To appear in J. Sprouse & N. Hornstein (eds.), *Experimental syntax and island effects*. Cambridge University Press.

On the Nature of Island Constraints.
II: Language Learning and Innateness

Colin Phillips
University of Maryland

1. Islands and Learning

Island effects have long been regarded as strong motivation for domain-specific innate constraints on human language. They are obscure and abstract, and they are a parade case of a linguistic phenomenon that is likely to be difficult to observe in the input that children must learn from. As such, they have been regarded as a good example of the need for Universal Grammar.

Basic island effects are illustrated in (1-3). Long-distance 'filler-gap' dependencies are found in many different constructions, including *wh*-questions (1a), relative clauses (1b), topicalization (1c), and comparatives (1d). These dependencies can be arbitrarily long, spanning two, three, or more clauses (2), leading to the common name 'unbounded dependencies'. But there are also a number of syntactic environments where these dependencies are blocked. Filler-gap dependencies may not cross the boundary of relative clauses (3a) and other types of complex noun phrases (3b-c), interrogative clauses (3d-e), subjects (3f), adjuncts (3g), non-parallel coordinate structures (3h), factive clauses (3i) and negative clauses (3j). (In the last two examples the relevant interpretation that is excluded is the one in which the interrogative word *why* is interpreted as modifying the embedded clause.) These various environments are known as islands (Ross 1967), because one cannot escape from them.

1 a. What did the journalist accuse a man of stealing ___?
b. This is a painting that the journalist accused a man of stealing ___?
c. Those chapters, most students agree that you can safely skip ___.
d. Mary isn't as fast as [John believes she was __ five years ago]

2 a. What does Wendolene like ___?
b. What does Wallace hope that Wendolene likes ___?
c. What does Gromit think that Wallace hopes that Wendolene likes ___?

3 a. * What did Wallace meet a woman [$_{\text{rel. cl.}}$ that hates ___]?
b. *What did John read the report [that Craig won ___]?
c. * Who did Robyn believe [Simon's news about ___]?
d. *What did Sue wonder [whether Joe wrote ___]?
e. * What does Helen know [who saw ___]?
f. *What did [the fact that Ellen remembered ___] surprise her children?
g. * Who did Susan watch TV [while talking to ___ on the phone]?
h. *What did [the Senate approve ___] and [the House reject the bill]?
i. *Why did they remember that the corrupt CEO had been acquitted ___?
j. *Why did they say that nobody left ___?

Island constraints are non-obvious properties of languages whose effects are not easy to observe in the primary language input to children. In addition, they have roughly similar effects across languages: some island effects are sufficiently consistent across languages to be good candidates for universals; and those that do vary across languages appear to draw from a standard menu of options. The cross-language similarities in island effects offer some hope that children might be left with little to learn about island constraints.

This general view of the learning challenge posed by island effects has met with tacit agreement. Linguists have either agreed that island effects motivate innate domain-specific constraints, and have searched for a set of unifying principles that could explain the diversity of island phenomena (Chomsky 1964, 1973, 1986; Lasnik & Saito 1992; Manzini 1992; for reviews see Szabolcsi & den Dikken 1999; Boeckx 2008), or they have argued for reductionist accounts of island effects, which deny the existence of island constraints and thereby seek to obviate the learning problem (Pritchett 1991; Kluender & Kutas 1993; Kluender, 1998, 2005, this volume; Hofmeister & Sag 2010; Hofmeister, Staum Casasanto, & Sag this volume). In the companion paper to the current article I discussed a series of challenges for the reductionist approach (and some for the grammatical account, too). Meanwhile, there has been much less interest in the alternative possibility that island constraints are indeed real grammatical restrictions, but that they are learned from the primary input to children. Also, most claims about the difficulty of learning island constraints from the input have been based on speculation about what is in the input to children, and so any Poverty of the Stimulus arguments have been based upon educated guesswork. Set against this background, recent work by Pearl and Sprouse (2011, this volume) is particularly noteworthy, as it proposes an account of how island constraints could be learned from the input to children, and at the same time it provides a clear idea of the nature of the corpus from which children must learn.

Pearl and Sprouse present a simple distributional learning model that is able to derive rating patterns for acceptable and unacceptable *wh*-extractions, using corpora of real child-directed speech. It is a very interesting proposal, and it represents an important step in the development of distributional learning models. Despite much enthusiasm for distributional models of language learning in developmental psychology and computational psycholinguistics (Saffran, Aslin, & Newport 1996; Gomez 2002; Maye, Werker, & Gerken 2002; Vallabha, McClelland, Pons, Werker, & Amano 2006; Solan, Horn, Ruppin, & Edelman 2005), these models have attracted relatively little attention or interest in most areas of linguistics. Although it is tempting to attribute this to linguists' ignorance or stubbornness, it more likely reflects the fact that most extant distributional learning models have focused on problems that working linguists consider to be too simple to be interesting, such as learning syntactic categories (Redington, Chater, & Finch 1998; Mintz 2006) subject-auxiliary inversion (Pullum & Scholz 2002; Reali & Christiansen 2005; Perfors, Tenenbaum, & Regier 2011) or *one*-substitution (Regier & Gahl 2004; Pearl & Lidz 2009). But islands are altogether different – nobody ever claimed that islands are too simple – and so Pearl and Sprouse's argument merits close attention. It is also useful that their model is simple and transparent, and that it is clear why it performs as it does given the input corpus. The corpus data, in particular, make it relatively easy to ask about the scalability of the learning model.

In Section 2 of this article I briefly summarize the distributional learning model. Sections 3-6 discuss a series of challenges for the model. Section 3 focuses on the difference between difficulty and unacceptability. Section 4 argues that Pearl and Sprouse's corpus analysis, together with what is known about the richness of island phenomena, actually reinforces the long-standing suspicion that the input data is too impoverished for children to learn island constraints without a strong set of learning constraints (i.e., Universal Grammar). Section 5 compares how different accounts of learning island constraints approach the problem of cross-language variation. It argues that the distributional learner encounters problems that a principles and parameters learner avoids, and that there may even be some situations where the distributional learner is best served by the absence of key examples from the input. Section 6 discusses the problem of generalizing and failing to generalize across dependency types, and Section 7 concludes.

2. Distributional Learning of Island Constraints

The core of Pearl & Sprouse's learning model is impressively simple. The learner parses *wh*-dependencies in the input and keeps track of the structural path between the fronted *wh*-phrase and the gap site in each input example. It then uses probabilistic information about attested and unattested structural paths in the input corpus to assign probabilities to possible and impossible *wh*-dependencies that were not encountered in the input. To illustrate, (4a) shows the structural paths associated with a simple main clause object question and (4b) shows a more complex example of a long *wh*-dependency with preposition stranding.

4	a.	[$_{\text{CP}}$ Who did [$_{\text{IP}}$ she [$_{\text{VP}}$ like ___]]]?	*parse*
		IP VP	*XPs crossed ('container nodes')*
		IP-VP	*structural path*
		start-IP-VP; IP-VP-end	*container node trigrams (CNTs)*
	b.	[$_{\text{CP}}$ Who did [$_{\text{IP}}$ she [$_{\text{VP}}$ think [$_{\text{CP}}$ [$_{\text{IP}}$ [$_{\text{NP}}$ the gift] [$_{\text{VP}}$ was [$_{\text{PP}}$ from ___]]]]]]]?	
		IP VP CP IP VP PP	
		IP-VP-CP-IP-VP-PP	
		start-IP-VP; IP-VP-CP; VP-CP-IP; CP-IP-VP; IP-VP-PP; VP-PP-end	

Pearl and Sprouse define structural paths in terms of the maximal projections (XPs) that dominate the gap site but do not dominate the fronted *wh*-phrase. These XPs are referred to as *container nodes*, and a structural path is expressed as a sequence of container nodes. The learner's task is not merely to memorize which exact structural paths occur in the input, but to generalize to structural paths that might be possible despite not occurring in the input. This is achieved by breaking down all attested structural paths into *container node trigrams* (CNTs), and then assigning an empirical probability to each trigram. The probabilities of the individual CNTs can then be used to derive probabilities for any conceivable structural path, simply by taking the product of the probabilities of the component CNTs.

(5) shows two salient examples of structural paths that are not attested in the input. Importantly, the 3-clause direct object extraction in (5a) is unattested in the corpus but is grammatically possible, whereas the subject island violation in (5b) is unattested in the corpus and is generally regarded as illicit. The structural path in (5a) is made up of CNTs

that are all well represented in the input corpus, and therefore the model assigns a relatively high probability to the 3-clause extraction. In contrast, the CNT IP-NP-PP in (5b) is not attested in the input corpus, and so the model assigns it a very low probability.[1] By defining a threshold of acceptability at a very low probability value, the distributional learner is able to successfully classify island violations as unacceptable and long-distance extractions as acceptable. Nevertheless, due to the fact that longer paths involve more CNTs, and consequently have lower probabilities, the model is also able to capture the fact that naïve participants generally rate longer *wh*-dependencies as less acceptable than shorter *wh*-dependencies (Gibson 1998; Hawkins 1999; Phillips, Kazanina, & Abada 2005; Sprouse, Wagers & Phillips 2012).

5 a. $[_{\text{CP}}$ What did $[_{\text{IP}}$ John $[_{\text{VP}}$ say $[_{\text{CP}}$ $[_{\text{IP}}$ he $[_{\text{VP}}$ heard $[_{\text{CP}}$ that $[_{\text{IP}}$ Bob had $[_{\text{VP}}$ said ___]]]]]]]]]?
IP-VP-CP-IP-VP-CP-IP-VP

b. *$[_{\text{CP}}$ What did $[_{\text{IP}}$ $[_{\text{NP}}$ the news $[_{\text{PP}}$ about ___]] $[_{\text{VP}}$ stun Harry]]]?
IP-NP-PP

Pearl and Sprouse's learner is able to model the patterns of acceptability ratings for the four different types of island tested by Sprouse and colleagues (2012): Complex NP (relative clause) islands, *wh*-islands (*whether* islands), subject islands, and adjunct islands. For each island the pattern of acceptability judgments is based on four sentence-types organized in a 2×2 manipulation of *wh*-dependency length and the presence of an island structure, as shown in the sample paradigm for a *wh*-island in (6). The numbers after each example represent the mean normalized acceptability ratings for each sentence type, and they show that the rating for the combination of an island structure and a long-distance *wh*-dependency is somewhat lower than would be predicted by the sum of the individual costs of the island structure and the *wh*-dependency. This is the superadditive property of island effects that competing theories seek to capture.

6	a.	Who ___ thinks that John bought a car?	– island / – long-distance	0.87
	b.	What do you think that John bought ___?	– island / + long-distance	0.22
	c.	Who ___ wonders whether John bought a car?	+ island / – long-distance	0.47
	d.	What do you wonder whether John bought ___?	+ island / + long-distance	-0.91

However, in order to capture the four island types the distributional learning model has to keep track of more than the category labels in the CNTs. In particular, it must separately track paths involving different types of complementizer, e.g., CP_{that}, $\text{CP}_{\text{whether}}$, CP_{null}, CP_{if}. This step of course requires that the model know that these distinctions among complementizer types could be relevant for characterizing the set of possible CNTs, and the differentiation among complementizer types presents the danger of creating a data sparseness problem. In fact, this danger turns out to be very real, as CNTs including the overt complementizer *that* are quite rare, both in child-directed speech corpora (2/11308

[1] Pearl & Sprouse's learner uses smoothed trigram probabilities, with the consequence that unobserved trigrams are assigned a non-zero probability. This allows the model to assign a probability to any conceivable trigram.

wh-dependencies) and in adult-directed speech corpora (5/8508 *wh*-dependencies). Nevertheless, the model is able to successfully predict acceptability patterns for *wh*-questions involving overt *that*.

The corpus of child-directed speech used by the model clearly does not represent a complete corpus of the *wh*-questions that an individual learner would be exposed to over the course of a few years. But based on the estimates that Pearl and Sprouse provide, the corpus is not too far different than what a real child must use. Their corpus of 66,000 child-directed utterances contains around 11,000 *wh*-questions. Based on the claim that children typically hear around a million utterances in a 3-year span (Hart & Risley 1995), they estimate that a typical child hears around 175,000 *wh*-questions between the ages of 2 to 5 years, which they take to be the period when children master constraints on *wh*-extraction (cf. de Villiers 1995; Crain & Thornton 1998). This means that real children plausibly encounter a corpus that is only one order of magnitude larger than what Pearl and Sprouse's model uses. This suggests that any data sparseness problems that might be found in Pearl and Sprouse's corpus are likely also present in the data that a typical child would encounter. For example, if a given structural path occurs only twice in Pearl and Sprouse's corpus, then increasing this to around 20 examples over a 3-year period, i.e., once every 2 months, should not count as particularly common, especially when we consider that children might not attend to all input sentences, or might misanalyze some of them (cf. Omaki 2010).

The Pearl and Sprouse model is an important attempt to apply simple distributional learning techniques to a problem that has generally been thought to lie beyond the reach of such models. Of course, this first step only addresses part of the richness of island phenomena. In addition to the four island types that Pearl and Sprouse test, children must come to know about the effects of additional islands (e.g., factives, negative islands), the consequences of extracting different kinds of phrases (subjects, objects, adjuncts, predicates, specific and non-referential expressions), and about similarities and differences of various kinds of unbounded dependencies (e.g., relative clauses, comparatives, topicalization; fronted vs. in-situ phrases). These types of island phenomena are briefly reviewed in the companion paper to the current article. It is therefore important to assess whether the distributional learning model can scale up to handle a wider range of phenomena.

3. Difficulty and Unacceptability are Different – Qualitatively

The Pearl and Sprouse model uses probabilities as surrogates for scalar ratings in acceptability judgment studies. This allows the model to show that long-distance *wh*-dependencies and island violations both affect probability estimates, but that they do so to different degrees, just as they affect human acceptability ratings to different degrees. But in expressing effects of dependency length and islands as shifts along a single probability dimension, the distributional learning model makes a stronger claim than does an experimenter who gathers judgments using a single rating scale. The experimenter who asks participants to rate difficult and illicit sentences using a single scale is providing raters with a simple task, but makes no commitment to the notion that difficulty and well-formedness do, in fact, correspond to a single underlying cognitive dimension. In contrast, the distributional learning model treats difficulty and ill-formedness as if they are really the same thing. I think that this conflation is unjustified, and it may obscure the most interesting finding of the model, but I also think that it can be easily fixed.

The input corpora for the distributional learning model are overwhelmingly dominated by simple *wh*-dependencies. In child- and adult-directed speech corpora around 90% of *wh*-dependencies involve single clause extractions; in the corpus of written text these simple *wh*-dependencies account for 96.3% of the corpus.[2] For this reason, it is no surprise that the model assigns a low probability to longer *wh*-dependencies. In fact, probabilities fall sharply as *wh*-dependencies grow in length. In the probabilities derived from child-directed speech a simple subject *wh*-question (e.g., *Who __ read the book?*) has a log probability of –1.26, and a bi-clausal object question with an overt complementizer (e.g., *What do you think that John read __?*) has a log probability of –13.06. On a log scale that is used as a surrogate for acceptability ratings this is a strikingly large drop in probability from adding only one clause. In fact, the low probability assigned to the bi-clausal queestion is due in large part to the overt complementizer *that*, which is very rare in the input corpus. Nevertheless, the log probability for the acceptable long-distance question is still much higher than the log probability assigned to island violations (range: -18 to -20), and so the model may be regarded as successful.

But the success in distinguishing long *wh*-dependencies from island violations might not generalize very far. Well-formed *wh*-dependencies can be arbitrarily long, spanning 2, 3, 4 or more clauses (7). Although adding more clauses makes it increasingly difficult to keep track of the full interpretation of the sentence, speakers have no difficulty parsing the syntax of the sentence and recognizing that it is well formed.

7 a. What will Priscilla read __?
b. What does Bill hope that Priscilla will read __?
c. What did Kathy say that Bill hopes that Priscilla will read __?
d. What does Robin expect Kathy to say that Bill hopes that Priscilla will read __?

In light of the low log probability assigned to a simple 2-clause extraction, it is likely that adding one or two more clauses to an object *wh*-question, as in (7c-d) would bring the log probability into the range of an island violation like (8). And adding more clauses to the *wh*-dependency would likely yield an even lower log probability than the island violation. This is an unwelcome result, as speakers easily recognize that long-distance object extraction has a different status than the island violation in (8). In effect, the distributional learner succeeds in distinguishing long-but-acceptable *wh*-dependencies from island violations only because it considers a limited range of dependency lengths.

8 *Who does Jack think [$_{CP}$ [$_{NP}$ the necklace for __] is expensive]?

There may be a straightforward numerical fix to the problem that long-but-acceptable *wh*-dependencies receive lower log probabilities than island violations. Island violations

[2] These figures appear to combine all *wh*-question types, including argument questions with *who*, *what*, and *which N*, and adjunct questions with *when*, *how*, and *why*. Based on figures in Zukowski & Larsen (2011) we can estimate that adjunct questions make up a substantial portion (20% or more) or the simple questions in the corpus.

include unattested CNTs, but they are assigned non-zero probabilities because of the smoothing that is applied to all probabilities derived from the corpus. The smoothing function could presumably be adjusted to keep the log probability of island violations below that of multi-clause *wh*-dependencies. But a fix of that nature avoids the more important point that there is a fundamental difference between long-distance dependencies and island violations. Multi-clause extractions can certainly be hard to interpret, but they are easy to parse and recognize as well-formed. They do not fail in the way that island violations do. This contrast can easily be detected intuitively, and it can also be seen in experimental findings on active filler-gap dependency processing. Active *wh*-dependency formation effects persist in longer *wh*-dependencies (Phillips et al. 2005; Wagers & Phillips 2009), but they disappear in island environments (Stowe 1986; Traxler & Pickering 1996; Omaki & Schulz 2011; Omaki et al. submitted). Active maintenance of the semantic features of a *wh*-phrase appears to decline rapidly as distance increases, but information about the syntactic category status of the *wh*-phrase persists (Wagers & Phillips submitted). Also, in their study of the relation between island effects and working memory capacity (WMC), Sprouse and colleagues found that there was a correlation between WMC and the dependency length effect, although there was no correlation between WMC and island effects. All of these observations indicate that the effects of longer *wh*-dependencies and island violations are qualitatively different.

An alternative way of distinguishing longer *wh*-dependencies and island violations, while retaining the general approach that Pearl and Sprouse adopt, could be to distinguish the types of probabilities that are assigned to each *wh*-dependency. It is possible that the probabilities that the model assigns to each *wh*-dependency, based on the product of CNT probabilities, are a reasonable approximation of the interpretability of the dependency. But the acceptability difference between long dependencies and island violations may be better captured by the probability of the least probable CNT in each dependency type. We could call this the *minimal CNT probability* (minCNT) for a *wh*-dependency. Island violations contain at least one CNT that has a very low probability because it is unattested in the input corpus. In contrast, long-but-acceptable *wh*-dependencies consist of many instances of CNTs that are widely attested in the input corpus. The minCNT measure would treat long-ish and very long *wh*-dependencies as in (7b-d) as equivalent, while classifying all of them as more acceptable than the island violation in (8).

Summarizing, Pearl and Sprouse's argument highlights their model's ability to distinguish long *wh*-dependencies from illicit *wh*-dependencies, matching empirical findings from rating studies. This is elegant, but it is perhaps a distraction from the more important finding of their study, which is that they can reliably distinguish acceptable and unacceptable structural paths using realistic input corpora. The key question, then, is whether this feature of the model can scale up to a wider range of island effects.

4. The Data Sparseness Problem is Real

4.1 Pearl & Sprouse's evidence against sparseness

The most important result of Pearl and Sprouse's study may be the simple finding that in corpora of natural speech illicit *wh*-dependencies fail to occur and licit *wh*-dependencies do occur. This is a finding about the informativeness of the input corpus. The second important finding is that the distinction between licit and illicit *wh*-dependencies can be captured rather well in terms of local sequences of nodes in the path between a *wh*-phrase and the gap site (CNTs in Pearl & Sprouse's model). This second finding is closely related to the fact that grammatical theories of islands typically capture generalizations about islands in terms of highly local properties of *wh*-dependencies (e.g., no extraction across a node of category $\mathrm{CP}_{+\mathrm{wh}}$; no extraction across a non-complement node). The combination of the informativeness property and the locality property is what makes it possible to use simple trigrams of nodes to make predictions about possible and impossible *wh*-dependencies that are not attested in the input. These two properties also are the reason why the model is able to succeed using a very simple distributional model. In fact, the details of the learning model probably do not matter very much. The locality property makes it easier to generalize beyond the input corpus, and the informativeness property makes the generalization succeed. These properties could be exploited by a variety of different learning models.

Moreover, the model succeeds despite working with a relatively small corpus of only around 11,000 *wh*-questions. A real child would need to learn from a corpus of *wh*-questions that is probably only around one order of magnitude larger in a 3-year period (see above). For this reason, Pearl and Sprouse's model is an important test-of-concept for the feasibility of learning from a realistic corpus. One reason for traditional skepticism about the learnability of island constraints involves the presumed sparseness of relevant data in the input. But perhaps the data sparseness problem is not so severe. We should therefore ask whether the corpus is sufficient to derive more of what native speakers know about islands, while preserving the locality and informativeness properties and avoiding problems of data sparseness.

Pearl and Sprouse point out that the non-sparseness of the input data for their model is primarily a function of the relation between (i) the number of CNTs that the learner must track, and (ii) the number of relevant examples in the input corpus. The number of CNTs that must be tracked can be estimated based on the number of container nodes that the model distinguishes (9a) and the number of container nodes in the path fragments that the model tracks (9b). The number of possible container node trigrams (729) is small relative to the number of *wh*-questions in the corpus (more than 11,000).

9 a. Container node (9): IP, VP, NP, PP, AdjP, $\mathrm{CP}_{\mathrm{null}}$, $\mathrm{CP}_{\mathrm{that}}$, $\mathrm{CP}_{\mathrm{if}}$, $\mathrm{CP}_{\mathrm{whether}}$
b. n-grams (n=3): XP_1, XP_2, XP_3
c. total n-grams (9^3): 729
d. corpus size: 11,308 *wh*-questions (child-directed corpus)

These calculations are encouraging at first sight, as they show why the distributional learner is able to recognize non-occurring CNTs even with a corpus that is somewhat smaller than the corpus that real children learn from. But even relatively modest changes to the model could make a big difference to this estimate. If the list of container nodes had to be expanded from 9 to 15 categories, and if the n-grams expanded from length 3 to length 4 nodes, then the total number of n-grams to be tracked would increase from 9^3 (729) to 15^4 (50,625), a figure that is rather large, even relative to the 175,000 *wh*-questions that Pearl and Sprouse estimate that children encounter in a 3-year period.

This makes it all the more important to assess how much detail would need to be tracked in order to ensure success on the full range of island phenomena.

4.2 Too many categories

Pearl and Sprouse show that in order to distinguish licit long-distance *wh*-dependencies from adjunct islands and *wh*-islands the distributional learner must distinguish at least 4 types of CPs in its CNTs (see (9a) above). There are good reasons to think that the learner may need to keep track of a number of additional categories.

One important feature that should be included is the distinction between CPs that are complements vs. modifiers of nouns. This is needed for distinguishing acceptable extractions from NP-complements from unacceptable extractions from relative clauses, as illustrated in (10).

10 a. Which city did the agent read the news [$_{\text{CP-comp}}$ that terrorists had attacked __]?
b. *Which city did the agent suspect the terrorists [$_{\text{CP-RC}}$ that __ had attacked __]?

The contrast in (10) could be captured by adding one more option to the list of CP types that the learning model must track. This is a small step, except that it requires the learner to separate CP container nodes based on an abstract property of the parse that is not reflected in the form of the complementizer itself (it is the word *that* in both cases). This may be a relatively innocent amendment to the model, in the respect that the learner already must learn from accurate abstract parses of the input sentences. But the more important question is how the learner knows that this property of CPs is one that it should keep track of, and how it distinguishes this property from many other properties of attested *wh*-dependencies that it might choose to keep track of. An attraction of the learning model that Pearl and Sprouse present is that it keeps track of relatively simple properties of structural paths, primarily category labels, and therefore it can plausibly be regarded as lacking detailed innate knowledge of *wh*-dependencies. If, instead, the learner needs to be equipped with a more detailed list of the syntactic properties that are relevant for constraints on *wh*-dependencies, then it becomes harder to maintain that the model is learning island constraints without innate domain-specific knowledge. A distributional learner with detailed knowledge of which features to keep track of could be an important component of a parametric/UG-constrained learner, and it would not be uninteresting. But it would not lack innate knowledge.

The feature that is needed for distinguishing complement CPs from relative clauses may be the tip of the iceberg, as many other features turn out to be important for distinguishing licit and illicit *wh*-dependencies. These include features of the structural path between the *wh*-phrase and the gap, such as (i) factivity (Kiparsky & Kiparsky 1971; Adams 1985; Rooryck 1992), (ii) tense (Huang 1982; Lasnik & Saito 1992), (iii) bridge vs. non-bridge verbs (Cattell 1978; Erteshik-Shir 1973) , (iv) operators such as negation and quantifiers (Rizzi 1990), and (v) definiteness (Fiengo & Higginbotham 1981; Postal 1998). In addition to features of the structural path of extraction, it is also important to distinguish the properties of the *wh*-phrase that undergoes extraction. Extraction possibilities differ based on (i) argument vs. adjunct *wh*-words, (ii) the referential specificity of the *wh*-phrase, e.g., *what* vs. *which*-N, and even (iii) the interpretation of the *wh*-phrase. For example, the *how* *many* question in (11a) is ambiguous, as shown by the paraphrases. But (11b), which replaces the verb *say* with the verb *forget*, which takes a factive complement, only allows one of the two readings (Rizzi 1990). This is just one among many properties that a successful distributional learner would need to track in order to derive island constraints from the input. In examples like (11) this may be particularly difficult, since the learner would need to accurately track the intended meaning of the ambiguous question, and could easily make errors.

11 a. How many books did you say that John read __? *Ambiguous*

How many books have the property that you say that John read them?
What is the number such that you say that John read that number of books?

b. How many books did you forget that John read __? *Unambiguous*

How many books have the property that you forgot that John read them?
** What is the number such that you forgot that John read that number of books?*

Taken together, these observations suggest that (i) many syntactic/semantic features are relevant to the characterization of island effects, beyond those that Pearl and Sprouse use in their model; and that (ii) the list of features to be tracked grows much larger if the goal is to have a learner that derives the set of relevant features, without the guiding hand of Universal Grammar.

4.3 Too little data

A feature that Pearl & Sprouse's learner already keeps track of via CNTs is the contrast between CPs with overt *that* vs. a null complementizer. This has advantages and disadvantages. The difference does matter in one specific environment, when the *wh*-phrase is extracted from the subject position immediately following overt *that*, i.e., the *complementizer-trace* constraint, as illustrated in (12). For this reason alone it is important that a distributional learner should track the status of the complementizer. However, in almost all other cases the difference between null and overt complementizers does not make a difference in English. As a result, most of the distributional information that the learner accumulates about null vs. overt *that* is unhelpful. Moreover, Pearl and Sprouse's frequency counts suggest that overt *that* is extremely rare in questions in the input – just 2 tokens out of the 11,000 *wh*-dependencies in the child-directed corpus – and so it is far from certain that the learner would even encounter sufficiently many cases to learn the complementizer-trace constraint illustrated in (12). Long-distance object questions with a null complementizer (12b) are roughly 80 times more frequent than object questions with an overt complementizer (12a). Long-distance subject questions (12d) are overall much less common than long-distance object questions. Therefore, the absence of subject questions with an overt complementizer (12c) likely would not stand out as a gap in the paradigm, even if the size of the corpus was multiplied 10-fold, yielding roughly the amount of input that a child might expect to encounter in a 3-year period. It seems, then, that the input corpus cannot be relied upon to provide sufficient input for children to learn about constraints on extraction – the data is too sparse.[3]

12 a. Who do you think that John met __? 2 / 11308 (child-directed corpus)
b. Who do you think John met __? 159 / 11308
c. * Who do you think that __ left? 0 / 11308
d. Who do you think __ left? 13 / 11308

So, not only is it difficult to guarantee that the learner would master the complementizer-trace constraint, but there is a danger that the rarity of the complementizer *that* might lead the learner to incorrectly conclude that the presence of the overt complementizer is grammatically excluded in environments where it is entirely fine.

Importantly, complementizer-trace effects do not rely on exotic or subtle linguistic features. They require the learner merely to track distinctions like subject vs. non-subject, and overt vs. null complementizer. Therefore we might expect it to be relatively easy for a distributional learner to recognize the absence of examples like (12c) in the input corpus. But if the corpus data is too sparse to ensure the learning of even this constraint, then there is little reason to be optimistic about the learning of distinctions that rely on more subtle features. This is unlikely to be a situation where we can hope that the problem can be solved by finding a more sophisticated mathematical model. If the relevant data is not well represented in the input corpus, then no amount of statistical magic can compensate for that.

One interesting – and to me surprising – feature of the three input corpora that Pearl and Sprouse analyze is that the *wh*-questions in the corpora are 100% grammatical. There are no errors in around 24,000 *wh*-dependencies. This perfection in the input data could be very useful from the learner's perspective, in the respect that it could allow the learner to take every input sentence as a reliable piece of evidence about the target language, and hence could successfully learn from CNTs that are quite rare, such as the CP$_{\text{that}}$-IP-VP sequence that is crucial for recognizing the acceptability of (12a), which occurs only twice in the child-directed corpus. If there were just a small amount of noise in the input corpus, then the learning model would struggle to take advantage of such rare cases. But even if the input corpus for *wh*-dependencies really is as perfectly grammatical as Pearl and Sprouse find in their sample, we should be cautious about a learner that proceeds under the assumption that all input sentences are grammatical. First, we know that the child's input corpus as a whole is not error free. The frequent agreement errors in natural speech are just one example of this (Eberhard, Cutting, & Bock 2005). So how could the learner know that the input corpus is more reliable for *wh*-dependencies than for some other types of dependencies? The learner presumably should be ready for some small level of error in all phenomena in the input. Second, it is probably not safe to assume that the child is able to

[3] Interestingly, the corresponing counts from the two adult-directed corpora that Pearl and Sprouse analyzed yield a slightly different picture. The counts corresponding to the examples in (12a-d) in the adult-directed speech corpus are 5, 30, 0, 52, from a total of 8508 *wh*-dependencies; in the adult-directed text corpus they are 2, 8, 0, 12, from a total of 4230 *wh*-dependencies. But even in the most 'helpful' corpus, the adult-directed speech corpus, we can estimate that the crucial object questions with overt *that* occur with sufficient frequency for a child to hear one roughly once every ten days. The child-directed speech corpus is far less helpful. So much for the virtues of child-directed speech.

perfectly encode the tens or hundreds of thousands of *wh*-dependencies in the input, given the child's limited language processing capacity, and hence the child may add noise to the corpus.

4.4 The limits of trigrams

As discussed in Section 4.1, the success of the distributional learner depends on the relation between the number of path fragments that the model must keep track of and the size of the input corpus. If the learner has a large corpus and relatively few path fragments to tally, then it is more likely that the learner will be able to reliably detect significant gaps in the corpus. (This is, of course, no guarantee that unacceptable structures stand out as significant gaps in the corpus, as discussed in Section 4.3.) The number of path fragments that the learner must track is a function of the number of categories that are distinguished in the *n*-grams, and the length of the *n*-grams. Pearl and Sprouse argue that *n*-grams of length 3, i.e., their CNTs are sufficient for describing constraints on *wh*-dependencies. But we should ask whether longer sequences are required?[4] The answer to this question is unlikely to depend on details of the probabilistic model, because we have already seen that the model's success depends on distinguishing CNTs that are acceptable but occur very rarely in the corpus from CNTs that are entirely unattested in the corpus. In effect, then, the question about the adequacy of trigrams reduces to the question of whether all cases of illicit *wh*-dependencies involve at least one illicit/unattested CNT. In other words, are all island violations attributable to the crossing of a specific local illicit chunk of structure?

The answer seems to be that most island violations can indeed be attributed to illicit local chunks of structure, but with a couple of notable exceptions.

One case where an island effect might not be reducible to an illicit local chunk of structure involves a contrast between English and Romance languages in the status of *wh*-islands. This contrast received much attention in early discussions of parametric syntax, and is still frequently cited in surveys of island phenomena. English generally disallows extraction of one *wh*-phrase across another, as shown in (13). But it is reported that Romance languages such as Italian (Rizzi 1982), Spanish (Torrego 1984), and French (Sportiche 1981) are more liberal, allowing escape from *wh*-islands, as illustrated in (14). In other respects, these languages appear to respect island constraints in a very similar fashion to English.

13 * What$_i$ do [$_{IP}$ you wonder [$_{CP}$ who$_j$ [$_{IP}$ __$_j$ likes __$_i$]?

14 Tuo fratello, [$_{CP}$ a cui$_i$ mi domando [$_{CP}$ che storie$_j$ abbiano raccontato __$_i$ __$_j$],
your brother, to whom I wonder which stories they-have told,
era molto preoccupato
was very worried

[4] The reader should not worry about the possible consequences of introducing more elaborate phrase structure representations that go beyond the traditional labels CP, IP, VP, NP, PP used in Pearl and Sprouse's model. Many additional maximal projections that are in vogue in contemporary syntax (Cinque 1999), but it should be straightforward to recapture the success of Pearl and Sprouse's model in more articulated syntactic representations, by tracking only the major category boundaries that are counterparts of the traditional maximal projections.

An influential early proposal by Rizzi (1982) was that the contrast in (13-14) can be captured by parameterizing Chomsky's classic *Subjacency Constraint* (Chomsky 1973). Chomsky had proposed that a number of different island effects in English could be explained by a constraint that banned *wh*-movement operations that crossed more than one *bounding node*, where the bounding nodes corresponded to NP and S (= IP). Under this account, the *wh*-island violation in (13) is ruled out because the *wh*-phrase *what* crosses two S/IP nodes in its movement path. Rizzi proposed that the acceptabilty of (14) could be explained by assuming that in Italian the bounding nodes are instead NP and S' (= CP), with the consequence that the fronted *wh*-phrase *a cui* in (14) crosses only one bounding node. A striking further prediction of this account is that extractions across two *wh*-phrases in Italian, as in (15), should be just as bad as their English counterparts. Rizzi reported that this prediction is correct.

15 * Questo argomento, [$_{CP}$ di cui$_{k}$ mi sto domandando [$_{CP}$ a chi potrei chiedere
this topic of which I am wondering to whom I-may ask
[$_{CP}$ quando dovro parlare __$_{k}$]]], mi sembra sempre piu complicato
when I'll-have-to speak to-me seems ever more complicated
'This topic, which I am wondering who I can ask
when I'll have to talk about, seems more and more complicated to me.'

If the Italian contrast between single and double *wh*-islands is accurate as described, then it is something that would be difficult to capture in terms of illicit local structural chunks, i.e., it would be a challenge for CNTs. If the learner received sufficient input to learn that a *wh*-dependency may cross a CP$_{+wh}$ category, then the model would likely predict that a *wh*-dependency may also cross two CP$_{+wh}$ categories, contrary to fact. The challenge lies in the fact that no individual part of the *wh*-dependency path in (15) is illicit by itself. Rather it is the combination of two pieces of the dependency, which are not local to one another, that is fatal.

However, it is not clear at present how serious is the challenge that cases like (15) present. First, the Italian facts are contested (Manzini 1992), and it has been claimed that English is not so different from Italian (Grimshaw 1986). Second, despite the notoriety of this early example of parametric variation, there appear to be few other cases of extractions that are ruled out by the combination of path-components that are each independently licit. As a result, the challenge for CNTs posed by examples like (15) remains uncertain.

A second challenge for CNTs involves another type of conspiracy between different parts of a *wh*-dependency, and it is a challenge that Pearl and Sprouse acknowledge. Parasitic gaps are a variety of multiple-gap construction in which one gap is inside a syntactic island and a second gap is in a perfectly licit position (Engdahl 1983; Culicover & Postal 2001). Examples of parasitic gaps inside a subject island (16b) and an adjunct island (17b) are shown below, together with matching examples of subject and adjunct island violations (16a, 17a). Surprisingly, the combination of a 'good' gap and a 'bad' gap in these constructions yields a result that is judged to be good. These phenomena are hard to capture in terms of CNTs, because their properties depend on chunks of structure that are not local to one another. If parasitic gap constructions occur sufficiently often in the child's input corpus, then a distributional learner might conclude from them that extraction from subject and adjunct clauses is acceptable more generally. If they occur too rarely in the input to be noticed, we instead face the problem that the distributional learner would treat parasitic gaps as impossible. It might be tempting to dismiss parasitic gaps as marginal phenomena, but the acceptability patterns are robust, and speakers show rapid sensitivity to their properties in on-line comprehension (Phillips 2006; Wagers & Phillips 2009).

16 a. * Which car did the attempt to fix __ ultimately damage the tools?
b. Which car did the attempt to fix __ ultimately damage __?

17 a. * Which theorem did Hubert prove the conjecture without understanding __?
b. Which theorem did Hubert prove __ without understanding __?

A related challenge for CNTs can be found in another type of multiple gap construction, coordinate structures. It is well known that a single *wh*-phrase can bind a gap in a pair of conjoined phrases (18a), and that if one conjunct contains a gap then the other conjunct must also contain a gap (18b). This generalization is known as the *Coordinate Structure Constraint* (Ross 1967; Gazdar 1981), and it applies regardless of the size of the conjuncts. For this reason it should be difficult to explain the contrast between (19a) and (19b) in terms of CNTs. Both sentences contain *wh*-dependencies consisting of CNTs that are all relatively frequent.

18 a. Which book did you buy ___ and read ___ in a single afternoon?
b.* Which book did you buy a newspaper and read ___ in a single afternoon?

19 a. Which book did Sally say that Jim bought ___ and Mary know that Jim read ___?
b. * Which book did Sally say that Jim bought a newspaper and Mary know that Jim read ___?

Summarizing this section, one of the most interesting findings of Pearl and Sprouse's project is that, for the specific island phenomena that they model, the input corpus appears to be sufficient to support successful learning. This challenges the widespread assumption that island constraints create a Poverty of the Stimulus problem. But based on considerations of the range of island phenomena that learners must come to know, I think that Pearl and Sprouse's corpus analyses may, in fact, help to strengthen the Poverty of the Stimulus argument for island constraints.

5. Cross-language Contrasts and the Parsing Problem

Innate domain-specific constraints contribute to theories of language learning in at least two ways. First, they offer one possible answer to the question of how children come to know details about language that it would be very hard to learn from the environment. Second, they provide an account of the limits on cross-language variation. If a linguistic constraint is simply built into the learner, then there is no need to learn the constraint, and the constraint should apply in all languages. Island constraints have long been regarded as excellent candidates for this kind of account: they are hard to learn, and they appear to apply in similar ways across languages. However, there are reasons to revisit both of these assumptions. There is growing evidence for cross-language variation in island constraints, suggesting that there may be more need for an account of how island constraints are learned from the input. And Pearl and Sprouse provide an explicit model that claims to do exactly that. Therefore, traditional claims about islands and innateness are certainly ripe for reassessment. In this section I consider how the distributional learning model might fare as an account of cross-language variation in island effects.

As a preliminary remark, it should be noted that cross-language variation in island effects is still rather limited. Although in most languages island effects have not been investigated in as much detail as they have in languages such as English, Japanese, Italian, and Chinese, it appears that in cross-language studies the same factors seem to be relevant for islandhood again and again. For example, relative clauses consistently resist extraction, so it is newsworthy when we discover a language or a construction that appears to allow extraction from relative clauses. It is standard to find that object extraction is more liberal than subject and/or adjunct extraction, and it is rare indeed to find situations where this contrast is reversed. Operators such as negation often interfere with *wh*-dependencies, but we tend not to find situations where pronouns create barriers for extraction. And there are many other cases like this. These cross-linguistic regularities are expected under an account that assumes innate island constraints (with the possibility of parameterization of some of those constraints). The regularities are more surprising under an account that claims that island constraints are simply derived from the input without domain-specific innate constraints.

As discussed in the companion paper to this article, current understanding of cross-language variation in island effects suggests that there are two types of variation. One type of variation involves cases where cross-language differences can be tied to independently motivated structural possibilities, which yield the surface appearance of varying island effects, but without the need to assume variation in the underlying constraints. I refer to this first type of variation as *surface island variation*. The second type of variation involves cases that – at least at present – cannot be reliably linked to independent structural variation, and hence suggest true variation in the island constraints themselves. I refer to these cases as *deep island violation*. I consider each of these two types of variation in turn, and compare how a distributional learner and a learner with innate constraints would fare.

5.1 Surface variation in islands and the parsing problem

First consider the case of contrasting complementizer-trace phenomena in English and Italian. English does not allow extraction of a subject from a position immediately following an overt complementizer (20a), but the corresponding sentence in Italian is perfectly fine. As indicated in (20b), this difference has been argued to reflect the availability of a post-verbal subject position in Italian (20c). Under this account, English and Italian both obey the restriction on gaps ('traces') that immediately follow complementizers, but Italian is able to circumvent the restriction by extracting a subject *wh*-phrase from the post-verbal position (Rizzi 1982; Roberts & Holmberg 2010). As such, the two languages differ in the acceptability of the corresponding surface strings, but they both obey the complementizer-trace restriction.

20 a. * Who did you say that __ wrote this book?

b. Chi$_i$ hai detto che ha scritto questo libro __$_i$?
who have.2sg said that has written this book

c. Hanno telefonato molti studenti.
have.3pl called many students

How would the distributional learner fare in learning the English and Italian generalizations? We have already seen that the distributional learner might struggle to learn the complementizer-trace constraint in English, due to the sparseness of the relevant examples in the input, specifically due to the rarity of *wh*-questions with overt *that*. If we make the plausible assumption that Italian adults ask their children roughly the same questions that English-speaking adults do, then we can estimate what the input to Italian children looks like, by combining the counts of null/overt complementizer *wh*-questions in (12) above. Italian children simply need to learn that long-distance subject questions are possible (Italian complementizers are obligatorily overt), so the learning task should be straightforward if there are sufficiently many long-distance subject questions in the input. Following Pearl and Sprouse's assumption that children learn island constraints from around three years' worth of input, which amounts to around an order of magnitude more data than are found in their corpus, we can estimate that Italian children might encounter long-distance subject questions about once per week. Is that sufficient for learning? Perhaps. We do not know how much input is needed, and once per week is neither in the 'obviously sufficient' or 'obviously insufficient' range.

However, even if the once-per-week subject question is sufficiently frequent, the Italian child faces a dilemma. We can safely assume that the child would quickly learn that Italian declarative sentences include both pre-verbal and post-verbal subjects, and it is reasonable to assume that the child might consider both of these subject positions as possible gap sites when parsing the *wh*-questions that he encounters in the input. So what conclusion should the child draw about the structure of long-distance subject questions? He might simply treat the sentence as ambiguous, i.e., as providing evidence for two different parses. Or he might probabilistically choose one of the two possible parses, again leading him to conclude that both pre-verbal and post-verbal subject gaps are possible in Italian. Existing evidence suggests that children have great difficulty entertaining multiple parses for a sentence (Trueswell, Sekerina, Logrip, & Hill 1999; Snedeker & Trueswell 2004; Leddon & Lidz 2006; Musolino & Lidz 2006), and that they exhibit a locality bias in parsing *wh*-questions (Omaki, Davidson White, Lidz, & Phillips, submitted). Therefore, the most likely scenario is that an Italian distributional learner would consistently analyze examples like (20b) as involving a pre-verbal subject gap, and hence would not conclude that there is a restriction on complementizer-trace sequences in Italian. But the learner would at least correctly treat (20b) as well formed, so perhaps we should not be concerned about this (more on this in a moment).

Compare this with what a 'Principles and Parameters' learner would do, armed with an innate constraint that rules out complementizer-trace sequences.[5] In English the learner would not need to learn anything about the restriction, and would only need to learn that English allows the option of overt and null complementizers. In Italian the learner would need to independently learn that the language allows both pre-verbal and post-verbal subjects. If the learner recognizes that Italian allows at least some form of long-distance *wh*-questions, then he should be able to generalize this to the possibility of long-distance subject questions launched from post-verbal subject position. This means that the learner could infer the acceptability of (20b) without actually observing examples like that. But if he did encounter such examples in the input, he would not treat them as unambiguous. The only parse that is consistent with his innate constraints would be one in which the gap is in the post-verbal subject position.

So the two learners should both correctly conclude that Italian (20b) is allowed, albeit with differing parses for the same sentence. Should we care about the different parses, if they both generate the same surface string? We probably should care, based on evidence from various related phenomena that show that the parse with a postverbal subject gap is the correct one (e.g., Rizzi 1982; Kenstowicz 1989). One such piece of evidence comes from the Fiorentine dialect of Italian (Brandi & Cordin 1989).

Fiorentino is like standard Italian in most respects, but one important difference is that Fiorentino requires a pre-verbal subject clitic that agrees in gender and number with the subject. This is shown in (21) for sentences with overt and null 3rd person singular subjects. Importantly, however, Fiorentino sentences with postverbal subjects use a default 3rd person masculine singular clitic, rather than agreeing with the postverbal subject (22).

21 a. Mario e parla
Mario SCL speaks
'Mario speaks.'

b. e parla
SCL speaks
'He speaks.'

c. * Parla
speaks

22 Gl ha telefonato delle ragazze.
SCL$_{MSg}$ has telephoned some girls$_{FPl}$
'Some girls telephoned.'

Since Fiorentino sentences show different subject-verb agreement patterns, depending on the position of the subject, it is possible to test the position of the gap in long-distance subject extractions. The evidence confirms the claim that long-distance subject extractions

[5] This discussion does not do justice to current work on the source and scope of variation in complementizer-trace phenomena (e.g., Rizzi 2006; Lohndal 2009).

originate in postverbal subject position, as the subject clitic must show default masculine singular agreement (23).

23 a. Quante ragazze tu credi che gli abbia parlato?
How-many girls you think that SCL$_{MSg}$ has$_{3Sg}$ spoken
'How many girls do you think have spoken?'

b. *Quante ragazze tu credi che le abbiano parlato?
How-many girls you think that SCL$_{FPl}$ have$_{3Pl}$ spoken

A Fiorentine learner with an innate constraint on complementizer-trace sequences, and with knowledge of the agreement patterns for preverbal and postverbal subjects, should automatically know that (23a) is the correct form for a long-distance subject question, rather than (23b), without directly encountering examples like (23a) in the input. In contrast, the distributional learner would likely conclude from examples with singular subjects that preverbal subject gaps are possible, and so should incorrectly conclude that (23b) is fine. Of course, it is conceivable that the learner might encounter many examples in the input of long-distance subject *wh*-questions that might allow him to directly observe that default agreement is required. But this is unlikely, due to a further ambiguity problem. Most long-distance subject questions in the input are likely to have a masculine singular subject (*who*, *what*, or *which*-N), and so the default subject-verb agreement clitic that is required in long-distance subject questions could easily be mistaken for full agreement with the subject. The most informative types of long-distance subject question, i.e., those with plural *wh*-phrases, are probably extremely rare in the input, since they would occur only in a subset of *which*-N questions Therefore it would likely be hard for a distributional learner to learn the appropriate generalization.

As a second example of surface variation in island effects, consider the cases of 'escapable' relative clauses in East Asian languages such as Chinese, Japanese, and Korean (Kuno 1973; Inoue 1976; Sohn 1980; Hasegawa 1981; Huang 1984; Kang 1986; Tsai 1997; Li 2002). These languages allow surface strings that appear to involve filler-gap dependencies that cross a relative clause boundary, challenging the universality of the ban on extraction from relative clauses. However, there are good linguistic arguments that the extractions from relative clauses are, in fact, illusory, and that they are instances of so-called *major subject constructions* (MSCs) in those languages (Sakai 1994; Han & Kim 2004; Hoshi 2004; Hsu 2006; Ishizuka 2009). MSCs allow a noun phrase topic that is outside a relative clause to license a null subject that is the relative clause, as shown by the Japanese example in (24a). When the RC-external noun phrase is extracted, it yields a surface word order that closely resembles an illicit extraction from an RC, but this is misleading. The true gap site is outside the RC, and so avoids the ban on extraction from RCs, and it is related to the null subject position inside the RC by an antecedent-pronoun relation (24b). Evidence for this analysis comes from demonstrations that constraints on MSCs also restrict the apparent extractions from relative clauses. This includes the restriction that the apparent gap be a subject, and restrictions on the argument structure of the predicate in the higher clause (e.g., *yogoreteiru* 'be dirty' in (24)). Thus, these languages do not show variation in the islandhood of relative clauses. Rather, they have an additional structural option that creates the illusion of acceptable extraction from relative clauses.

24 a. [$_{\text{IP}}$ sono sinsi$_i$-ga [$_{\text{NP}}$ [$_{\text{CP}}$ pro$_i$ __$_j$ kiteiru] [yoohuku$_j$]]-ga yogoreteiru]
that gentleman-NOM pro wearing-is suit-NOM dirty-is
'That gentleman is such that the suit that he is wearing is dirty.'

b. [$_{\text{CP}}$ Op$_i$ [$_{\text{IP}}$ __$_i$ [$_{\text{NP}}$ [$_{\text{CP}}$ pro$_i$ __$_j$ kiteiru] yoohuku$_j$]-ga yogoreteiru] [sinsi$_i$]]
Op pro wearing-is suit-NOM dirty-is gentleman
'The gentleman who the suit that he is wearing is dirty.'

What would a distributional learner conclude about the escapable RCs in these languages? In light of what we have learned about the sparseness of the data that children must learn from, there is a clear danger that the learner would not encounter any relevant examples like this. In fact, that might turn out to be the best option for this learner. If the learner does encounter examples of these extractions, then a couple of possibilities present themselves. If the learner does not already know about MSCs, then he would presumably parse a sentence like (24b) as involving a *wh*-dependency that crosses a RC boundary. This would then count as evidence that extraction from RCs is possible in general in the target language, leading to substantial overgeneralization. If instead the learner already has learned about the possibility of MSCs in the target language, then sentences like (24b) would count as ambiguous, since they could be parsed either as long-distance extractions from inside a RC or as local extractions in an MSC. This would again raise the danger that the learner would count at least some of the examples as evidence that extraction from RCs is allowed in general in the language, again creating a risk of overgeneralization.[6]

In contrast, a Principles and Parameters learner equipped with knowledge of a universal ban on extraction from RCs would fare differently. If that learner encountered a sentence like (24b) it should either analyze it as a speech error, or it should analyze it as an instance of local extraction in an MSC. The option that (24b) is an example of acceptable extraction from an RC should not be available to this learner. In fact, this learner should be able to correctly parse (24b) without ever encountering such examples in the input. As long as the learner independently knows the properties of MSCs, and knows that local extraction is possible, it should be able to infer that local extraction from an MSC is possible.

Interestingly, the best outcome for a distributional learner might be to never encounter examples like (24b) in the input. If this learner can independently learn that MSCs are possible, and that local *wh*-dependencies are possible, then it should be able to accept a local *wh*-dependency in an MSC, and so it should also correctly accept examples like (24b)

[6] I mentioned above that children appear to show a locality bias in interpreting ambiguous *wh*-dependencies (Omaki et al. submitted), just as adults do. We might therefore predict that children would favor the parse of (24b) as involving local extraction in an MSC, rather than long-distance extraction from an RC, thereby avoiding the danger of overgeneralization. But this prediction is not so straightforward, as the locality bias holds when all other aspects of the competing parses are equivalent. It is uncertain whether learners would favor a parse involving an MSC and a shorter relativization over a parse with a longer relativization and no MSC. Also, on-line studies with Japanese adults (Aoshima, Phillips, & Weinberg 2004) and children (Omaki et al. submitted) shows that the relevant notion of locality for Japanese is not one that favors structurally shorter *wh*-dependencies. Rather, it favors dependencies that satisfy the thematic or scope requirements of the *wh*-phrase as quickly as possible. Due to the head-final nature of Japanese, this means that structurally longer dependencies may be favored over structurally shorter dependencies.

only when they match the independently learned properties of MSCs. In this instance, then, the greatest risk to the distributional learner might come from actually encountering positive examples of escapable relative clauses, as only then does the danger of misparsing present itself. It remains to be seen whether there are other cases where a distributional learner is better served by failing to encounter a key example in the input. We normally assume that distributional learners should fare better when they receive more input, but this might not always be true.

5.2 Deep variation in islands

The examples above of cross-language variation from Romance and East Asian languages involve surface variation in island effects, rather than genuine variation in island constraints. But there are also cases where current evidence suggests that there is genuine variation in island constraints. For example, extraction from complex subjects is generally degraded-to-unacceptable in English and many other languages, but there are some languages in which this is possible. For example, Stepanov (2007) gives examples of acceptable subject extraction from Russian (25), Hungaran (Kiss 1987), Palauan (Georgopoulos 1991), and other languages. Other cases of apparently genuine variation in island constraints involve the islandhood of certain types of adjunct clauses (e.g., Japanese and Malayalam vs. Russian and Malay: Yoshida 2006) and the presence of island effects in argument *wh*-in-situ questions (e.g., Chinese vs. Hindi: Malhotra 2009).

25 a. * What do you wish that [to buy __] would be no trouble at all.

b. Cto by ty xotel ctoby kupit' ne sostavljalo by nikakogo truda?
what SUBJ you wanted that-SUBJ to-buy not constitute SUBJ no labor
'What would you want that [to buy __] would not be any trouble?'

In these cases the distributional learner and the learner with innate constraints face a similar task. If the learner's task in such cases is to choose between a more restrictive and a more liberal grammar, then both types of learner should adopt the more restrictive grammar, unless they encounter positive evidence of the *wh*-dependencies that are possible only in the more liberal grammar. For the distributional learner, this is because the learner only allows *wh*-dependencies that can be built using CNTs that it has encountered in its prior experience. The distributional learner is an inherently conservative learner. For a learner with innate constraints, this is because the learner likely has a built in bias to select the more restrictive of a pair of grammatical alternatives. In the case of a language that allows more liberal extraction, both types of learner need the input corpus to contain positive examples of the more liberal extraction. If these examples do not reliably occur in the input, then neither learner should be able to converge on the target language. To my knowledge, we currently have no good evidence on the presence or absence of such examples in child-directed speech in the relevant languages.

6. Generalizing across Dependency Types

Finally, I should highlight an important goal for any distributional learner that seeks to discover constraints on extraction without the help of innate domain-specific knowledge. An central finding from the past 40 years of syntax research is that *wh*-dependencies are just one among a class of unbounded dependencies that obey very similar constraints. Relativization, topicalization, comparatives, and adjective-*though* constructions are all subject to the same island constraints as *wh*-dependencies. In the transformational grammar literature these dependencies are known by the unfortunately opaque name *A' ('A-bar') dependencies*. Meanwhile, there are other types of long-distance dependencies, including raising, bound variable anaphora and dependencies involving resumptive pronouns, that are not subject to the same constraints as *wh*-dependencies. How do learners come to know which long-distance dependencies are underlyingly the same and which are different? How do they know that evidence for restrictions on one type of dependency can be treated as evidence that the same restriction applies to another dependency that they might encounter less often?

In a theory in which learners are equipped with innate domain-specific constraints, the learner starts with the knowledge that there is a limited set of linguistic dependency types, and his task is simply to identify which constructions in the ambient language exemplify which classes of dependency. Once a dependency has been classified as an A' dependency, the learner can immediately transfer what he has learned about *wh*-dependencies to this other type of dependency.

In Pearl and Sprouse's distributional learner it is less clear how cross-classification of dependencies might occur. It is probably not a viable option to simply assume that the constraints on each type of dependency are learned separately. As we have seen, the distributional learner faces a serious data sparseness problem even for *wh*-dependencies, which, together with relativization dependencies, probably make up the vast majority of the A' dependencies in the input corpus. Therefore the data sparseness problem is probably even more acute for other types of A' dependencies, making it all the more important for the learner to be able to combine evidence across all kinds of A' dependencies. It remains to be seen how this can be achieved, and whether it can be done without falsely generalizing to other types of long-distance dependencies that are not subject to island constraints, such as forwards and backwards anaphora.

7. Conclusion to Part II

Pearl and Sprouse's model represents a very interesting step forward in discussions about distributional learning of syntactic phenomena. The model shifts the debate about distributional alternatives to innate linguistic knowledge into a domain where the debate belongs, i.e., phenomena that linguists have regarded as providing good evidence for innate linguistic constraints. The model is simple and transparent, and it is not difficult to relate it to proposals in the formal syntax literature, all of which make the model eminently testable. In addition, Pearl and Sprouse have done a great service by providing a comprehensive analysis of the *wh*-dependencies in corpora whose scale is not too far removed from the input that a child must learn from. This also makes it feasible to assess what information about *wh*-dependencies is available to real children.

Pearl and Sprouse argue that their model is able to derive island constraints from the input data without the benefit of innate domain-specific knowledge. They emphasize that their model learns that longer *wh*-dependencies have a different status than shorter *wh*-dependencies, but that the model distinguishes this dependency-length effect from island effects, matching human judgment data. However, I think that the information in Pearl and Sprouse's study ultimately strengthens the case for innate constraints rather than weakening it. Although the model assigns different probabilities to long dependencies and island violations, it is probably insufficient to treat this as a mere quantitative difference. The corpus analyses suggest that the data sparseness problem for learning island constraints is, in fact, quite serious. And although some aspects of cross-language variation in island effects do need to be learned, the data sparseness uncovered by Pearl and Sprouse's analyses demonstrate how valuable it is for learners to be guided by universal constraints and by information that they learn from other constructions.

I should emphasize that the arguments outlined here are not intended as arguments against distributional mechanisms in language learning. Rather, they are arguments against the utility of distributional learning in the absence of a strong set of learning biases.

Acknowledgements

This work was supported in part by NSF Grant BCS-0848554.

References

Adams, M. (1985). Government of empty subjects in factive clausal complements. *Linguistic Inquiry, 16*, 305-313.

Aoshima, S., Phillips, C., & Weinberg, A. S. (2004). Processing filler-gap dependencies in a head-final language. *Journal of Memory and Language, 51*, 23-54.

Boeckx, C. (2008). Islands. *Language and Linguistics Compass, 2*, 151-167.

Brandi, L. & Cordin, P. (1989). Two Italian dialects and the null subject parameter. In O. Jaeggli & K. Safir (eds.), *The null subject parameter*, pp. 111-142. Dordrecht: Kluwer.

Cattell, R. (1978). On the source of interrogative adverbs. *Language, 54*, 61-77.

Chomsky, N. (1964). *Current issues in linguistic theory*. The Hague: Mouton.

Chomsky, N. (1973). Conditions on transformations. In S. Anderson & P. Kiparsky (eds.), *A festschrift for Morris Halle*, pp. 232-286. New York: Holt, Rinehart, & Winston.

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

Cinque, G. (1999). *Adverbs and functional heads: A cross-linguistic perspective*. Oxford University Press.

Crain, S. & Thornton, R. (1998). *Investigations in Universal Grammar*. Cambridge, MA: MIT Press.

Culicover, P. & Postal, P., eds. (2001). *Parasitic gaps*. Cambridge, MA: MIT Press.

de Villiers, J. (1995). Questioning minds and answering machines. In D. MacLaughlin & S. McEwan (eds.), *Proceedings of the 19th Boston University Conference on Language Development*, pp. 20-36. Somerville, MA: Cascadilla Press.

Engdahl, E. (1983). Parasitic gaps. *Linguistics and Philosophy, 5*, 5-34.

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

Fiengo, R. & Higginbotham, J. (1981). Opacity in NP. *Linguistic Analysis, 7,* 395-421.

Gazdar, G. (1981). Unbounded dependencies and coordinate structure. *Linguistic Inquiry, 12*, 155-184.

Georgopoulos, C. (1991). *Syntactic variables: Resumptive pronouns and binding in Palauan.* Dordrecht: Kluwer.

Gibson, E. (1998). Linguistic complexity: Locality of syntactic dependencies. *Cognition, 68*, 1-76.

Gomez, R. (2002). Variability and detection of invariant structure. *Psychological Science, 13*, 431-436.

Grimshaw, J. (1986). Subjacency and the S/S' parameter. *Linguistic Inquiry, 17*, 364-369.

Han, C. & Kim, J. (2004). "Double relative clauses" in Korean? *Linguistic Inquiry, 35*, 315-337.

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

Hasegawa, N. (1981). A lexical interpretive theory with emphasis on the role of subject. PhD dissertation, University of Washington.

Hawkins, J. A. (1999). Processing complexity and filler-gap dependencies across languages. *Language, 75*, 224-285.

Hofmeister, P., Staum Casasanto, L., & Sag, I. (this volume). Islands in the grammar: Evidence and non-evidence.

Hofmeister, P. & Sag, I. A. (2010). Cognitive constraints on syntactic islands. *Language*, *86,* 366-415.

Hoshi, K. (2004). Parameterization of the external D-system in relativization. *Language, Culture, and Communication, 33*, 1-50.

Hsu, C.-C. N. (2006). Issues in head-final relative clauses in Chinese: Derivation, processing, and acquisition. PhD dissertation, University of Delaware.

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

Inoue, K. (1976). *Henkei-bunpo to nihongo.* Tokyo: Taishukan.

Ishizuka, T. (2009). CNPC violations and possessor raising in Japanese. Ms. UCLA.

Kang, Y.-S. (1986). *Korean syntax and Universal Grammar*. PhD dissertation, Harvard University.

Kenstowicz, M. (1989). The null subject parameter in modern Arabic dialects. In O. Jaeggli & K. Safir (eds.), *The null subject parameter*, pp. 263-275. Dordrecht: Kluwer.

Kiparsky, P. & Kiparsky, C. (1971). Fact. In M. Bierwisch & K. Heidolph (eds.), *Progress in linguistics*. The Hague: Mouton.

Kiss, K. É. (1987). *Configurationality in Hungarian.* Dordrecht: Reidel.

Kluender, R. (1998). On the distinction between strong and weak islands: A processing perspective. In P. Culicover & L. McNally (eds.), *Syntax and Semantics 29: The limits of syntax*, pp. 241-279. San Diego, CA: Academic Press.

Kluender, R. (2005). Are subject islands subject to a processing account? In V. Chand, A. Kelleher, A. J. Rodriguez, & B. Schmeiser (eds.), *Proceedings of the 23rd West Coast Conference on Linguistics*, pp. 475-499.

Kluender, R. (this volume). *Title …*

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

Kuno, S. (1973). *The structure of the Japanese language.* Cambridge, MA: MIT Press.

Lasnik, H. & Saito, M. (1992). *Move alpha: Conditions on its application and output*. Cambridge, MA: MIT Press.

Leddon, E. M. & Lidz, J. L. (2006). Reconstruction effects in child language. In D. Bamman, T. Magnitkaia, & C. Zaller (eds.), *Proceedings of the 30th Annual Boston University Conference on Language Development,* pp. 328-339. Somerville, MA: Cascadilla Press.

Lohndal, T. (2009). Comp-t effects: Variation in the position and features of C. *Studia Linguistica, 63*, 204-232.

Malhotra, S. (2009). On *wh*-quantifier interactions. Ms. University of Maryland.

Manzini, M. R. (1992). *Locality: A theory and some of its empirical consequences*. Cambridge, MA: MIT Press.

Maye, J., Werker, J. F., & Gerken, L.-A. (2002). Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition, 82*, B101-B111.

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.

Musolino, J. & Lidz, J. (2006). Why children aren't universally successful with quantification. *Linguistics, 44*, 817-852.

Omaki, A. (2010). *Commitment and flexibility in the developing parser.* PhD dissertation, University of Maryland.

Omaki, A., Davidson White, I., Goro, T., Lidz, J., & Phillips, C. (submitted). No fear of commitment: Children's incremental interpretation in English and Japanese *wh*-questions.

Omaki, A. & Schulz, B. (2011). Filler-gap dependencies and island constraints in second language sentence processing. *Studies in Second Language Acquisition, 33*, 563-588.

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*, 235-265.

Pearl, L. & Sprouse, J. (submitted). Syntactic islands without Universal Grammar: A computational model of the acquisition of constraints on long-distance dependencies.

Pearl, L. & Sprouse, J. (*this volume*). Computational models of acquisition for islands. In J. Sprouse & N. Hornstein (eds.): *Experimental syntax and island effects*. Cambridge University Press.

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

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

Phillips, C., Kazanina, N., & Abada, S. (2005). ERP effects of the processing of syntactic long-distance dependencies. *Cognitive Brain Research, 22*, 407-428.

Postal, P. (1998). *Three investigations of extraction*. Cambridge, MA: MIT Press.

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

Reali, F. & Christiansen, M. (2005). Uncovering the statistical richness of the stimulus: Structure dependence and indirect statistical evidence. *Cognitive Science, 29*, 1007-1028.

Redington, M., Chater, N., & Finch, S. (1998). Distributional information: A powerful cue for acquiring syntactic categories. *Cognitive Science, 22*, 425-469.

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

Rizzi, L. (1982). *Issues in Italian syntax.* Dordrecht: Foris.

Rizzi, L. (1990). *Relativized minimality.* Cambridge, MA: MIT Press.

Rizzi, L. (2006). On the form of chains: Criterial positions and ECP effects. In L. L.-S. Cheng & N. Corver (eds.), *Wh-movement: Moving on*, pp. 97-133. Cambridge, MA: MIT Press.

Roberts, I. & Holmberg, A. (2010). Introduction: Parameters in minimalist theory. In T. Biberauer, A. Holmberg, I. Roberts, & M. Sheehan (eds.), *Parametric variation: Null subjects in minimalist theory*, pp. 1-58. Cambridge, UK: Cambridge University Press.

Rooryck, J. (1992). Negative and factive islands revisited. *Journal of Linguistics, 28*, 343-373.

Ross, J. R. (1967). *Constraints on variables in syntax.* PhD dissertation, MIT.

Saffran, J., Aslin, R., & Newport, E. (1996). Statistical learning by 8-month old infants. *Science, 274*, 1926-1928.

Sakai, H. (1994). Complex NP constraint and case conversions in Japanese. In M. Nakamura (ed.), *Current topics in English and Japanese*, pp. 179-203. Tokyo: Hituzi Syobo.

Snedeker, J. & Trueswell, J. C. (2004). The developing constraints on parsing decisions: The role of lexical-biases and referential scenes in child and adult sentence processing. *Cognitive Psychology, 49*, 238-299.

Sohn, H. (1980). Theme prominence in Korean. *Korean Linguistics, 2*, 2-19.

Solan, Z., Horn, D., Ruppin, E., & Edelman, S. (2005). Unsupervised learning of natural languages. *Proceedings of the National Academy of Sciences, 102*, 11629-11634.

Sportiche, D. (1981). Bounding nodes in French. *The Linguistic Review, 1*, 219-246.

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

Stepanov, A. (2007). The end of CED: Minimalism and extraction domains. *Syntax, 10*, 80-126.

Stowe, L. A. (1986). Parsing WH-constructions: Evidence for on-line gap location. *Language and Cognitive Processes, 3*, 227-245.

Szabolcsi, A. & den Dikken, M. (1999). Islands. *GLOT International, 4*, 3-8.

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.

Trueswell, J. C., Sekerina, I., Hill, N. M., & Logrip, M. L. (1999). The kindergarten path effect: Studying on-line sentence processing in young children. *Cognition, 73*, 89-134.

Tsai, W.-T. D. (1997). On the absence of island effects. *Tsing Hua Journal of Chinese Studies, 27*, 125-149.

Vallabha, G. K., McClelland, J. L., Pons, F., Werker, J. F., & Amano, S. (2007). Unsupervised learning of vowel categories from infant-directed speech. *Proceedings of the National Academy of Sciences, 104*, 13273-13278.

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

Wagers, M. W. & Phillips, C. (submitted). Going the distance: memory and decision making in active dependency construction.

Yoshida, M. (2006). *Constraints and mechanisms in long-distance dependency formation*. PhD dissertation, University of Maryland.

Zukowski, A. & Larsen, J. (2011). *Wanna* contraction in children: Retesting and revising the developmental facts. *Language Acquisition, 18*, 211-241.

Department of Linguistics
1401 Marie Mount Hall
College Park, MD 20742

colin@umd.edu