Computational models of acquisition for islands

Lisa Pearl & Jon Sprouse

1. Why look at language acquisition?

Though it is not always directly stated, the debate at the center of this volume is in many ways driven by language acquisition considerations. Long-distance dependencies are themselves relatively complex, as they involve context-sensitive grammatical operations (e.g., wh-movement or slash-passing). The existence of context-sensitive operations alone increases the complexity of the hypothesis space of possible grammars that must be considered by children during the acquisition process. If island effects are indeed the result of grammatical constraints, then the hypothesis space increases yet again, as the grammar must also contain complex constraints on context-sensitive operations. A common hypothesis in the generative syntax literature is that this level of complexity (constraints on context-sensitive grammatical operations) cannot be learned directly from the input that children receive (i.e., this is a poverty of the stimulus problem). As such, many generative syntacticians have postulated the existence of innate domain-specific knowledge about the form that such constraints must take. In other words, the grammatical approach to island effects has often correlated with a nativist, or Universal Grammar (UG) based, view of language acquisition. In this way, a reductionist approach to island effects could be seen as a type of simplifying approach to the grammar, as it could eliminate the need for one set of innate constraints on the shape of human grammars. Because of this, it seems to us that discussions of “parsimony” and “simplification” in the reductionist literature either directly or indirectly concern the presumed problem that occurs during language acquisition.

Given the amount of research that has been conducted on the debate between grammatical and reductionist approaches to island effects, it seems important at this stage to determine exactly what type of innate knowledge (if any) would be necessary to learn the grammatical constraints that give rise to island effects, given the input that children receive during language acquisition. Such an investigation will help determine exactly what is at stake in this debate. If grammatical island constraints cannot be learned from the input available to children without innate domain-specific knowledge (UG), then this debate has direct implications for the language acquisition process. However, if grammatical island constraints can be learned from the input available to children without UG-like knowledge, then this debate is simply one empirical question among the hundreds that must be answered in order to have a complete theory of language.

In this chapter, we examine child-directed speech input in order to formalize the apparent induction problem that has been claimed by linguists. We then explore a statistical learning model of island constraints that is based upon the frequency of certain abstract structures in the input. The model is tested on input derived from child-directed speech (from CHILDES: MacWhinney (2000)) as well as input derived from adult-directed speech (Switchboard section of Treebank-3: Marcus et al. 1999) and adult-directed text (Brown section of Treebank-3: Marcus et al. 1999). We use this statistical model to investigate the types of learning biases that are necessary to learn these constraints from the input, with the goal of determining whether any innate domain-specific biases (i.e., UG) are necessary. Our results suggest that a learner only requires the following biases to learn syntactic island constraints from child-directed input, none of which are considered specific to the nativist/UG approach to language acquisition:

(i) perceive the input with a phrase-structure-based representation of sentences (i.e., a parser)

(ii) track the frequency of sequences of three phrase structure nodes (trigrams of phrase structure nodes), and their associated probability of occurring

(iii) construct a longer dependency by combining trigrams of phrase structure nodes, and assess that dependency's grammaticality based on that combination

The fact that syntactic island constraints can be learned from realistic child-directed and adult-directed input without any nativist/UG-specific abilities suggests that the grammatical versus reductionist debate has no implications for the debate between nativists and non-nativists, but is instead just one question among many required to fully understand the human language system.

2. The induction problem

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

(1), Subject islands (2), Whether islands (3), and Adjunct islands (4). Sprouse et al. (2012) used a factorial definition of island effects for each island type (see Sprouse (*this volume*) for discussion of the value of the factorial definition of island effects). For our purposes, this simply means that each island type was defined by four sentence types (4 island types x 4 sentence types = 16 sentence types). An example of each sentence type and the resulting container node sequence is given in (1)–(4): (a) matrix gap, non-island structure, (b) embedded gap, non-island structure, (c) matrix gap, island structure, (d) embedded gap, island structure.

(1) Complex NP islands

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

(2) Subject islands

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

(3) Whether islands

a. Who __ thinks that Jack stole the necklace? MATRIX | NON-ISLAND
b. What does the teacher think that Jack stole __ ? EMBEDDED | NON-ISLAND

c. Who __ wonders whether Jack stole the necklace? MATRIX | ISLAND

d. *What does the teacher wonder whether Jack stole __ ? EMBEDDED | ISLAND

(4) Adjunct islands

a. Who __ thinks that Lily forgot the necklace? MATRIX | NON-ISLAND

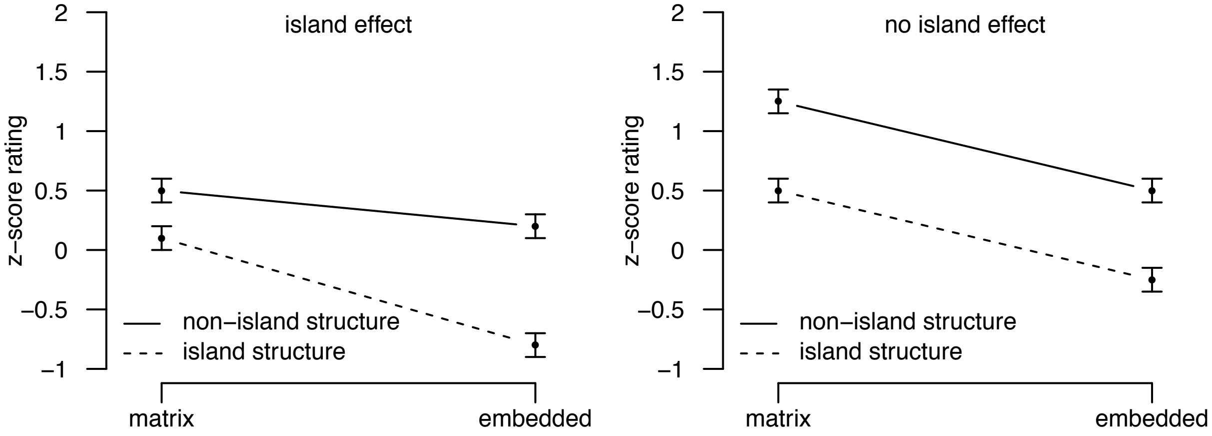
b. What does the teacher think that Lily forgot __ ? EMBEDDED | NON-ISLAND

c. Who __ worries if Lily forgot the necklace? MATRIX | ISLAND

d. *What does the teacher worry if Lily forgot __ ? EMBEDDED | ISLAND

The factorial definition of island effects makes the presence of an island effect visually salient: if we plot the acceptability of the four sentence types in a configuration known as an interaction plot, the presence of an island effect shows up as two non-parallel lines, which indicates a statistical interaction of the two factors in the definition (the left panel of Figure 1); the absence of an island effect shows up as two parallel lines, which indicates no interaction of the two factors in the definition (the right panel of Figure 1).

Figure 1: Example graphs showing the presence (left panel) and absence (right panel) of island effects using the factorial definition (see also Sprouse (*this volume*)).



Sprouse et al. (2012) found that adult judgments demonstrated an island effect for all four island types, which means that knowledge of these syntactic islands is indeed necessary to acquire.

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 a well-known corpus of child-directed speech from the CHILDES database (MacWhinney, 2000) to annotate with phrase structure tree information: the Adam corpus from the Brown data set (Brown, 1973). 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

---

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

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 this corpus is comprised of child-directed speech to one child between the ages of two and five years old, with 124,285 word tokens total. In all the utterances, there were 4,795 wh-dependencies. Table 1 shows the number of examples 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]

	MATRIX | NON-ISL	EMBED | NON-ISL	MATRIX | ISLAND	*EMBED | ISLAND*
Complex NP	1	113	0	0
Subject	1	7	0	0
Whether	1	113	0	0
Adjunct	1	113	3	0

---

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

From Table 1, we can see that these utterance types are fairly rare in general, with the most frequent type (LONG | NON-ISLAND) appearing 0.02% of the time (113 of 4,795). Secondly, we see that being grammatical doesn't necessarily mean an utterance type will occur in the input. Specifically, while both the MATRIX | NON-ISLAND and MATRIX | ISLAND utterance types are grammatical, they rarely occur in the input (1 for MATRIX | NON-ISLAND, 3 for some of the MATRIX | ISLAND types). This is problematic from a learning standpoint, if a learner is keying grammaticality directly to input frequency. Unless the child is very sensitive to small frequency differences (1 or 3 out of 4,795 is less than 0.001% of the relevant input), the difference between the frequency of grammatical MATRIX | ISLAND or MATRIX | NON-ISLAND utterances and that of ungrammatical EMBEDDED | ISLAND utterances is very small for Adjunct island effects. It's even worse for Complex NP, Subject, and Whether island effects, since the difference between grammatical MATRIX | ISLAND utterances and ungrammatical EMDBEDDED | 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 grammatical knowledge 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 section 4.

3. The acquisition process

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 complete description of the acquisition process must contain at least the following:

(i) a specification of the child's representation(s) of the hypothesis space
(ii) a representation of the input that is available to children (the *intake* (Fodor 1998a))
(iii) the updating procedure that is used to navigate the hypothesis space

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. For example, almost all theories assume that children must have a bias to represent their hypotheses about linguistic structures as abstract phrase structure trees. Nativist/UG-based theories may go even further and assume an even more abstract hypothesis space, perhaps in the form of primitives necessary for innate syntactic constraints (e.g., bounding nodes for the Subjacency condition (Chomsky 1973)). Similarly many theories assume that children have a bias to use probabilistic reasoning to update their beliefs about which structures are grammatical (e.g.,Tenenbaum & Griffiths 2001, Griffiths & Tenenbaum 2005, Gerken 2006, Xu & Tenenbaum 2007, Frank et al. 2009). Nativist/UG-based theories may again go even further by assuming that a single occurrence of a given structure is enough to instantiate a given grammar (e.g., triggers (Lightfoot 1991, Gibson & Wexler 1994, Niyogi & Berwick 1996, Fodor 1998a, Dresher 1999, Lightfoot 2010, among others)). Formally modeling these allows us to see the effect of any given learning bias on acquisition, and determine which biases are necessary. Once we have that, we can then investigate the nature of the necessary biases to determine if they qualify as unique to nativist/UG-based approaches to acquisition, or are shared by non-nativist theories of acquisition.

The question of whether a given learning bias is nativist or non-nativist in nature is actually quite a bit more complex than is often assumed in the syntactic literature. For example, there at least three dimensions to learning biases that may be relevant (Pearl & Mis 2011, submitted):

(i) Are they *innate* (and so part of the human biological endowment) or *derived* from prior experience (probably prior experience with language data)?

(ii) Are they *domain-specific* (and are only used for learning language) or *domain-general* (and are used when learning anything)?

(iii) Are they about *the hypothesis space* (and so may restrict the learner's hypotheses explicitly) or about *the learning mechanism* (and so may restrict the learner's hypotheses implicitly)?

Clearly, learning biases could involve any logically possible combination of these dimensions. 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*. Note that only learning biases that are both *innate* and *domain-specific* are candidates for UG. An explicit constraint against syntactic islands would be just this kind of bias, since it would be *innate* (it's explicitly built in) and *domain-specific* (it applies only to language). 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 islands.

4. The acquisition process for islands

We turn now to a proposal about how to learn the grammatical status of various long-distance dependencies in English. 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 (5a), 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 the gap location, 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. Since container nodes play an integral role in all syntactic formulations of island constraints, they therefore 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) that tracks the container nodes of the gap location (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. For (5a), the container nodes would be the sequence in (5b), where the gap location of the displaced NP *who* is dominated by the matrix VP and then the matrix IP. We can represent this dominance information as a sequence of container nodes, as in (5c). Another example is shown in (6a)-(6c), with the utterance "Who did she think the gift was from?" Here, the gap position of the displaced NP *who* is dominated by several nodes (6b). This can be represented by the container node sequence in (6c).

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

(6) 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 (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 representing the input in a particular way. 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 representation of *the hypothesis space* (phrase structure trees), though the units over which this process operates are likely *derived*.

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.[4] For example, the container node sequences from (5c) would be represented as a sequence of trigrams as in (7c), and the container node sequences from and (6c) would be represented as a sequence of trigrams as in (8c):

(7) a. [$_{CP}$ Who did [$_{IP}$ she [$_{VP}$ like [$_{NP}$ __]]]]]?

b. IP VP

c. start-IP-VP-end =

start-IP-VP

IP-VP-end

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

b. IP VP CP IP VP PP

---

[4] Note that this means a learner is learning from data containing dependencies besides the one of interest. 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 an implicit learning bias that expands the relevant intake set of the learner – all dependencies are informative, not just the ones being judged as grammatical or ungrammatical.

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.[5] For example, in (3), the sequence IP-VP would have a probability equal to the product of the trigram *start-IP-VP* and the trigram *IP-VP-end*.

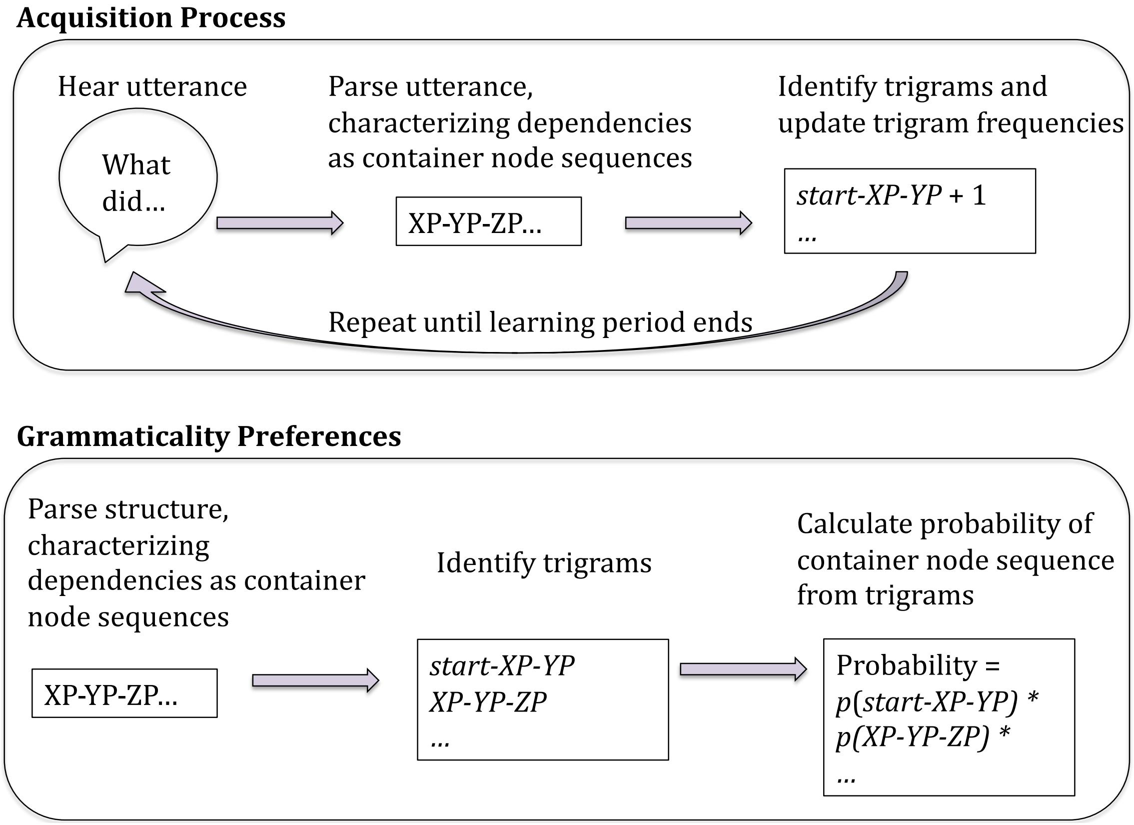
All other things being equal, this automatically makes longer dependencies less probable than shorter dependencies since more probabilities are multiplied together for longer dependencies, and those probabilities are always less than 1. Note, however, that the frequency of the individual trigrams comprising 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

---

[5] We note that the learner 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.

dependency length. The learning algorithm and calculation of grammaticality preferences are schematized in figure 2, and two examples of grammaticality preferences are shown in (9) and (10).

Figure 2. Steps in the acquisition process and calculation of grammaticality preferences.



(9) “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

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)

(10) *"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

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 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.[6] 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 (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). This is less than the number of vocabulary items children know by the time they would be learning grammaticality preferences about dependency structures, and so 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 thus very important for this learning algorithm to be psychologically plausible. We suggest that learners may adopt an

---

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

initial strategy of using the basic-level phrasal constituents noted above (derived from parsing), which is minimally taxing memory-wise. Later, they may back off to finer-grained distinctions, such as noting the complementizer used for a CP (e.g., *that*, *whether, if,* null, etc.) and making a container node out of the CP and its identifying complementizer (e.g., $\mathrm{CP}_{\mathrm{that}}$ vs. $\mathrm{CP}_{\mathrm{whether}}$ vs. $\mathrm{CP}_{\mathrm{if}}$ vs. $\mathrm{CP}_{\mathrm{null}}$, 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 "backing-off" 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*.

Given this learning algorithm, a child can generate a grammaticality preference for a given dependency at any point during learning, based on the input observed already. In particular, the child will gauge a dependency's grammaticality by calculating its probability, as derived from the frequency of the trigrams that comprise the grammaticality (see figure 1). A relative grammaticality judgment 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 island structures vs. non-island structures. The 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 acquisition process along the relevant dimensions for the UG debate: domain-specific vs. domain-general, and innate vs. derived. Note that none of the learning biases (or their components) appear to be both necessarily innate and domain-specific simultaneously, and therefore none of these biases (or their components) appear to be part of a nativist/UG-based approach to the acquisition of island constraints. In other words, the learning model that we have constructed here is not based on any Universal Grammar assumptions.

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 into a phrase structure tree	*			*
Extract container nodes	*			*
Identify trigrams		*	*	
Update probability of each trigram		*	*	
Calculate probability of utterance's dependency		*	*	

5. Learning about islands from realistic input

We turn now to specific case studies of learning preferences about structural dependencies. First, we consider the input to our modeled learners. 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. Table 3 and Table 4 describe the composition of three corpora: child-directed speech from the Brown-Adam corpus (Brown 1973) 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.

	Brown-Adam	Switchboard	Brown
total # utterances	26141	74576	24243
total *wh*-dependencies	4795	8508	4230

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

Sequence and Example Utterance	Brown-Adam	Switchboard	Brown
IP	11.7%	17.2%	33.0%
Who saw it?	560	1464	1396
IP-VP	78.3%	73.0%	63.3%
What did she see?	3755	6215	2677
IP-VP-AdjP-IP-VP	0.0%	<0.1%	0.1%
What are you willing to see?	0	1	5
IP-VP-AdjP-IP-VP-PP	0.0%	<0.1%	0.0%
What are you willing to go to?	0	1	0
IP-VP-AdjP-PP	0.0%	<0.1%	<0.1%

What are they good for?	0	1	1
IP-VP-CP$_{\text{null}}$-IP	0.1%	0.6%	0.3%
Who did he think stole it?	7	52	12
IP-VP-CP$_{\text{null}}$-IP-VP	2.2%	0.4%	0.2%
What did he think she stole?	107	30	8
IP-VP-CP$_{\text{null}}$-IP-VP-IP-VP	<0.1%	<0.1%	0.0%
What did he think she wanted to steal?	3	3	0
IP-VP-CP$_{\text{null}}$-IP-VP-IP-VP-IP-VP-PP	0.0%	<0.1%	0.0%
Who did he think she wanted to pretend to steal from?	0	1	0
IP-VP-CP$_{\text{null}}$-IP-VP-PP	0.4%	<0.1%	<0.1%
What did he think she wanted it for?	19	5	1
IP-VP-CP$_{\text{that}}$-IP-VP	0.0%	<0.1%	<0.1%
What did he think that she stole?	0	5	2
IP-VP-CP$_{\text{that}}$-IP-VP-IP-VP	0.0%	<0.1%	0.0%
What did he think that she wanted to steal?	0	1	0
IP-VP-CP$_{\text{that}}$-IP-VP-PP	0.0%	<0.1%	0.0%
Who did he think that she wanted to steal from?	0	1	0
IP-VP-IP	0.0%	<0.1%	0.0%
Who did he want to steal the necklace?	0	2	0
IP-VP-IP-VP	3.2%	3.4%	1.3%
What did he want her to steal?	155	287	57
IP-VP-IP-VP-IP-VP	0.0%	<0.1%	<0.1%
What did he want her to pretend to steal?	0	6	1
IP-VP-IP-VP-IP-VP-PP	0.0%	<0.1%	0.0%
Who did he want her to pretend to steal from?	0	6	0
IP-VP-IP-VP-NP	<0.1%	0.0%	0.0%
What did he want to say about it?	3	0	0
IP-VP-IP-VP-NP-IP-VP	0.0%	0.0%	<0.1%
What did he have to give her the opportunity to steal?	0	0	1
IP-VP-IP-VP-NP-PP	0.0%	<0.1%	0.0%
What did she want to steal more of?	0	1	0
IP-VP-IP-VP-PP	0.2%	0.4%	<0.1%
What did she want to steal from?	10	33	4
IP-VP-IP-VP-PP-PP	0.0%	0.0%	<0.1%
What did she want to get out from under?	0	0	1

IP-VP-NP	0.4%	0.1%	0.1%
What did she say about the necklace?	19	10	5
IP-VP-NP-IP-VP	0.0%	<0.1%	<0.1%
What did he give her the opportunity to steal?	0	1	2
IP-VP-NP-PP	<0.1%	<0.1%	0.0%
What was she a member of?	2	6	0
IP-VP-PP	3.2%	4.3%	1.3%
Who did she steal from?	154	369	57
IP-VP-PP-CP_{null}-IP	0.0%	<0.1%	0.0%
What did she feel like was a very good place?	0	1	0
IP-VP-PP-IP-VP	0.0%	<0.1%	0.0%
What did she think about buying?	0	3	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-IP-VP	0.0%	<0.1%	0.0%
What is she in the habit of doing?	0	1	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 doesn't seem to differ much from adult-directed speech with respect to the proportional frequency of these two sequences (child-directed (Brown-Adam): 78.3%/11.7%, adult-directed (Switchboard): 73.0%/17.2%). Adult-directed written text tends to be biased slightly more towards main clause subject dependencies, though main clause object dependencies 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, and in fact absent in the child-directed speech sample we examined. This will become relevant when we examine the learned grammaticality preferences for dependencies involving the complemetizer *that*.

We can test our modeled learners by comparing their learned grammaticality preferences to empirical data on adult grammaticality judgments available in Sprouse et al. (2012) (see also Sprouse (*this volume*)). Recall that Sprouse et al. (2012) examined four island types, using a factorial definition of island effects for each island type. The resulting container node sequence for each type is given in (11)–(14): (a) matrix gap, non-island structure, (b) embedded gap, non-island structure, (c) matrix gap, island structure, (d) embedded gap, island structure.

(11) Complex NP islands

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

(12) Subject islands

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

(13) Whether islands

a. IP MATRIX | NON-ISLAND

b. IP-VP-CP/CP CP_{that}-IP-VP EMBEDDED | NON-ISLAND

c. IP MATRIX | ISLAND

d. *IP-VP-CP/$\text{CP}_{\text{whether}}$-IP-VP EMBEDDED | ISLAND

(14) Adjunct islands

a. IP MATRIX | NON-ISLAND

b. IP-VP-CP/CP_{that}-IP-VP EMBEDDED | NON-ISLAND

c. IP MATRIX | ISLAND

d. *IP-VP-CP/CP_{if}-IP-VP EMBEDDED | ISLAND

Recall also that the factorial definition of island effects makes the presence of an island effect visually salient: if we plot the acceptability of the four sentence types in a configuration known as an interaction plot, the presence of an island effect shows up as two non-parallel lines, which indicates a statistical interaction of the two factors in the definition (the left panel of Figure 1); the absence of an island effect shows up as two parallel lines, which indicates no interaction of the two factors in the definition (the right panel of Figure 1).

To evaluate the success of our learners, we can plot the predicted grammaticality preferences in a similar interaction plot: if the lines are non-parallel, indicating an interaction, similar to the graph in the left panel of Figure 1, then the learner has acquired island constraints; if the lines are parallel, indicating no interaction, similar to the graph in the right of Figure 1, then the learner did not acquire island constraints.

To ground 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 Brown-Adam and Brown-Eve (Brown 1973) and estimating the proportion of *wh*-dependencies to total utterances, we set the learning period to 165,000 data points. So, our learners will encounter 165,000 data points containing dependencies, drawn randomly from a distribution characterized by the corpora in table 3.

All our modeled learners will follow the learning algorithm and grammaticality preference calculation outlined in figure 2. 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 grammaticality – more probable dependencies are more grammatical, while less probable dependencies are less grammatical. Though the inferred grammaticality 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.

Because the result of a grammaticality preference calculation is often a very small number (due to multiplying many probabilities together), we will calculate the log probability. This allows for easier comparison of grammaticality judgments. All of the log probabilities are negative. The more positive numbers (i.e. closer to zero) represent "more grammatical" structures while more negative numbers (i.e., farther from zero) represent "less grammatical" structures.[7] To make a direct comparison of these log probabilities with acceptability judgments, Figure 3 plots the experimentally obtained judgments for the dependencies from Sprouse et al. (2012), while figure 4 shows that model-derived log probabilities of the dependencies, based on child-directed input and figure 5 shows the model-derived probabilities of the dependencies, based on adult-directed input. Crucially at this stage of the model evaluation, we are only assuming basic-level container node distinctions (i.e., CP rather than $\text{CP}_{\textit{that}}$, etc.). This means that all CP nodes are represented as CP, irrespective of what complementizer is used. As we will see, this assumption has ramifications for the success of the learner.

---

[7] 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 grammatical dependencies are more surprising.

Figure 3: Experimentally derived acceptability judgments for all four island types from Sprouse et al. (2012) (N=173).

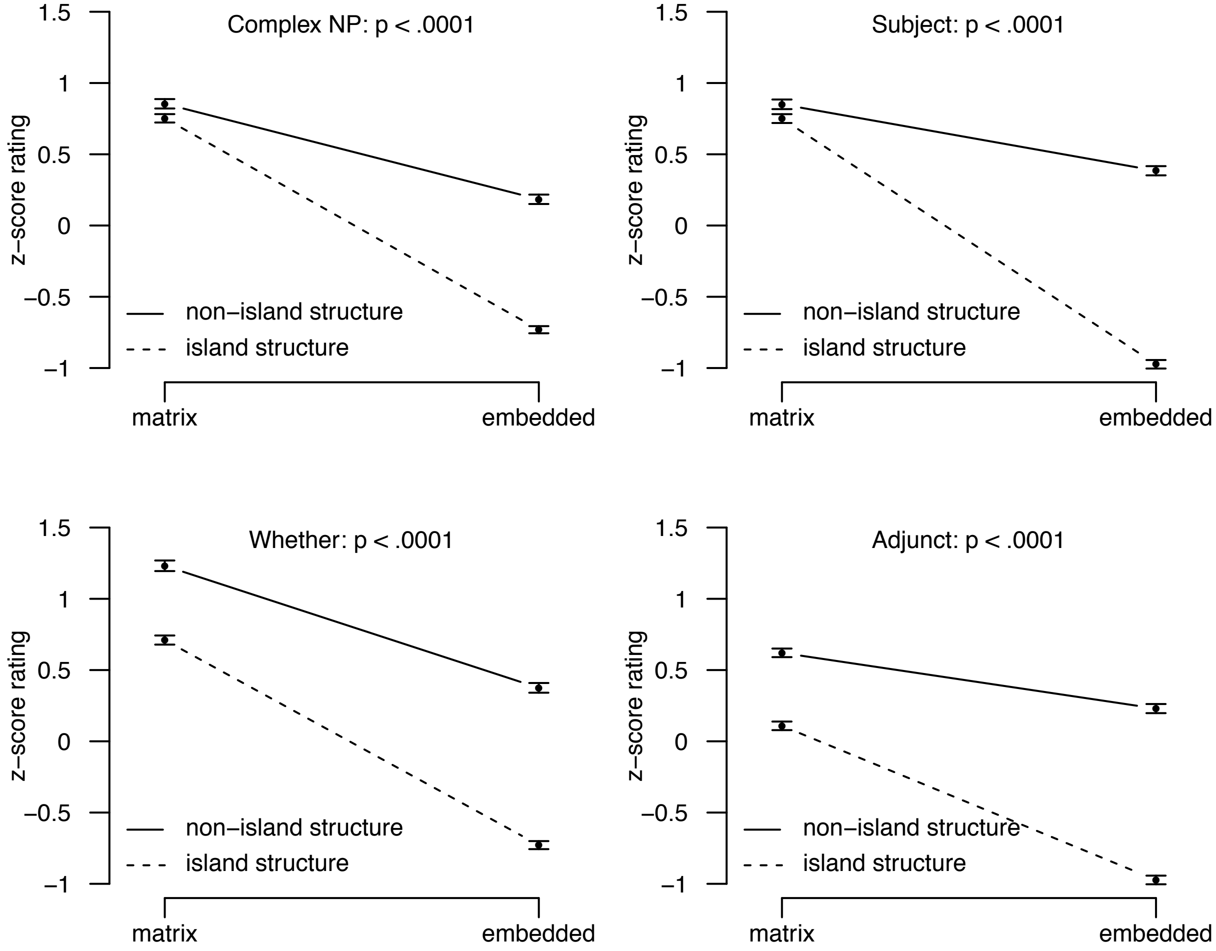


Figure 4: Log probabilities derived from child-directed speech (Brown-Adam corpus: Brown 1973) for a learner that does not discriminate CP node types. The apparent lack of dashed "island structure" line in the Whether and Adjunct island graphs indicates that the line is identical to the solid "non-island" structure line, as can be seen from the overlapping endpoints.

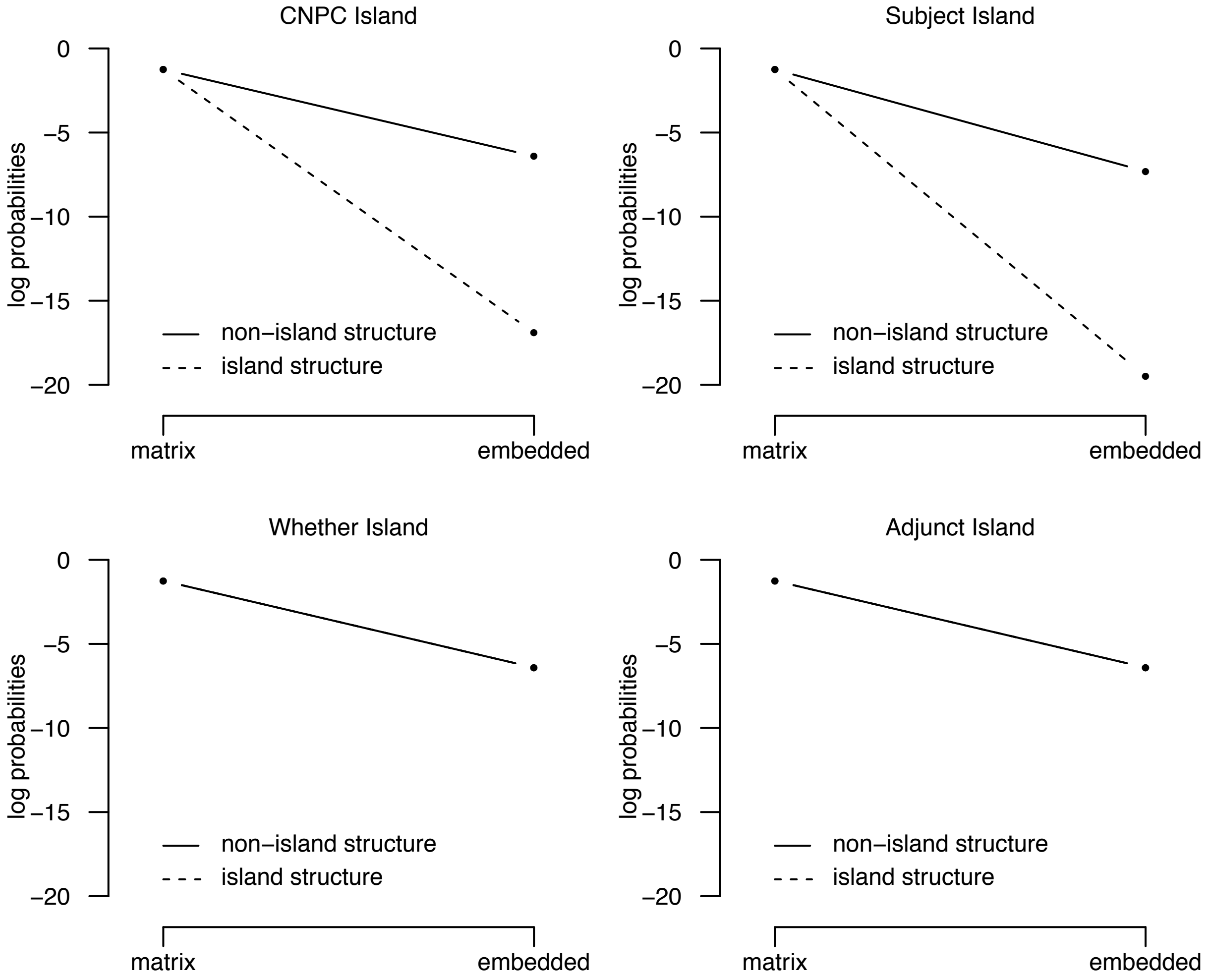


Figure 5: Log probabilities derived from adult-directed speech (Treebank-3 corpus: Marcus et al. 1999) for a learner that does not discriminate CP node types. The apparent lack of dashed "island structure" line in the Whether and Adjunct island graphs indicates that the line is identical to the solid "non-island" structure line, as can be seen from the overlapping endpoints.

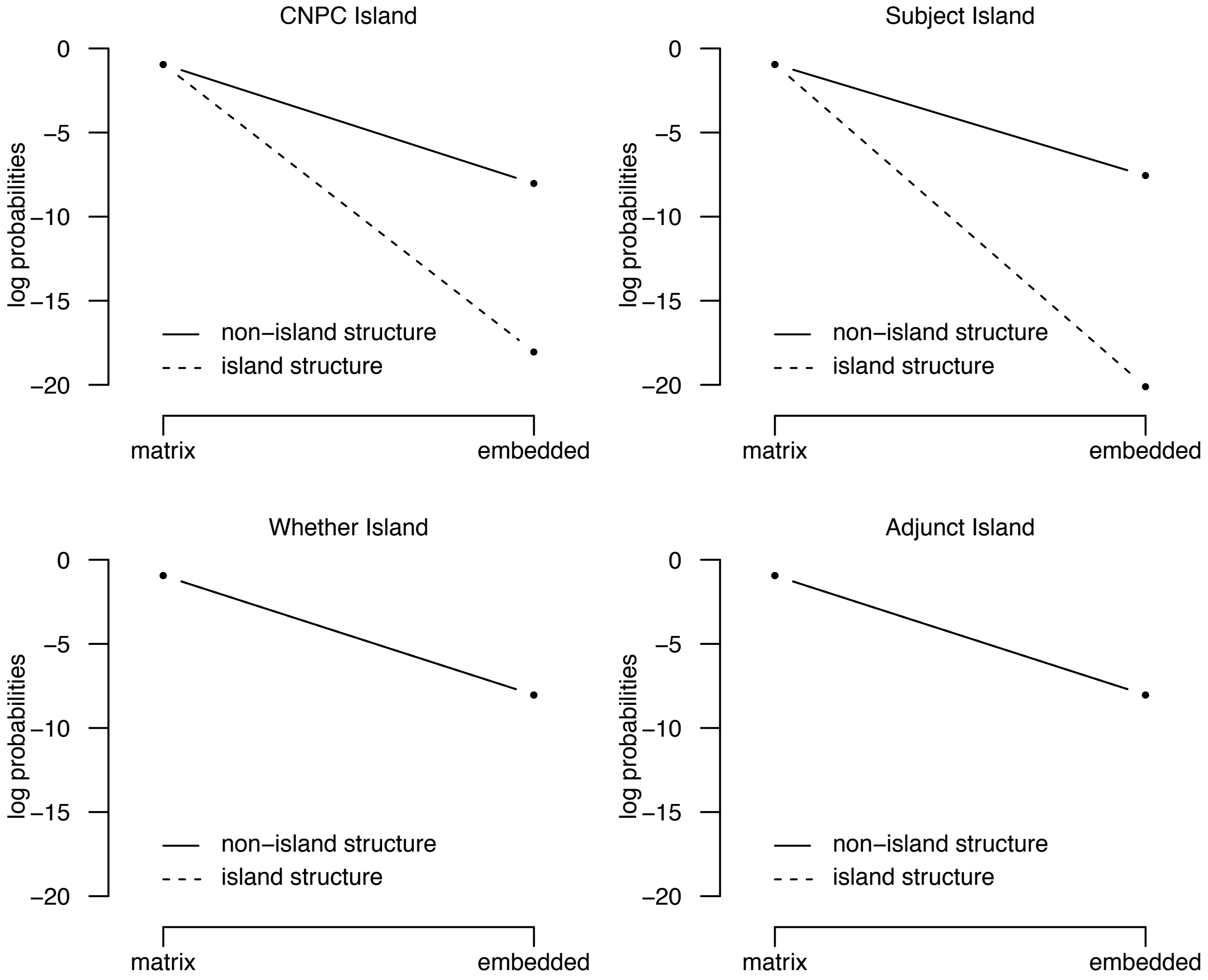


Figure 3 shows that the experimentally collected judgments for all four island types show the non-parallel lines that are indicative of an interaction, and therefore indicative of island effects. Figures 4 and 5 show that our learning model using child-directed speech (Figure 4) or adult-directed speech (Figure 5),with no distinction between CP node types, can learn the correct grammaticality preference for two of the four islands examined, CNPC and Subject islands, as both of these island types show the non-parallel lines that indicate an interaction. However, this learner fails to distinguish Whether and Adjunct islands from the control structures, as the lines are not only parallel, indicating no interaction, but also identical (resulting in graphs that appear to only contain one line). Upon closer inspection, this is not surprising because the learner does not distinguish between structures with the sequence IP-VP-CP-IP-VP, which means that Whether and Adjunct island violations, which contain specific types of CPs ($\mathrm{CP}_{\mathit{whether}}$ and $\mathrm{CP}_{\mathit{if}}$, are treated identically to grammatical utterances containing $\mathrm{CP}_{\mathit{null}}$ or $\mathrm{CP}_{\mathit{that}}$, such as "What did he think (that) she saw?".

To remedy this, we created a second learning model that allowed for finer distinctions among the CP nodes. In particular, this model 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-$\mathrm{CP}_{\mathrm{whether}}$-IP-VP and adjunct islands as IP-VP-$\mathrm{CP}_{\mathrm{adjunct}}$-IP-VP (e.g., IP-VP-$\mathrm{CP}_{\mathrm{if}}$-IP-VP). Grammatical structures will appear as IP-VP-$\mathrm{CP}_{\mathrm{null}}$-IP-VP or IP-VP-$\mathrm{CP}_{\mathrm{that}}$-IP-VP, which will allow our learner to distinguish these from the islands. Figures 6 and 7 represent the results of this kind of learner, given child-directed and adult-directed data as input respectively.

Figure 6: Log probabilities derived from child-directed speech (Brown-Adam corpus: Brown 1973) for a learner that does discriminate CP 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 by examining the overlapping line endpoints.

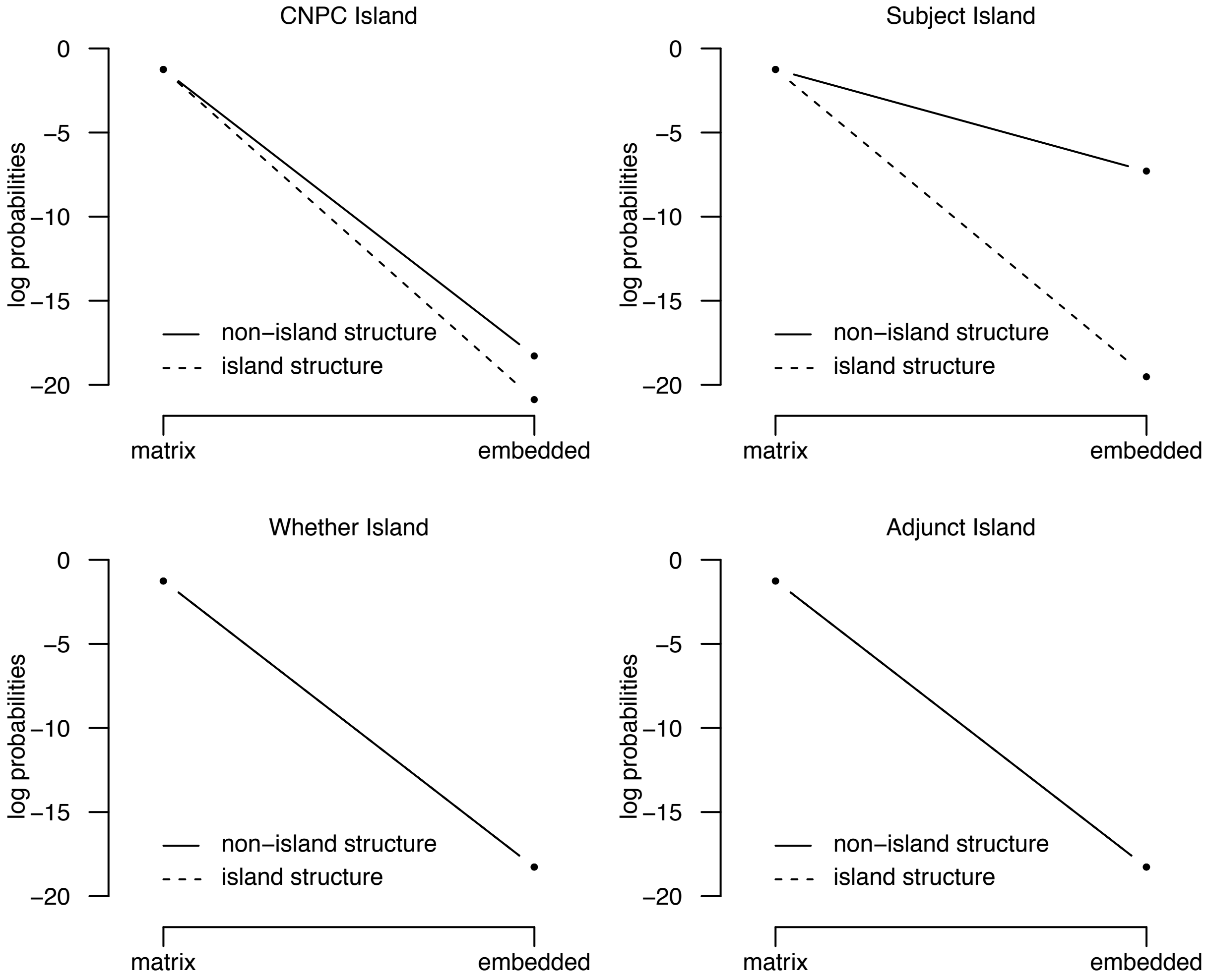
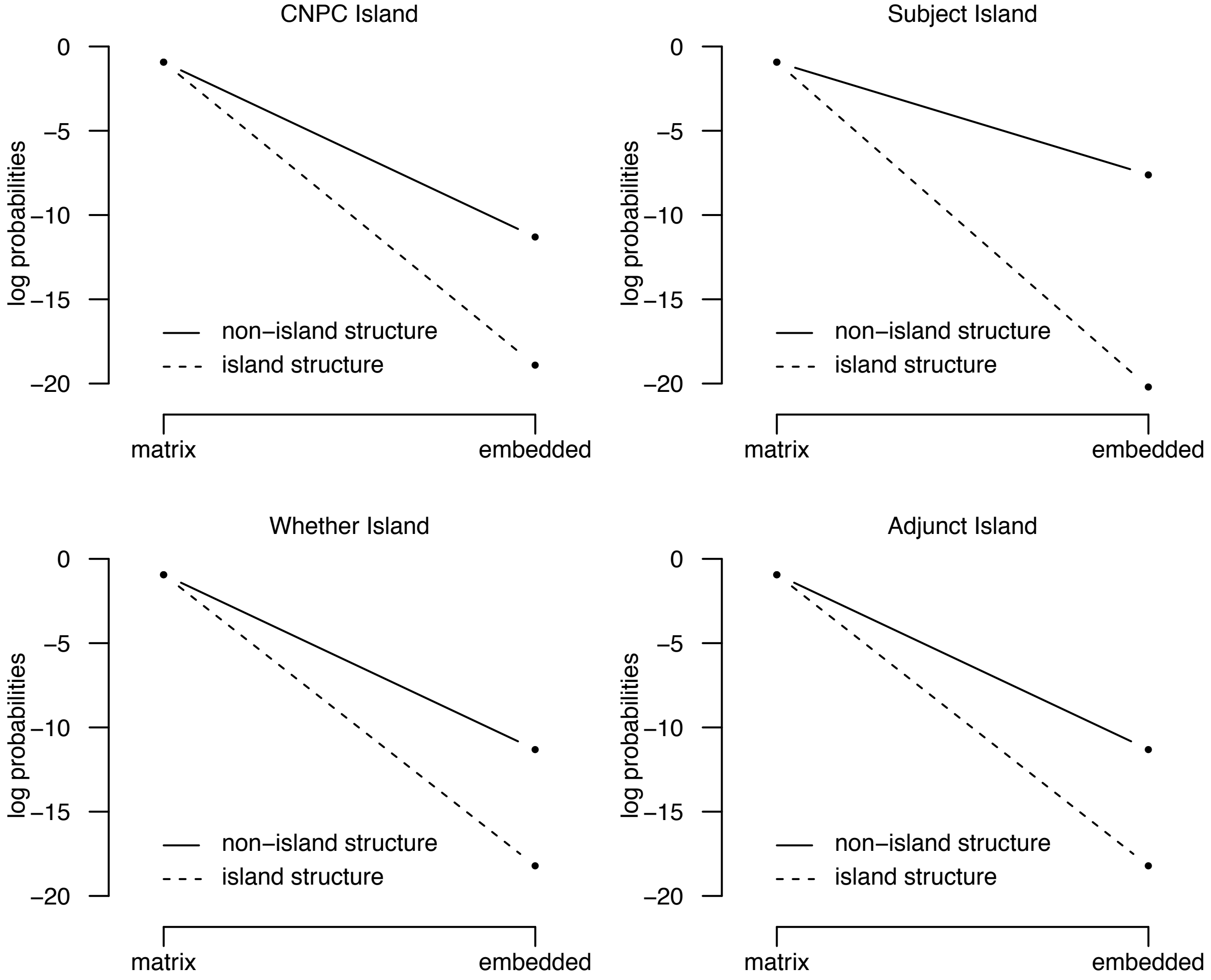


Figure 7: Log probabilities derived from adult-directed speech (Treebank3 corpus: Marcus et al. 1999) for a learner that does discriminate CP types.



Compared to our results from learners with undifferentiated CP container nodes, we see in Figure 7 that the learner using adult-directed data would end up with the correct grammaticality preferences for all four islands (compare Figure 7 to Figure 3). The learner using child-directed data, however, still fails to capture the ungrammaticality of Whether and Adjunct islands, as indicated by the parallel (and identical) lines in Figure 6.[8] This turns out to have a transparent explanation. Recall from the child-directed speech corpus (in Table 3) that there are no examples of complementizer *that* in *wh*-dependencies (as compared to the 136 examples of the null complementizer). Since there are also no examples of complementizers like *whether* and *if*, this makes the learner equate *wh*-dependencies involving *that* with equivalent *wh*-dependencies involving *whether* or *if*. This problem did not occur in the learner using adult-directed data since there were 9 examples of *wh*-dependencies with complementizer *that*. While this is still not that many, it is (crucially) not zero and so makes the adult-data learner judge wh-dependencies with *that* as more grammatical than equivalent wh-dependencies with *whether* or *if.* Thus, the adult-data learner can capture the relative ungrammaticality of Whether and Adjunct islands, while the child-data learner cannot.

It is possible that child-directed speech generally has at least some *wh*-dependencies involving the complementizer *that*, and so a learner using more balanced child-directed speech would also be able to learn that Whether and Adjunct islands are ungrammatical, as compared to equivalent structures using *that*. However, it may well be that complementizer *that* in *wh*-dependencies is less frequent in child-directed speech than in adult-directed data, so we might expect children to acquire the judgment that Whether and Adjunct islands are ungrammatical later than they acquire the judgment that CNPC and Subjects islands are ungrammatical.

---

[8] Also note that the CNPC island effect is not nearly as strong as before, though it is still present. This is because the control structure (EMBEDDED | NON-ISLAND) involved the complementizer *that*, which never appeared in the child-directed input. Thus, the control structure was also perceived as fairly ungrammatical – though notably less ungrammatical than the CNPC island structure.

To sum up, we find that a psychologically plausible learner that tracks the probabilities of certain abstract representations of wh-dependencies in the input is able to reproduce adult judgments about the (un)grammaticality of islands. In order to capture adult judgments about all four islands investigated, the learning model requires adult-directed input and a certain level of specification in the representation. However, the learning model can capture two of the four islands (CNPC and Subject) with only a basic level of detail and learning from either child-directed or adult-directed data. Notably, while this learning model assumes a particular representation of the input, it does not require innate, domain-specific knowledge about grammatical and ungrammatical structures. Instead, this representation can be derived from other properties of the learning system, which involve innate, domain-general learning abilities and derived, domain-specific representations (see Table 1).

6. Discussion & conclusion

In this chapter, we have proposed a statistical model for the acquisition of syntactic constraints on *wh*-dependencies that does not rely on innate, domain-specific knowledge of island constraints. Instead, our psychologically plausible learning model is able to implicitly derive knowledge of islands from the input using a series of relatively uncontroversial assumptions, such as the ability to parse sentences into phrase structure trees, the ability to track the nodes that contain the gap location of a *wh*-dependency, the ability to track the frequency of trigrams of container nodes, and the ability to construct a grammaticality preference for a dependency based on its trigrams. This suggests that children (and adults) do not need innate, domain-specific knowledge about islands, which in turn suggests that explicit constraints against island structures do not have to be part of Universal Grammar. In addition, we find that the learning model capable of doing this doesn't even need to involve sophisticated probabilistic inference abilities, such as Bayesian updating (e.g., Feldman et al. 2009, Foraker et al. 2009, Frank et al. 2009, Goldwater et al. 2009, Pearl et al. 2011, Perfors et al. 2011). Instead, the probabilistic learning component is fairly simple and involves tracking frequencies of particular linguistic representations that are small in size (trigrams of container nodes).

However, these results do raise interesting questions about how feasible this learner would be for the full range of constraints on *wh*-dependencies. 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 (15a and 16a), the addition of another gap outside of the island seems to eliminate the unacceptability (15b and 16b) (see Phillips, 2006 for experimentally collected acceptability judgments):

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

(16) 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 (15) would be as in (17). The sequence for both the ungrammatical gap in (15a) and the grammatical (parasitic) gap in (15b) are identical, and in fact would be as (un)acceptable as other adjunct islands, such as those using the complementizer *if*.

(17)

a. *Which book did [$_{\text{IP}}$ you [$_{\text{VP}}$ laugh [$_{\text{CP}}$ without [$_{\text{IP}}$ [$_{\text{VP}}$ reading __]]]]]?

Ungrammatical gap sequence: IP-VP-CP$_{\text{without}}$-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.

The implications of these findings for the grammar versus reductionism debate are substantial. Many of the reductionist proposals for capturing island effects without grammatical constraints have at their heart the notion that fewer grammatical constraints will lead to "simpler" grammars, and thus less motivation for innate, domain-specific learning biases (i.e., Universal Grammar). However, as we have just seen, syntactic constraints on *wh*-dependencies can be learned from input available to children without the need for innate, domain-specific biases ((i.e., Universal Grammar). Therefore there appears to be little psychological motivation to "simplify" grammatical theories above and beyond the quest for truth in science, which in this case would be the desire to accurately characterizing the grammatical system itself. We believe that this changes the nature of this debate significantly, as the question about the right characterization of island effects is no longer tied to assumptions about the nature of language acquisition, but is instead simply one question among many that must be answered to arrive at a complete understanding of the human language faculty.

Acknowledgements

This research was supported in part by National Science Foundation grant BCS-0843896 to Lisa Pearl and Jon Sprouse. We have benefited greatly from discussions with audiences at Input & Syntactic Acquisition 2009 and Input & Syntactic Acquisition 2012. All errors remain our own.

References

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

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

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

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

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

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.

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 Piatelli Palmarini, M., Uriagereka, J., and Salaburu, P. (eds.), *Of Minds and Language: The Basque Country Encounter with Noam Chomsky*, pp.256-277. New York: 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.

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.

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

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.

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*, pp. 159–166.

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

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. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*.

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*. Mahwah, NJ: Lawrence Erlbaum Associates.

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.* Philadelphia: Linguistic Data Consortium.

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

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

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. In *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.

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.

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.

Sprouse, J. *this volume.* Defining the terms of the debate: Reductionist theories and the superadditive nature of island effects. In Sprouse, J. and Hornstein, N. (eds.), *Experimental Syntax and Island Effects*. Cambridge University Press.

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

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

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

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.

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 Chan, H., Jacob, H., & Kapia, E. (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.