assumes that the phrase structure has already been inferred, and learning phrase structure may require sophisticated probabilistic inference methods. However, once the phrase structure is available, no sophisticated inference is required to learn syntactic island constraints, which is the learning process explicitly modeled here.

eliminate some of the concerns that have been raised about the psychological plausibility of Bayesian inference as a realistic learning mechanism for humans (e.g., see McClelland, Botvinick, Noelle, Plaut, Rogers, Seidenberg, & Smith, 2010 for a recent review).

Finally, it is also interesting to note that at least for the *wh*-dependency constructions and level of syntactic abstraction studied here, the distributional differences between child-directed speech and adult-directed speech appear to be fairly minimal. This is an important methodological point for researchers of syntactic acquisition, as it's often the case that large samples of syntactically annotated adult-directed speech data are more easily accessible and readily available than syntactically annotated child-directed speech data. At the level of syntactic dependencies, it appears that adult-directed speech data could serve as a reasonable proxy for child-directed speech data. It may be the case that this is also true of other abstract syntactic structural relationships, though future research is clearly necessary.

5.8. Deriving developmental predictions from computational models

As discussed briefly in section 4.2, the computational learning model proposed here is technically agnostic about the time-course of the implementation of the learning biases necessary to successfully acquire syntactic island constraints (i.e., our model simply assumes that all of the learning biases are present). However, it should still be noted that one of the more interesting consequences of learning models that combine several distinct learning biases is that it is logically possible that the learning biases are implemented at different times, resulting in specific learning trajectories. For example, it is logically possible that the bias to use subcategorized CP container nodes only arises after acquisition of syntactic islands has failed using basic level CP container nodes. If children initially treat all CP container nodes as identical, then there will be a period early in the acquisition of syntactic islands during which children will perceive dependencies spanning Complex NP and Subject island structures as ungrammatical, while simultaneously perceiving dependencies spanning Whether and Adjunct island structures as grammatical (closely mirroring the results of the learning model in section 4.5.1). At a later point in the acquisition process children would then "expand" to the more detailed container node representation, and learn Whether and Adjunct island constraints. Of course, it is also possible that the subcategorized CP bias is in place early enough that such a stage never occurs; the point here is not that this is a unique prediction of our model, but rather that models that rely on the interaction of several different learning biases can be used to map out the hypothesis space for the time-course of syntactic acquisition (for experiments investigating the time course of syntactic island acquisition, see De Villiers & Roeper , 1995; De Villiers, Roeper, Bland-Stewart, & Pearson, 2008; and Goodluck, Foley, & Sedivy, 1992; and see Roeper & de Villiers, 2011 for a recent review of the *wh*-question acquisition literature).

It should also be noted that learning models that are based on the distribution of the input will also be sensitive to temporal changes in that distribution. We saw a hint of such a change in section 5.5, as the dispreference for object gaps in embedded clauses headed by *that* was weaker when learned from adult-directed speech, because the adult-directed speech corpora contains more tokens of complementizer *that*. Now that structurally-annotated child-directed corpora are available, it will be possible to look for small changes in the distribution of relevant structures in child-directed and adult-directed speech that may be responsible for some aspects of the time-course of syntactic acquisition.

6. Conclusion

By examining a particular acquisition problem that is considered a strong motivation for UG, we have been able to concretely determine that it may not, in fact, require UG to solve after all (though UG-like learning biases are certainly one solution to the problem). After first verifying that there was an induction problem for children, we used a simple statistical learner sensitive to abstract syntactic representations to demonstrate how knowledge of syntactic island constraints can be implicitly derived from the frequencies of those representations in both child-directed and adult-directed input. Crucially, this learning model considered indirect positive evidence and so expanded the set of data considered relevant, thus alleviating the apparent induction problem. The results of this learner suggest that the complex learning biases necessary to acquire complex syntactic phenomena may in principle be derived from the interaction of independently motivated (non-UG) biases, thus reducing the motivation for the UG hypothesis in these cases. Moreover, these phenomena can be learned without the need for complex probabilistic inferential mechanisms such as Bayesian inference. Beyond that, these results also reaffirm the empirically supported analyses that characterize syntactic theory. Because this learning model requires a combination of distinct learning biases, it can also be used to explore the hypothesis space of potential time-courses of syntactic island acquisition. We believe that these results highlight how explicit computational modeling studies of acquisition can contribute to our understanding of language abilities and knowledge in the human mind.

Acknowledgements

We would like to thank Colin Phillips, Norbert Hornstein, Bob Berwick, Bob Frank, Virginia Valian, Alexander Clark, Misha Becker, Anne Hsu, Kamil Ud Deen, Jeff Lidz, Charles Yang, Julian Pine, Terry Regier, William Sakas, Amy Perfors, two anonymous reviewers, the attendees of the Input & Syntactic Acquisition workshop held at the LSA in 2012 and at UC Irvine in 2009, and the audience at the Ecole Normale Supérieure in 2011 for numerous comments and suggestions on previous versions of this work. All errors remain our own. We would also like to thank Tom Roeper for discussion of the experimental evidence in children for syntactic island knowledge. In addition, we are very grateful to Jessica Lee, Uma Patel, Christine Thrasher, and other members of the Computation of Language Laboratory who aided in the syntactic annotation of the child-directed speech. This work was supported in part by NSF grant BCS-0843896.

References

Abrusan, M. (2011). Presuppositional and Negative Islands: A Semantic Account. *Natural Language Semantics, 19(3)*, 257–321.

Aslin, R., Saffran, J., & Newport, E. (1998). Computation of Conditional Probability Statistics by 8-Month-Old Infants. *Psychological Science, 9(4)*, 321-324.

Baker, C. (1978). *Introduction to generative transformational syntax.* Englewood Cliffs, NJ: Prentice Hall.

Baker, C. (1981). *The Logical Problem of Language Acquisition.* Cambridge: MIT Press.

Bates, E., Bretherton, I., & Snyder, L. (1988). *From first words to grammar: Individual differences and dissociable mechanisms*. Cambridge, MA: Cambridge University Press.

Berwick, R., Pietroski, P., Yankama, B., & Chomsky, N. (2011). Poverty of the Stimulus Revisited. *Cognitive Science, 35,* 1207-1242.

Boeckx, C. & Grohmann, K. K. (2007). Remark: Putting Phases in Perspective. *Syntax, 10,* 204–222.

Braunwald, S. (1978). Context, word and meaning: Toward a communicational analysis of lexical acquisition. In A. Lock (Ed.), *Action, gesture and symbol: The emergence of language*, 485-527. London: Academic Press.

Brown, R. (1973). *A first language: The early stages*. Cambridge, MA: Harvard University Press.

Chomsky, N. (1965). *Aspects of the theory of syntax*. Cambridge: The MIT Press.

Chomsky, N. (1973). Conditions on transformations. In S. Anderson & P. Kiparsky (Eds.), *A Festschrift for Morris Halle*, (pp. 237-286). New York: Holt, Rinehart and Winston.

Chomsky, N. (1980). *Rules and representations*. Oxford: Basil Blackwell.

Chomsky, N. (1981). *Lectures on Government and Binding.* Dordrecht: Foris.

Chomsky, N. (1986). *Barriers*. Cambridge: The MIT Press.

Chomsky, N. (1988). *Language and problems of knowledge: The managua lectures.* Cambridge, MA: MIT Press.

Chomsky, N. (2001). Derivation by phase. In M. Kenstowicz (Ed.), *Ken Hale: A life in language*, (pp. 1-52). Cambridge, MA: MIT Press.

Cowart, W. (1997). *Experimental syntax: Applying objective methods to sentence judgments*. Thousand Oaks, CA: Sage.

Crain, S. (1991). Language acquisition in the absence of experience. *Behavioral and Brain Sciences, 14,* 597–612.

Crain, S., & J. Fodor. (1985). How can grammars help parsers? In D. Dowty, L. Karttunen, & A. Zwicky (Eds.), *Natural language parsing: psycholinguistic, computational, and theoretical approaches,* (pp. 94–128). Cambridge University Press.

Crain, S., & Pietroski, P. (2002). Why language acquisition is a snap. *The Linguistic Review, 19*, 163–183.

de Villiers, J. G. & T. Roeper. (1995). Relative clauses are barriers to Wh-movement for young children. *Journal of Child Language, 22*, 389-404.

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

Deane, P. (1991). Limits to attention: a cognitive theory of island phenomena. *Cognitive Linguistics, 2*, 1-63.

Denison, S., Reed, C., & Xu, F. (2011). The emergence of probabilistic reasoning in very young infants. *Proceedings of the 33rd Annual Meeting of the Cognitive Science Society,* Boston, MA.

Dewar, K. & Xu, F. (2010). Induction, Overhypothesis, and the Origin of Abstract Knowledge: Evidence from 9-Month-Old Infants, *Psychological Science, 21(12),* 1871-1877.

Dresher, E. (1999). Charting the learning path: Cues to parameter setting. *Linguistic Inquiry*, *30*, 27-67.

Dresher, E. (2003). Meno's paradox and the acquisition of grammar. In S. Ploch (Ed.), *Living on the edge: 28 papers in honour of Jonathan Kaye (Studies in Generative Grammar 62),* (pp. 7–27). Berlin: Mouton de Gruyter.

Engdahl, E. (1980). Wh-constructions in Swedish and the relevance of subjacency. In J. T. Jensen (Ed.), *Cahiers Linguisticques D'Ottawa: Proceedings of the Tenth Meeting of the North East Linguistic Society*, (pp. 89-108). Ottawa, ONT: University of Ottawa Department of Linguistics.

Engdahl, E. (1983) Parasitic Gaps. *Linguistic Inquiry, 6(1),* 5–34.

Erteschik-Shir, N. (1973). *On the nature of island constraints.* Cambridge, MA: MIT dissertation.

Feldman, N., Griffiths, T., & Morgan, J. (2009). The influence of categories on perception: Explaining the perceptual magnet effect as optimal statistical inference. *Psychological Review, 116,* 752-782.

Fodor, J. D. (1998a). Unambiguous Triggers. *Linguistic Inquiry, 29*, 1-36.

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

Fodor, J. D. (2009). Syntax Acquisition: An Evaluation Measure After All? In M. Piatelli Palmarini, J. Uriagereka, & P. Salaburu. (Eds.), *Of Minds and Language: The Basque Country Encounter with Noam Chomsky*, Oxford University Press.

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

Frank, M.C., Goodman, S., & Tenenbaum, J. (2009). Using Speakers' Referential Intentions to Model Early Cross-Situational Word Learning. *Psychological Science, 20(5),* 578-585.

Frazier, L. & Flores d'Arcais, G. (1989). Filler driven parsing: A study of gap filling in Dutch. *Journal of Memory and Language, 28*, 331–344.

Gerken, L. (2006). Decision, decisions: infant language learning when multiple generalizations are possible. *Cognition*, *98*, B67-B74.

Gibson, E. & Wexler, K. (1994). Triggers, *Linguistic Inquiry, 25,* 355-407.

Goldberg, A. (2007). *Constructions at work.* Oxford: Oxford University Press.

Goldwater, S., T. Griffiths, & M. Johnson. (2009). A Bayesian Framework for Word Segmentation: Exploring the Effects of Context. *Cognition, 112(1)*, 21-54.

Goodluck, H., Foley, M., & Sedivy, J. (1992). Adjunct islands and acquisition. In H. Goodluck (Ed.), *Islands constraints,* (pp. 181-194). Dordrecht: Kluwer.

Graf Estes, K., Evans, J., Alibali, M., & Saffran, J. (2007). Can Infants Map Meaning to Newly Segmented Words? *Psychological Science, 18(3)*, 254-260.

Griffiths, T. & Tenenbaum, J. (2005). Structure and strength in causal induction. *Cognitive Psychology, 51,* 334-384.

Hagstrom, P. (1998). Decomposing Questions. Doctoral dissertation. MIT, Cambridge, MA.

Hale, J. (2001). A probabilistic Earley parser as a psycholinguistic model. In *Proceedings of the Second Meeting of the North American Chapter of the Association for Computational Linguistics*, 159–166.

Hart, B., & Risley, T. R. (1995). *Meaningful differences in the everyday experience of young American children.* Baltimore, MD: P.H. Brookes.

Hornstein, N., & Lightfoot, D. (1981). Introduction. In N. Hornstein (Ed.), *Explanation in linguistics: The logical problem of language acquisitions* (pp. 9–31). London: Longman.

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

Jaeger, T.F & Snider, N. (2008). Implicit learning and syntactic persistence: Surprisal and Cumulativity. *Proceedings of the 30th Annual Meeting of the Cognitive Science Society*, 1061-1066.

Jaeger, T. F. (2010). Redundancy and Reduction: Speakers Manage Syntactic Information Density. *Cognitive Psychology, 61(1),* 23-62.

Klein, D. & Manning, C. (2002). A Generative Constituent-Context Model for Improved Grammar Induction. *Proceedings of the 40th Annual Meeting for the Association for Computational Linguistics.* Association for Computational Linguistics: Stroudsburg, PA, 128-135.

Kluender, R., & Kutas, M. (1993). Subjacency as a processing phenomenon. *Language and Cognitive Processes, 8*, 573-633.

Lasnik, H. & Saito, M. (1984). On the nature of proper government. *Linguistic Inquiry, 15,* 235-289.

Legate, J. & Yang, C. (2007). Morphosyntactic learning and the development of tense. *Language Acquisition, 14(3),* 315-344.

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, *106,* 1126–1177.

Levy, R. (2011). Integrating surprisal and uncertain-input models in online sentence comprehension: formal techniques and empirical results. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*.

Lidz, J., Waxman, S., & Freedman, J. (2003). What infants know about syntax but couldn't have learned: experimental evidence for syntactic structure at 18 months. *Cognition*, *89*, B65–B73.

Lightfoot, D. (1989). The child's trigger experience: degree-0 learnability. *Behavioral and Brain Sciences, 12,* 321–334.

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

Lightfoot, D. (2010). Language acquisition and language change. *Wiley Interdisciplinary Reviews: Cognitive Science, 1*, 677-684. doi: 10.1002/wcs.39.

MacWhinney, B. (2000). *The CHILDES project: Tools for analyzing talk*. Lawrence Erlbaum Associates, third edition.

MacWhinney, B. (2004). "A multiple process solution to the logical problem of language acquisition". *Journal of Child Language, 31,* 883–914.

Manning, C. & Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. Cambridge, MA: MIT Press.

Marcus, M., Santorini, B., Marcinkiewicz, M., & Taylor, A. (1999). *Treebank-3.* Linguistic Data Consortium, Philadelphia.

McClelland, J. L., Botvinick, M. M., Noelle, D. C., Plaut, D. C., Rogers, T. T., Seidenberg, M. S., & Smith, L. B. (2010). Letting structure emerge: Connectionist and dynamical systems approaches to understanding cognition. *Trends in Cognitive Sciences*, *14*, 348-356.

McKinnon, R. & Osterhout, L. (1996). Event-related potentials and sentence processing: Evidence for the status of constraints on movement phenomena. *Language and Cognitive Processes, 11(5)*, 495-523.

McMurray, B. & Hollich, G. (2009). Core computational principles of language acquisition: can statistical learning do the job? Introduction to Special Section. *Developmental Science, 12(3),* 365-368.

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

Mintz, T. (2006). Finding the verbs: distributional cues to categories available to young learners. In K. Hirsh-Pasek & R.M. Golinkoff (Eds.), *Action Meets Word: How Children Learn Verbs,* (pp. 31-63). New York: Oxford University Press.

Mitchener, W. & Becker, M. (2011). Computational Models of Learning the Raising-Control Distinction. *Research on Language and Computation, 8(2),* 169-207.

Nishigauchi, T. (1990). *Quantification in the Theory of Grammar.* Dordrecht: Kluwer.

Niyogi, P. & Berwick, R. (1996). A language learning model for finite parameter spaces. *Cognition, 61,* 161-193.

Pearl, L. (2008). Putting the Emphasis on Unambiguous: The Feasibility of Data Filtering for Learning English Metrical Phonology. In H. Chan, H. Jacob, & E. Kapia (Eds.) *BUCLD 32: Proceedings of the 32nd Annual Boston University Conference on Child Language Development,* (pp.390-401), Somerville: MA: Cascadilla Press.

Pearl, L. (2011). When unbiased probabilistic learning is not enough: Acquiring a parametric system of metrical phonology. *Language Acquisition, 18(2)*, 87-120.

Pearl, L., Goldwater, S., & Steyvers, M. (2011). Online Learning Mechanisms for Bayesian Models of Word Segmentation, *Research on Language and Computation,* special issue on computational models of language acquisition. DOI 10.1007/s11168-011-9074-5.

Pearl, L., & Lidz, J. (2009). When domain-general learning fails and when it succeeds: Identifying the contribution of domain-specificity. *Language Learning and Development, 5(4),* 235–265.

Pearl, L., & Lidz, J. (in press). Parameters in Language Acquisition. In K. Grohmann & C. Boeckx (Eds.), *The Cambridge Handbook of Biolinguistics.* Cambridge: Cambridge University Press.

Pearl, L. & Mis, B. (2011). How Far Can Indirect Evidence Take Us? Anaphoric One Revisited. *Proceedings of the 33rd Annual Conference of the Cognitive Science Society.* Boston, MA: Cognitive Science Society.

Pearl, L. & Mis, B. (submitted). What Indirect Evidence Can Tell Us About Universal Grammar: Anaphoric One Revisited. Ms., University of California, Irvine.

Pearl, L. & Sprouse, J. (forthcoming) Computational Models of Acquisition for Islands, In J. Sprouse & N. Hornstein (Eds), *Experimental Syntax and Islands Effects*. Cambridge University Press.

Pearl, L. & Weinberg, A. (2007). Input Filtering in Syntactic Acquisition: Answers from Language Change Modeling, *Language Learning and Development, 3(1),* 43-72.

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

Pelucchi, B., Hay, J., & Saffran, J. (2009a). Statistical Learning in Natural Language by 8-Month-Old Infants. *Child Development, 80(3)*, 674-685.

Pelucchi, B., Hay, J., & Saffran, J. (2009b). Learning in reverse: Eight-month-old infants track backward transitional probabilities. *Cognition, 113(2),* 244-247.

Phillips, C. (2006). The real-time status of island constraints. *Language, 82*, 795-823.

Phillips, C. (2012a). On the Nature of Island Constraints I: Language Processing and Reductionist Accounts. In J. Sprouse and N. Hornstein (eds.), *Experimental syntax and island effects.* Cambridge University press.

Phillips, C. (2012b). On the Nature of Island Constraints II: Language Learning and Innateness. In J. Sprouse and N. Hornstein (eds.), *Experimental syntax and island effects.* Cambridge University press.

Pullum, G., & Scholz, B. (2002). Empirical assessment of stimulus poverty arguments. *The Linguistic Review, 19,* 9–50.

Regier, T., & Gahl, S. (2004). Learning the unlearnable: The role of missing evidence. *Cognition, 93,* 147–155.

Reinhart, T. (1997). Quantifier Scope: How Labor is Divided Between QR and Choice Functions. *Linguistics and Philosophy, 20,* 335-397.

Rizzi, L. (1982). Violations of the wh-island constraint and the subjacency condition. In L. Rizzi (Ed.), *Issues in Italian Syntax.* Dordrecht, NL: Foris.

Rizzi, L. (1991). *Relativized minimality.* Cambridge, MA: MIT Press.

Roeper, T., & de Villiers, J. (2011). The Acquisition Path for Wh-Questions. In J. de Villiers & T. Roeper (Eds.), *Handbook of Generative Approaches to Language Acquisition*, *Studies in Theoretical Psycholinguistics 41*, (pp. 189-246). Springer: New York.

Ross, J. (1967). Constraints on variables in syntax. Doctoral dissertation, MIT, Cambridge, Mass.

Saffran, J., Aslin, R., & Newport, E. (1996). Statistical Learning by 8-Month-Old Infants. *Science, 274,* 1926-1928.

Saffran, J. R., Hauser, M., Seibel, R. L., Kapfhamer, J., Tsao, F., & Cushman, F. (2008). Grammatical pattern learning by infants and cotton-top tamarin monkeys. *Cognition, 107*, 479-500.

Saffran, J. R., Johnson, E. K., Aslin, R. N., & Newport, E. L. (1999). Statistical learning of tone sequences by human infants and adults. *Cognition, 70(1)*, 27-52.

Sagae, K., Davis, E., Lavie, A., MacWhinney, B., & Wintner, S. (2010). Morphosyntactic annotation of CHILDES transcript. *Journal of Child Language, 37(3),* 705-729.

Sakas, W.G. & Fodor, J.D. (2001). The structural triggers learner. In Bertolo, S. (Ed.) *Language Acquisition and Learnability*, (pp. 172-233). Cambridge, UK: Cambridge University Press.

Sampson, G. (1989). Language acquisition: growth or learning? *Philosophical Papers, 18*, 203-240.

Sampson, G. (1999). Collapse of the language nativists. *The Independent*, *April 9, 1999*, 7.

Scholz, B., & Pullum, G. (2002). Searching for arguments to support linguistic nativism. *The Linguistic Review, 19,* 185–223.

Schütze, C. (1996). *The empirical base of linguistics: Grammaticality judgments and linguistic methodology*. Chicago: The University of Chicago Press.

Schütze, C. & Sprouse, J. (2011). Judgment Data. In D. Sharma & R. Podevsa (Eds.), *Research Methods in Linguistics*.

Soderstrom, M., Conwell, E., Feldman, N., & Morgan, J. (2009). The learner as statistician: three principles of computational success in language acquisition. *Developmental Science, 12(3),* 409-411.

Sprouse, J. (2012). Deriving competing predictions from grammatical approaches and reductionist approaches to island effects. In J. Sprouse & N. Hornstein, (Eds.), *Experimental Syntax and Island Effects*. Cambridge University Press.

Sprouse, J. & Almeida, D. (2011). The role of experimental syntax in an integrated cognitive science of language. In K. Grohmann & C. Boeckx (Eds.) *The Cambridge Handbook of Biolinguistics*.

Sprouse, J. & Hornstein, N. (2012). *Experimental syntax and island effects*. Cambridge University Press.

Sprouse, J., M. Wagers, & C. Phillips. (2012). A test of the relation between working memory capacity and syntactic island effects. *Language, 88, 82-124.*

Stowe, L. (1986). Parsing wh-constructions: Evidence for on-line gap location. *Language and Cognitive Processes, 1*, 227–245.

Suppes, P. (1974). The semantics of children's language. *American Psychologist, 29,*103– 114.

Szabolcsi, A. & Zwarts, F. (1993). Weak islands and an algebraic semantics of scope taking. *Natural Language Semantics, 1*, 235-284.

Tenenbaum, J. & Griffiths, T. (2001). Generalization, similarity and Bayesian inference. *Behavioral and Brain Sciences, 24,* 629-640.

Tomasello, M. (2004). What kind of evidence could refute the UG hypothesis?, *Studies in Language, 28(3),* 642-645.

Torrego, E. (1984). On Inversion in Spanish and Some of Its Effects, *Linguistic Inquiry, 15,* 103-129.

Traxler, M.J., & Pickering, M.J. (1996). Plausibility and the processing of unbounded dependencies: An eye-tracking study. *Journal of Memory and Language, 35,* 454-475.

Tribus, M. (1961). *Thermodynamics and Thermostatics: An Introduction to Energy, Information and States of Matter, with Engineering Applications*. New York, NY.: D. Van Nostrand Company Inc.

Truswell, R. (2007). Extraction from adjuncts and the structure of events. *Lingua*, *117*, 1355–1377.

Tsai, W.-T. (1994). On nominal islands and LF extraction in Chinese. *Natural Language and Linguistic Theory, 12*, 121-75.

Valian, V. (1991). Syntactic subjects in the early speech of American and Italian children. *Cognition, 40,* 21–81.

Viau, J., & Lidz, J. (2011). Selective learning in the acquisition of Kannada ditransitives. *Language.*

Wagers, M., & Phillips, C. (2009). Multiple dependencies and the role of the grammar in real-time comprehension. *Journal of Linguistics, 45,* 395-433.

Wang, H. & Mintz, T. (2008). A Dynamic Learning Model for Categorizing Words Using Frames. In H. Chan, H. Jacob, & E. Kapia (Eds.), *BUCLD 32 Proceedings*, (pp. 525-536). Somerville, MA: Cascadilla Press.

Xu, F., & Tenenbaum, J. (2007). Word learning as Bayesian inference. *Psychological Review, 114*, 245-272.

Yang, C. (2002). *Knowledge and Learning in Natural Language.* Oxford: Oxford University Press.

Yang, C. (2004). Universal Grammar, statistics, or both? *Trends in Cognitive Sciences, 8(10),* 451-456.