Syntactic Islands without 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

Corresponding Author
Lisa Pearl
University of California, Irvine
Department of Cognitive Sciences
2314 Social & Behavioral Sciences Gateway Building
Irvine, CA 92697-5100
1-(949) 824-0156 Phone
1-(949) 824-2307 Fax
lpearl@uci.edu

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, sometimes called syntactic islands. 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 model that successfully learns the pattern of acceptability judgments observed in formal experiments, based on realistic input data. Crucially, while this modeled learner does require several types of learning biases to work in concert, it does not require any (clearly) innate, domain-specific biases. This suggests that syntactic islands constraints can in principle be learned without relying on UG. We discuss the consequences of this learner for the learning bias debates, as well as questions raised by the nature of the linguistic knowledge that is required by this learner.

Keywords
child-directed speech, computational modeling, language acquisition, probabilistic learning, syntactic islands, Universal Grammar

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 in fact 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; Lightfoot, 1989; Crain, 1991), 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 resolve it, and the central question is simply about the 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 the ones below:

(i) a sensitivity to the distributional data in the available input

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

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)

Notably, many of these proposed learning biases can be (and have been) combined with aspects of the UG hypothesis (e.g., statistical learning using distributional data of representations defined

by UG: Legate & Yang, 2007; Mitchener & Becker, 2011; Pearl, 2011; Pearl & Lidz, 2009; Pearl & Mis, 2011; Sakas & Fodor, 2001; Yang, 2002; Yang 2004). However, again, often the idea of UG opponents is to supplant all UG biases with biases that would clearly not be part of UG. To this end, it is worth clarifying what makes a learning bias part of UG. We suggest that learning biases may be categorized along (at least) three dimensions:

(i) Are they *domain-specific* or *domain-general* ?
(ii) Are they *innate* or *derived* from prior experience?
(iii) 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. All of the learning biases mentioned above, for example, are domain-general (either innate or derived).

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 how an ideal learner using Bayesian inference can choose structure-dependent 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 structure-dependent representations; instead, this knowledge can be derived from a domain-general sensitivity to the distributional properties of the data. Notably, children must still know that structure-dependent representations are possible – but they do not need to have competing representations ruled out a priori.

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 how 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 expanded the range of informative data beyond the traditional data set of unambiguous data. This kind of learner suggested 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 so long as the child knew to ignore certain kinds of ambiguous data, and they proposed a learning preference for learning in cases of local uncertainty in order to achieve this. 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 empirical results from children so 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) suggested that an additional strategy was still needed to reach the adult knowledge state, perhaps similar to the one proposed in Foraker et al. (2009). Foraker at el. (2009) demonstrated that an ideal Bayesian learner who also has detailed linguistic knowledge about the link between semantic interpretation and certain syntactic structures (syntactic complements and syntactic modifiers) can use the difference in distribution for *one* with these structures to converge on the correct knowledge for *one*. Though the learning mechanism is domain-general, it is unclear if

the detailed linguistic knowledge necessary can be derived through domain-general means or would instead be part of UG.

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 electronic 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., structure-dependence and anaphoric *one*) can in principle be learned without innate, domain-specific biases (although there are some questions as to whether the end-states of these learning models are identical to the end-states that linguists hypothesize for adult speakers; see Pearl & Mis (2011, submitted) for this issue with respect to anaphoric *one*).

Although these findings have substantially advanced our understanding of the acquisition of some aspects of syntax, the UG hypothesis is still the dominant assumption for a large segment of the field of linguistics. We believe there may be two reasons for this. First, the phenomena that have been investigated so far are not considered central to the syntactic theories of UG proponents. In other words, 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, 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 is relatively scarce. For example, 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 we need to test syntactic learning models with, and (ii) investigating the learning biases required to learn a set of phenomena that is undeniably central to (UG-based) syntactic theories – 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. Next, we syntactically annotated three 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 not only identified the data from which syntactic islands must be learned, but 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 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. We note that this learner does require several types of learning biases to work in concert for the acquisition of syntactic island constraints (in particular, combining domain-general statistical learning methods with more abstract domain-specific representations, similar to previous acquisition models (Foraker et al., 2009; Legate & Yang, 2007; Mitchener & Becker, 2011; Pearl & Lidz, 2009; Pearl, 2011; Pearl & Mis, 2011; Pearl & Mis, *submitted*; Pearl & Sprouse, *forthcoming;* Perfors et al., 2011; Regier & Gahl,

2004; Yang, 2002; Yang, 2004). However, it crucially does not require any (clearly) innate, domain-specific biases. Given that the UG hypothesis requires that at least one of the necessary learning biases is innate and domain-specific, even if other necessary learning biases are not, we take this as evidence that syntactic island constraints can in principle be learned from child-directed speech without UG. Though this statistical learner does not require any clearly innate, domain-specific biases, it should be noted that it does rely on several types of fine-grained linguistic knowledge, such as a distinction between different types of Complementizer Phrase (CPs). While we are reluctant to label this fine-grained linguistic knowledge as UG, questions still remain as to how this fine-grained linguistic knowledge is itself learned. As such, we will suggest that these sophisticated biases may arise based on the interaction of the other independently motivated 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 three 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 three child-directed speech corpora and also on adult-directed speech and text corpora. Section 5 provides a general discussion of the consequences of this learner for the learning bias debates, as well as questions raised by the nature of the linguistic knowledge that is required by this learner. 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 can 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 held that the dependency between a displaced element (e.g., a *wh*-word) and the gap position could not cross two or more *bounding nodes* (Chomsky, 1973; Huang, 1982; Lasnik & Saito, 1984; and many others). The definition of *bounding nodes* could vary from language to language in order to account for the various patterns of island effects that had been observed cross-linguistically. For example, the bounding nodes in English were argued to be NP (Noun Phrase) and IP (Inflection Phrase) (Chomsky, 1973), and bounding nodes in Italian and Spanish were argued to be NP and CP (Complementizer Phrase) (Rizzi, 1980; Torrego, 1984). Crucially, this framework assumed that the Subjacency Condition itself was part of UG, as were the possible options for bounding nodes (NP, IP, or CP). The language learner then simply needed to determine which bounding nodes were 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 *phases*) 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 to us 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 asterisks/no-asterisks 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), Subject islands (2b), Whether islands (2c), and Adjunct islands (2d).

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 (LENGTH: short, long; STRUCTURE: 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).

(6) Complex NP islands

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

(7) Subject islands

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

(8) Whether islands

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

(9) Adjunct islands

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

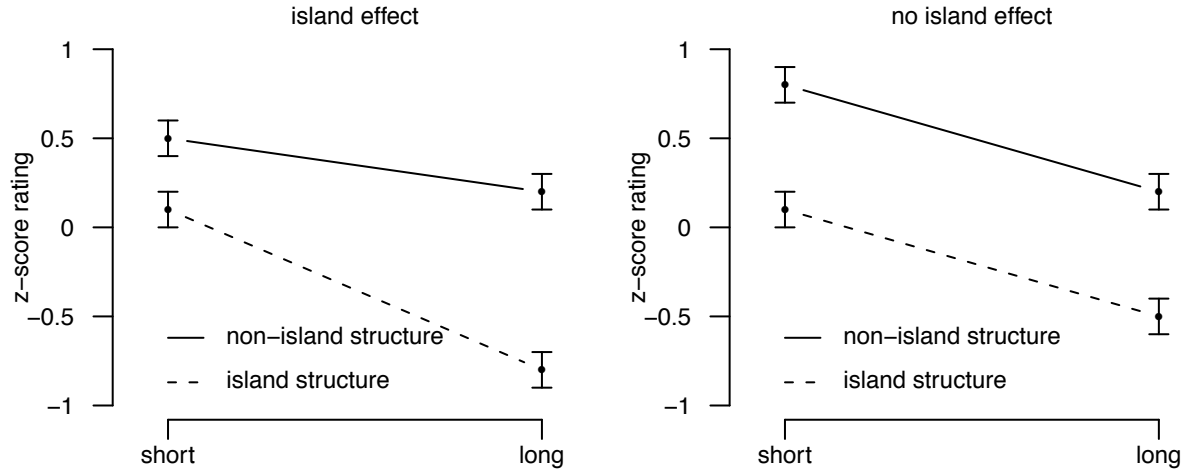
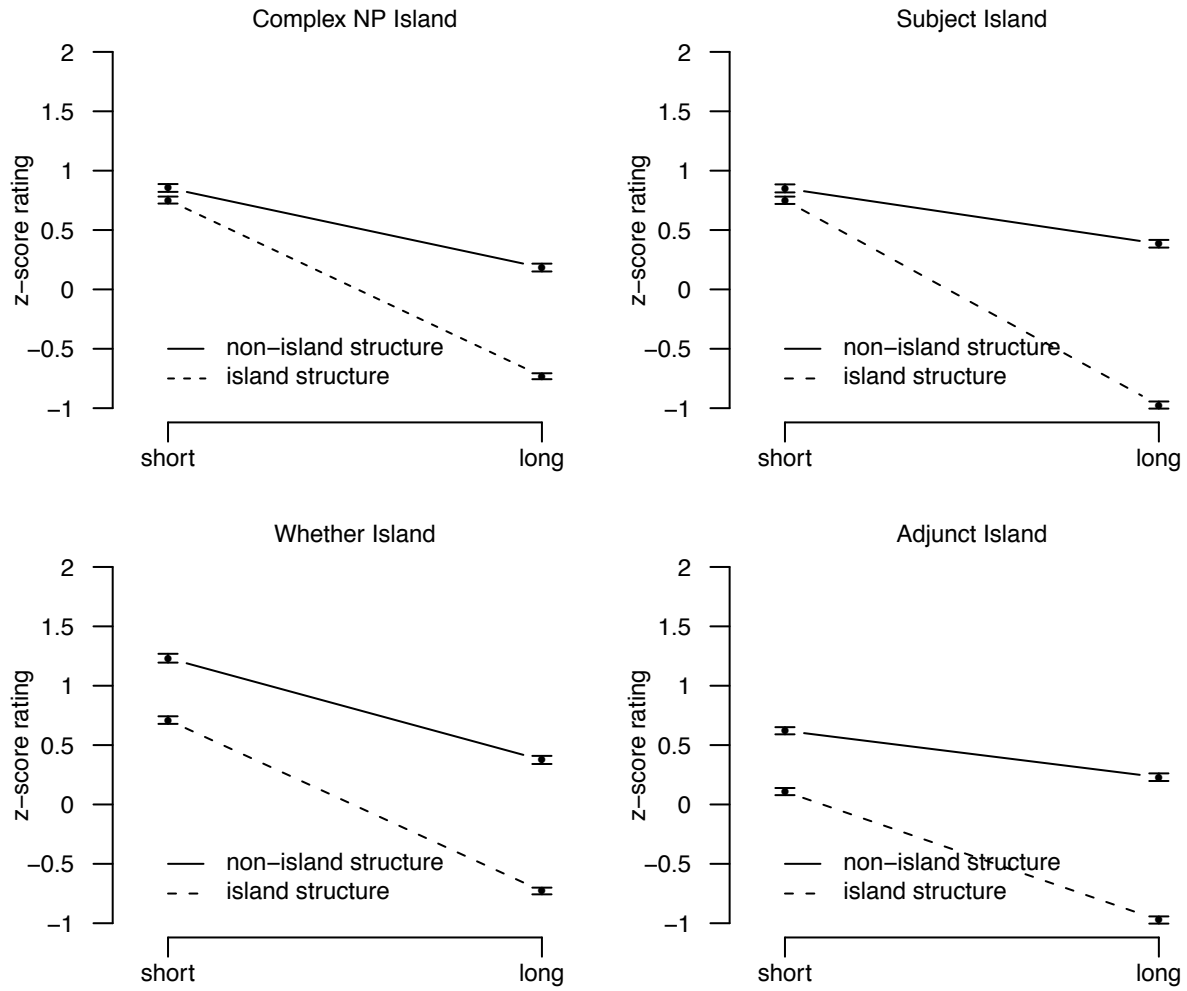


Figure 2 plots the experimentally obtained judgments for the island types investigated in Sprouse et al. (2012), which shows that adult speakers appear to have implicit knowledge of these four syntactic islands. We can thus use the superadditive interactions for the four island types in Figure 2 as an 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 three well-known corpora of child-directed speech from the CHILDES database (MacWhinney, 2000) to annotate with phrase structure tree information: the Adam and Eve corpora from the Brown data set (Brown, 1973), and the Valian dataset (Valian, 1991). We first automatically parsed the child-directed speech utterances using a freely available syntactic parser (the Charniak parser[1]), yielding the basic phrase tree structures. 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.[2] However, in case they were not, we 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).

The data from these three corpora comprise child-directed speech to 23 children between the ages of one and four years old, with 340,913 word tokens total. Of all the utterances, 14,260 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).

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.[3]

	SHORT | NON-ISLAND	LONG | NON-ISLAND	SHORT | ISLAND	*LONG | ISLAND*
Complex NP	4	177	0	0
Subject	4	13	0	0
Whether	4	177	0	0
Adjunct	4	177	3	0

From Table 1, we can see that these utterance types are fairly rare in general, with the most frequent type (LONG | NON-ISLAND) appearing 0.01% of the time (177 of 14,260). Secondly, we see that being grammatical doesn't necessarily mean an utterance type will occur in the input. Specifically, while both the SHORT | NON-ISLAND and SHORT | ISLAND utterance types are grammatical, they rarely occur in the input (4 for SHORT | NON-ISLAND, between 0 and 3 for SHORT | ISLAND). This is problematic from a learning standpoint, if a learner is keying grammaticality intuitions directly to input frequency. Unless the child is very sensitive to small

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

[2] 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.

[3] Note that the number of SHORT | 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 LONG | 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)

frequency differences (3 or 4 out of 14260 is less than 0.001% of the relevant input), the difference between the frequency of grammatical SHORT | ISLAND or SHORT | NON-ISLAND utterances and that of ungrammatical LONG | 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 SHORT | ISLAND utterances and ungrammatical LONG | ISLAND structures is nonexistent. Since neither utterance type appears in the input, how would this learner classify one as grammatical and the other ungrammatical? Thus, it appears that child-directed speech input presents an induction problem to a learner attempting to acquire adult grammaticality intuitions about 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 grammaticality judgments. 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. We explore this option in the learning algorithm we describe in the next section.

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 that solves it in order for them to become adults who have the acceptability judgments observed in Sprouse et al. (2012).[4] 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

[4] 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, 2011).

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; Perfors et al., 2011); 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 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 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 its *container nodes*. A simple way to identify the container nodes is simply those phrasal constituents currently unclosed (opened with a left bracket), given the understood position of the dependencies. 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 can extract this information from the phrase structure trees is actually a well-established fact of the behavior of the human sentence parser.

(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.

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.[5] For example, the container node

[5] 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

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):

(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.[6] For example, in (9), 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[7] are schematized in Figure 3, and two examples of grammaticality preferences are shown in (14) and (15).

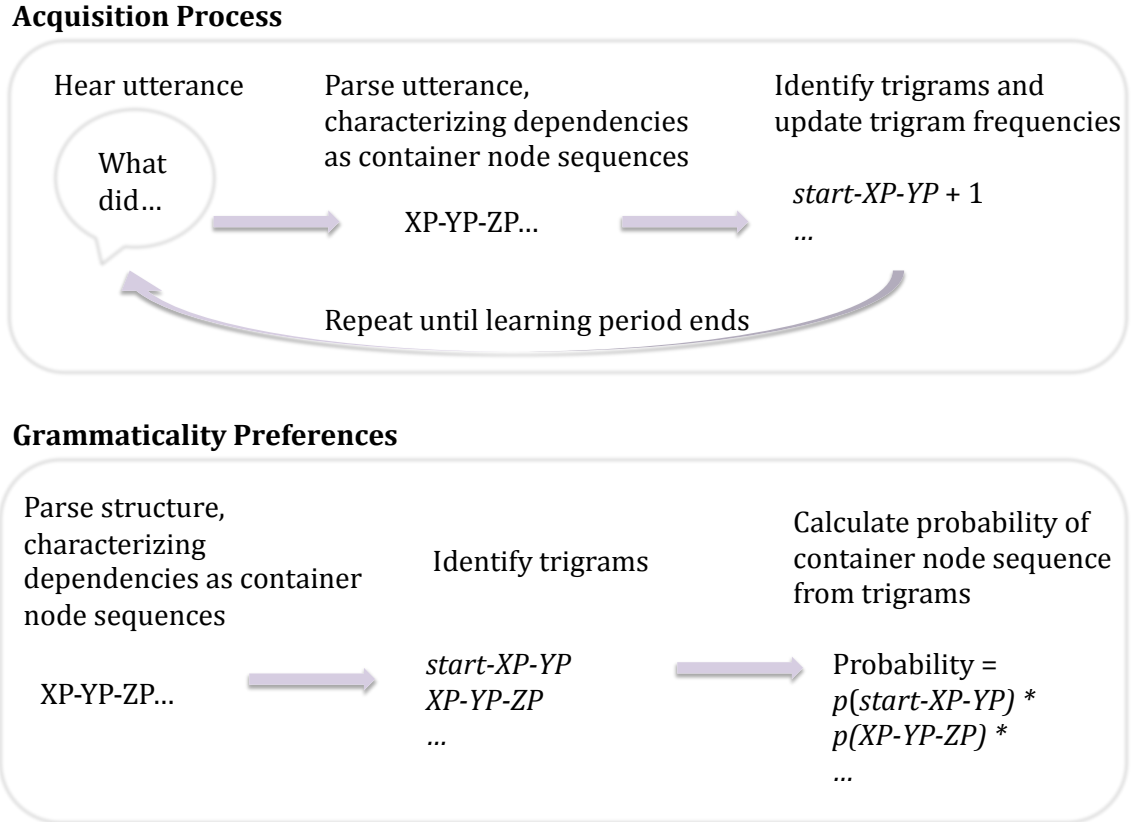
---

expands the relevant intake set of the learner – all dependencies are informative, not just the ones being judged as grammatical or ungrammatical.

[6] 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.

[7] 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

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)

---

of section 4, an acceptability judgment is the result of several factors, of which the 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 machine learning 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.[8] However, that is not likely to be 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[9], 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 leaning model. One 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, such as noting the complementizer used for a CP (e.g., *that*, *whether, if,* null, etc.) and subcategorizing CP container nodes based on the specific

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

[9] 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.

complementizers (e.g., CP_{that} vs. $CP_{whether}$ vs. CP_{if} vs. CP_{null}, etc.). 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.* Another possibility is that learners could subcategorize CP container nodes from the outset, perhaps because children's linguistic experience has already highlighted the fact that different complementizers have different semantic and pragmatic implications 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 5.5 for a discussion of the relationship between computational learning models and hypotheses about the time-course of acquisition).

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, and therefore none of these biases appear to be part of a UG-based approach to the acquisition of island constraints. In other words, the learning model that we have constructed here is not based on the UG hypothesis.

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 requires a bias that is both domain-specific and innate simultaneously.

Description of process	**Domain-specific**	**Domain-general**	**Innate**	**Derived**
Parse utterance & identify dependencies	*			*
Identify container nodes	*			*
Extract trigram sequences		*	*	
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 three corpora: child-directed speech from the Adam and Eve corpora from Brown (1973) and the Valian corpus (Valian, 1991) 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	65932	74576	24243
total *wh*-dependencies	11308[10]	8508	4230

[10] Note that this is smaller than the number of utterances containing both *wh*-words and verbs reported in section 3. This is because not every utterance that contains a *wh*-word and a verb actually has a *wh*-dependency (e.g., *What about seeing the movie tomorrow?*)

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?	11.3% 1274	17.2% 1464	33.0% 1396
IP-VP What did she see?	80.4% 9092	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_{null}-IP Who did he think stole it?	0.1% 13	0.6% 52	0.3% 12
IP-VP-CP_{null}-IP-VP What did he think she stole?	1.4% 159	0.4% 30	0.2% 8
IP-VP-CP_{null}-IP-VP-IP-VP What did he think she wanted to steal?	<0.1% 6	<0.1% 3	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-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?	0.2% 20	<0.1% 5	<0.1% 1
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	<0.1%	<0.1%	0.0%

Who did he want to steal the necklace?	2	2	0
IP-VP-IP-VP What did he want her to steal?	3.3% 369	3.4% 287	1.3% 57
IP-VP-IP-VP-IP-VP What did he want her to pretend to steal?	0.0% 0	<0.1% 6	<0.1% 1
IP-VP-IP-VP-IP-VP-PP Who did he want her to pretend to steal from?	0.0% 0	<0.1% 6	0.0% 0
IP-VP-IP-VP-NP What did he want to say about it?	<0.1% 5	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?	0.3% 30	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?	0.4% 45	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% 2	<0.1% 6	0.0% 0
IP-VP-PP Who did she steal from?	2.5% 282	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	0.0%	<0.1%	0.0%

Where was she at in the building?	0	2	0
IP-VP-PP-NP-PP	0.0%	0.0%	0.0%
What do you put it on top of?	2	0	0
IP-VP-PP-NP-PP-IP-VP	0.0%	<0.1%	0.0%
What is she in the habit of doing?	0	1	0
IP-VP-PP-PP	<0.1%	0.0%	0.0%
What does he eat out of?	1	0	0
IP-VP-PP-VP	<0.1%	0.0%	0.0%
What did he think about stealing?	1	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: 80.4%/11.3%, 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) and Valian (Valian 1991), and estimating the proportion of *wh*-dependencies (11,308) to total utterances (65,932), we set the learning period to 175,000 *wh*-dependency data points. So, our learners will encounter 175,000 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.[11]

(16) Complex NP islands

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

(17) Subject islands

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

(18) Whether islands

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

(19) Adjunct islands

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

[11] 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).

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 4. 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 intuitions 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 of acceptability judgments. All of the 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.[12] 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. 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.

[12] 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. 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.

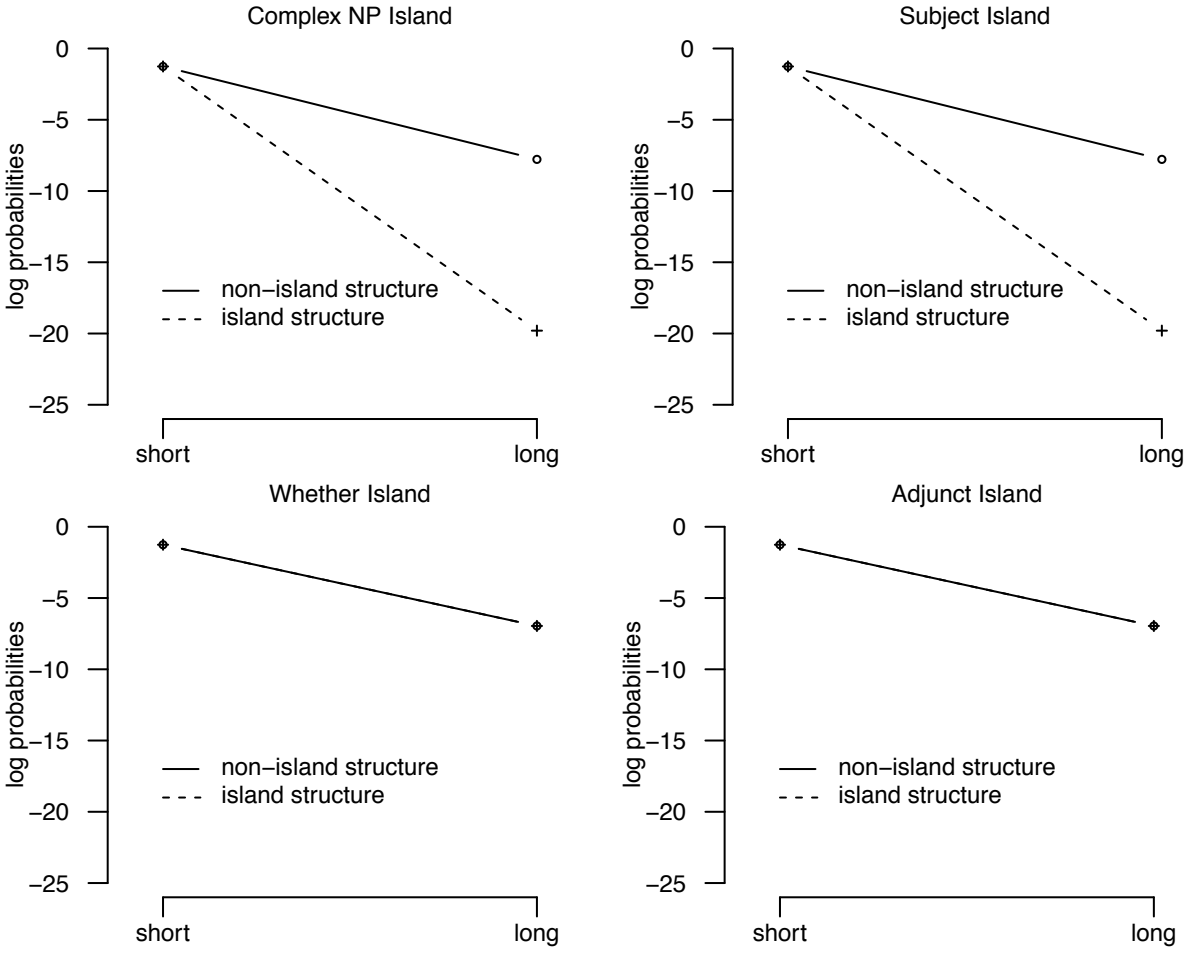


Figure 5. Log probabilities derived from adult-directed speech and text for a learner that does not discriminate CP node types. The apparent lack of dashed "island structure" line in the Whether and Adjunct island graphs indicates that the line is identical to the solid "non-island" structure line, as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

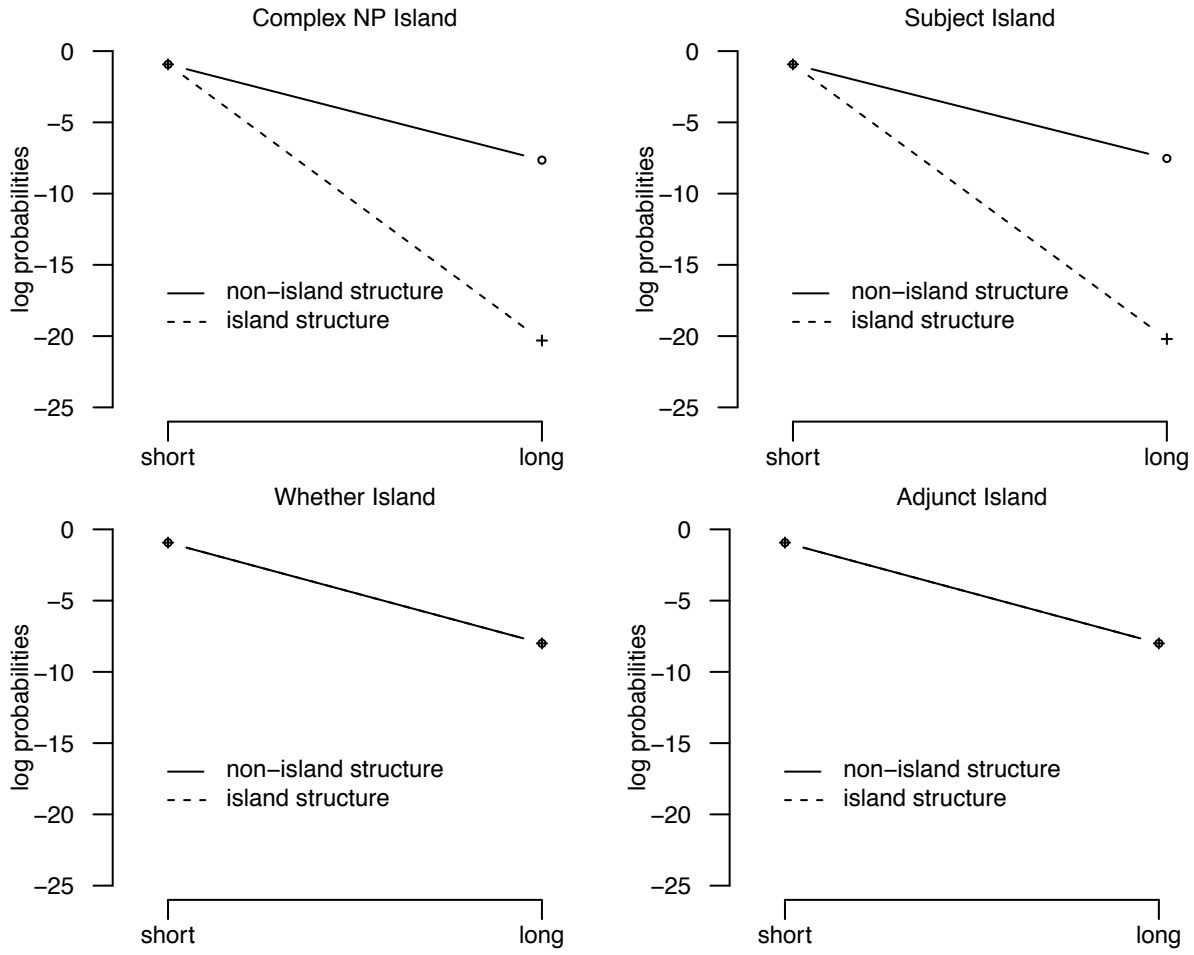


Table 5. Inferred acceptability of different *wh*-dependencies from Sprouse et al. (2012), represented with log probability.

		Child-directed speech	Adult-directed speech & text
Control dependencies			
matrix subject	IP	-1.26	-0.93
embedded subject	IP-VP-CP-IP	-7.76	-7.53
embedded object	IP-VP-CP-IP-VP	-6.96	-8.00
Island-spanning dependencies			
Complex NP	IP-VP-NP-CP-IP-VP	-18.01	-18.08
Subject	IP-VP-CP-IP-NP-PP	-19.85	-20.17
Whether	IP-VP-CP-IP-VP	-6.96	-8.00
Adjunct	IP-VP-CP-IP-VP	-6.96	-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	15.94	18.59	5.70	5.70
	embedded subject	--	12.07	--	--
	embedded object	10.24	--	0.00	0.00
	Adult-directed speech & text				
	matrix subject	17.14	19.24	7.07	7.07
	embedded subject	--	12.64	--	--
	embedded object	10.07	--	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-$CP_{\textit{null/that}}$-IP-VP and structures with the sequence IP-VP-$CP_{\textit{whether/if}}$-IP-VP , which means that Whether and Adjunct island violations, which contain specific types of CPs ($CP_{\textit{whether}}$ and $CP_{\textit{if}}$), are treated identically to grammatical utterances containing $CP_{\textit{null}}$ or $CP_{\textit{that}}$, 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_{adjunct}-IP-VP (e.g., 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 linguistically sophisticated 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.

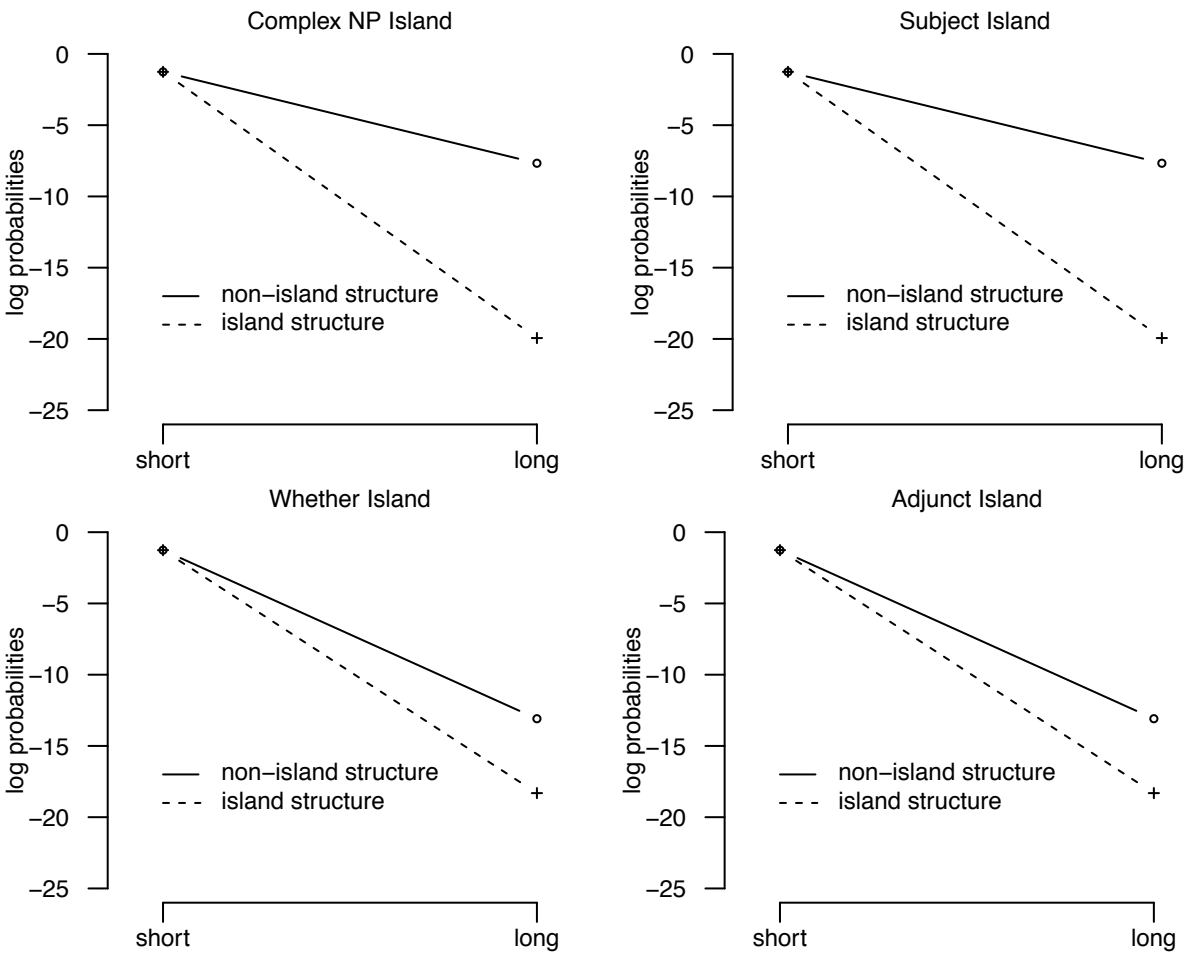


Figure 7: Log probabilities derived from adult-directed speech and text for a learner that discriminates CP types.

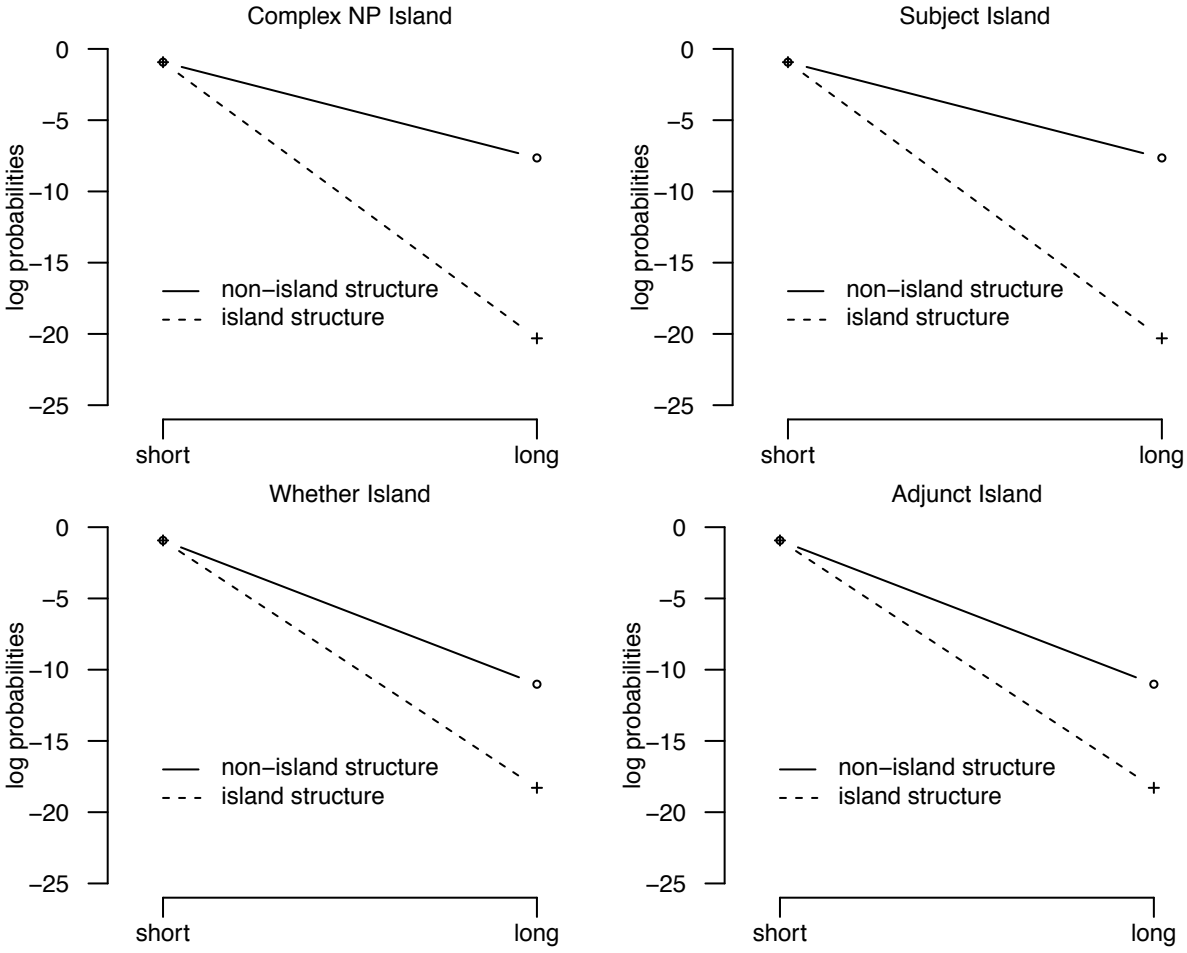


Table 7. Inferred grammaticality of different *wh*-dependencies from Sprouse et al. (2012), represented with log probability.

		Child-directed speech	Adult-directed speech & text
Control dependencies			
matrix subject	IP	-1.26	-0.93
embedded subject	IP-VP-CP_{null}-IP	-7.68	-7.65
embedded object	IP-VP-CP_{that}-IP-VP	-13.06	-11.02
Island-spanning dependencies			
Complex NP	IP-VP-NP-CP_{that}-IP-VP	-19.22	-18.87
Subject	IP-VP-CP_{null}-IP-NP-PP	-19.94	-20.31
Whether	IP-VP-$CP_{whether}$-IP-VP	-18.32	-18.29
Adjunct	IP-VP-CP_{if}-IP-VP	-18.32	-18.29

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	17.97	18.68	17.06	17.06
	embedded subject	--	12.26	--	--
	embedded object	6.13	--	5.22	5.22
	Adult-directed speech & text				
	matrix subject	17.94	19.38	17.36	17.36
	embedded subject	--	12.67	--	--
	embedded object	7.85	--	7.27	7.27

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 5 times more grammatical than the island-spanning dependencies. 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.

5. General Discussion

In this study, we investigated an acquisition problem previously believed to 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 (in a similar spirit to models by Foraker et al., 2009; Pearl & Mis, 2011; submitted; Perfors et al., 2011; and Regier & Gahl, 2004) is able to reach the target knowledge state, where dependencies spanning syntactic islands are perceived as ungrammatical. The statistical learning model itself included two derived, domain-specific learning biases, and three innate, domain-general learning biases, but crucially did not contain any clear instances of innate, domain-specific learning biases (see also Table 2). This suggests that syntactic island effects can in principle be learned without the UG hypothesis. However, these results do raise interesting questions about the role of sophisticated linguistic knowledge in the learning process (and relatedly, how that linguistic knowledge is learned), as well as how feasible this learner would be for the full range of constraints on *wh*-dependencies. We turn to these questions presently.

5.1. Is tracking trigram sequences of container nodes 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. However, to our knowledge no one has proposed combining these two assumptions into one: that children track trigrams of container node sequences. This is clearly an example of a relatively sophisticated learning bias, though it is also clearly less sophisticated than canonical UG hypotheses (e.g., the Subjacency Condition in section 2). Furthermore, this particular learning bias is difficult to classify according to the taxonomy laid out in section 1. On the one hand, the ability to track trigram sequences is likely innate and domain-general. On the other hand, the identification of container nodes is likely derived and domain-specific. The question then is what the status of the interaction of the two is. One possibility is that the interaction of two existing learning biases is the result of another learning bias (an innate, domain-general bias to combine existing biases), in which case this particular bias would simply be the consequence of three separate biases, none of which is part of the UG hypothesis. This raises an interesting possibility that many of the phenomena that have appeared to require the UG hypothesis may in fact be learnable through the complex interaction of non-UG learning biases.

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 in fact built upon a series of independent (and non-UG) biases.

5.3. The problem of parasitic gaps

Though this statistical model demonstrates that syntactic islands can in principle be learned from child-directed input, this particular model cannot capture certain 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 $\_\__{\text{true}}$ [before reading $\_\__{\text{parasitic}}$]?

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

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).

We believe the grammaticality of parasitic gap constructions pose 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 relevant syntactic island violation. In other words, our learner would infer that parasitic gap constructions are ungrammatical. 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. We are currently examining additional syntactically-annotated child-directed corpora to answer this (and other) questions.

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

First and foremost, it appears that syntactic island effects – a set of phenomena that are central to (UG-based) syntactic theories – do not in principle require UG to be learned. However, it is also interesting to note that the acquisition of syntactic island effects did not require altering the syntactic analysis of island constraints. More specifically, the (implicit) output of the learning model looks very similar to existing theories of syntactic islands: 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. In this case, we relied upon several uncontroversial assumptions, and the idea that several simple learning biases can interact to produce more sophisticated learning biases.

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). 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

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.

5.5. 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).

6. Conclusion

By examining a particular acquisition problem considered as motivation for UG, we have been able to concretely determine that it does not, in fact, require UG to solve (though UG-like learning biases are certainly one solution to the problem). After first verifying that there was an induction problem for children, we then 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. In addition to only using learning biases that would not be considered part of the UG hypothesis, this type of learner also 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 be derived from the interaction of independently motivated (non-UG) biases, thus reducing the motivation for the UG hypothesis. 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 all 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 Tom Roeper, the attendees of the Input & Syntactic Acquisition workshop held at UCI 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. This work was supported in part by NSF grant BCS-0843896.

Role of the funding source

The NSF provided the support that enabled the authors to annotate the syntactic acquisition data for this study, and supported the use of computational modeling methodology to investigate the acquisition of island constraints. In addition, the NSF provided support for the collection of acceptability judgment data in Sprouse et al. (2012) that were used as an empirical baseline for the current study.

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.

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. (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.

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.

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. (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. & 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.

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). Defining the terms of the debate: Reductionist theories and the superadditive nature of 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., M. Wagers, & C. Phillips. (2012). A test of the relation between working memory capacity and syntactic island effects. *Language.*

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.

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.