Islands in the grammar? Standards of evidence

Philip Hofmeister
Department of Language and Linguistics
University of Essex
Colchester, UK
E-mail: phofme@essex.ac.uk

Laura Staum Casasanto
SUNY at Stony Brook
Stony Brook, NY
E-mail: laura.staum@gmail.com

Ivan A. Sag
Department of Linguistics
Stanford University
Stanford, CA
E-mail: sag@stanford.edu

When considering how a complex system operates, the observable behavior depends upon both architectural properties of the system and the principles governing its operation. The behavior of computer chess programs, for instance, depends upon both the processing speed and resources of the computer and the programmed rules that determine how the computer selects its next move. Despite having very similar search techniques, a computer from the 1990s might make a move that its 1970s forerunner would overlook simply because it had more raw computational power. From the naïve observer's perspective, however, it is not clear if a particular move is dispreferred or overlooked because of computational limitations or the search strategy and decision algorithm. In the case of computers, evidence for the source of any particular behavior can ultimately be found by inspecting the code and tracking the decision process of the computer.

But with the human mind, such options are not yet available. The preference for certain behaviors and the dispreference for others may theoretically follow from cognitive limitations or from task-related principles that preclude certain kinds of cognitive operations, or from some combination of the two. This uncertainty gives rise to the fundamental problem of finding evidence for one explanation over the other. Such a problem arises in the analysis of syntactic island effects – the focus of this volume. These involve the low acceptability ratings elicited by sentences with long-distance dependencies into certain syntactic configurations, as well as the general rarity of such utterances in attested speech or text (Chomsky 1962, 1973, 1977, 1981, 1986; Ross 1967). What makes island constructions of particular interest is the fact that the traditional syntactic analyses behind them have been highly influential in the development of syntactic theories – movement-based analyses of syntactic and semantic phenomena, in particular.

As our knowledge of language processing has increased, the possibility of an alternative to grammar-based accounts of island effects has emerged: perhaps these low judgments are a consequence of the processing difficulty comprehenders face when reading or listening to these sentences, as argued for by Deane (1991), Kluender (1991, 1998), and Kluender and Kutas (1993). Building on this work, we have provided evidence for the role of processing pressures in island phenomena elsewhere, as reflected by reading times (Hofmeister 2007; Hofmeister, Jaeger, Sag, Arnon, and Snider 2007; Sag, Hofmeister, & Snider 2007; Hofmeister & Sag 2010). Here, our goal is to assess certain arguments that have been made to the effect that grammatical constraints MUST be involved in island phenomena. Our criticism of these arguments is not meant to exclude the possibility that grammatical constraints play a role in island effects – in the absence of precisely formulated, observationally adequate grammatical accounts of island effects, any such argument is most likely futile. In spite of this difficulty, it is our contention that independently motivated processing factors can successfully explain a substantial amount of the judgment variation that has been used to motivate island constraints within grammar. Hence, a number of such grammar-internal structural constraints need not be assumed as being part of the human biological endowment for language.

Over the course of this discussion, we weigh the adequacy of several types of evidence invoked in support of grammatical accounts of island phenomena. These include satiation effects, the relationship between working memory and judgments of acceptability, and data from filled-gap paradigms and plausibility manipulations.

Much of the relevant evidence relies at least partially on acceptability judgments. This reliance is not particular to arguments about island effects – acceptability judgments form the bedrock of syntactic theory. But as numerous linguists and researchers have pointed out, they are an imperfect window into the mind (Chomsky & Miller 1963; Miller & Chomsky 1963; Fodor 1978; Schütze

1996; Cowart 1997; Hofmeister & Sag 2010; Staum Casasanto, Hofmeister, & Sag 2010). That is, one-dimensional scores derived from acceptability experiments potentially reflect the contributions of multiple factors, making it extremely difficult to identify the role of any one factor – ungrammaticality, contextualization difficulty, semantic anomaly, processing difficulty, etc. The difficulties of interpreting acceptability judgments are, of course, not unique to that method. Reading times, eye movements, reaction times, and other dependent measures are all indirect ways of quantifying internal cognitive states of a speaker or listener, which we have no direct access to. As a result, all such data potentially reflect the interaction of many diverse factors, which can make it challenging to attribute observed variation in one dimension to a specific type of constraint.

In order to assess the kinds of evidence relevant to the debate over the status of island effects (henceforth, the islands debate), the state of our knowledge about many of them, especially acceptability judgments, must be reevaluated. Researchers (ourselves included) have attempted to use acceptability judgments to adjudicate contentious issues (as in the islands debate) without a well-informed understanding of how judgments vary in uncontroversial contexts. Part of our analysis here, therefore, is informed by experimental data comparing sentences which exhibit uncontroversial processing difficulty, but no other deficit, to those containing grammatical violations. By looking at these clear cases, we can provide a principled basis for our treatment of controversial examples in terms of processing or grammar.

In the end, we argue that many of the findings from studies putatively supporting grammar-based interpretations of island phenomena have plausible, alternative interpretations rooted in specific, well-documented processing mechanisms. In other instances, we suggest that the experimental measures are not understood well enough at present for them to be used as decisive evidence for either grammar-based or processing-based of island effect. Though the grammar-external explanations that we isolate render many familiar grammatical proposals obsolete (e.g. the SUBJACENCY CONDITION, the SUPERIORITY/MINIMAL LINK CONDITION), we can of course not conclude that grammatical constraints have no role to play in the explanation of island phenomena. However, based on a review of the available evidence, it is likely that a prerequisite for future progress in the islands debate is a better understanding of the available sources of linguistic evidence.

1 Island Effects

The classic data surrounding island effects indicate that unacceptability results when dependencies enter into certain syntactic configurations, such as relative clauses (1) and interrogative clauses (2) (Ross 1967; Chomsky 1973, 1977, 1981, 1986).

(1) This is the puzzle that we met the mathematician [who solved ___].

(2) What did Craig wonder [whether the doctors knew ___]?

Within the class of islands, however, changes to lexical items, the similarity of discourse references, the syntactic and semantic content of the dislocated items, and other manipulations can radically alter the acceptability of the islands, even though they leave the grammatical structure intact (see Hofmeister & Sag 2010 for a review). Indeed, many linguists have argued that island effects go away under certain circumstances (Ross 1967; Chomsky 1973; Cinque 1990; Rizzi 1990; Pesetsky 1987, 2000; Deane 1991). The example in (3) from Ross (1967), for instance, violates the subjacency

condition, while (4) counterexemplifies Chomsky's (1973) Subject Condition. Moreover, extraction out of non-finite clauses appears in general to be more acceptable than extraction out of finite clauses, as shown in (5) ('A > B' indicates that A is more acceptable than B):

(3) Which reports does the government prescribe the height of the lettering on (the covers of) __?

(4) . . . many also were seized and sold into slavery, of which only some ___ had been ransomed at the time he wrote.

(5) a. Which questions did he know how to answer __? >
b. Which questions did he know how he should answer __?

The typical response from defenders of a grammar-based view of islands is that counterexamples such as these have some special characteristic that nullifies or weakens syntactic island constraints. For instance, a recurring claim regarding counterexamples is that they exhibit different syntactic structure, often in the form of hidden constituents or structure, such as null resumptive pronouns or an invisible syntactic controller (Georgopoulos 1985, 1991; Saah & Goodluck 1995; Phillips in press). Such proposals have failed to provide independent evidence for these otherwise unmotivated assumptions.

In contrast, the variation in acceptability judgments for island-violating sentences frequently seems to correlate with differences in factors known to affect processing difficulty, e.g. depth of embedding, similarity-based interference, structural ambiguity, collocational frequency, implausibility (Deane 1991; Kluender 1991; Kluender 1998; Hofmeister & Sag 2010). For instance, using syntactically and semantically richer *wh*-phrases (e.g. *which book* rather than *what*) typically results in better-sounding sentences (Hofmeister & Sag 2010). But this change also significantly improves processing at retrieval sites in long-distance dependencies, as shown by reading time studies using sentences not containing island structures (Hofmeister 2007, 2011). In fact, the defining features of islands have also been linked to processing difficulty. Besides the general cost of filler-gap dependency processing, processing clause boundaries generally leads to lower acceptability ratings and increased processing times (Frazier & Clifton 1989; Kluender & Kutas 1993; Kluender 1998). And as pointed out in Hofmeister & Sag (2010), islands such as *wh*-islands and complex NP islands often create garden-path environments.

An attractive feature, then, of a processing-based view of these phenomena is that general processing principles can be used to account not only for the major acceptability contrasts, but also for the observable variation associated with manipulations that leave the island structure intact.[1] In contrast, analyses couched in terms of grammatical principles generally leave this variation unexplained, or dismiss it as uninteresting peripheral data. Some grammatical accounts of islands,

[1]Phillips (this volume) suggests that this is not a point in favor processing accounts, on the grounds that an account rooted in independently motivated processing constraints fails to explain why 'filler-gap dependencies have a much stronger impact in island contexts than they do in non-island contexts'. Such 'superadditivity' effects, however, are not uncommon in studies of processing difficulty, attention, and multi-tasking: processing costs combine superadditively due to either resource limitations or processing bottlenecks (Pashler 1994). A processing account of island constraints is thus very much at home with these superadditivity effects. Elsewhere, we show that superadditive decrements in acceptability are, in fact, a hallmark of combining significant sentence processing costs (Staum Casasanto et al. 2010).

however, have provided detailed accounts of how such putative counterexamples might be explained (Pesetsky 1987, 2000; Rizzi 1990; Cinque 1990; Manzini 1992). Unfortunately, these accounts almost all stand and fall with ad hoc, unmotivated assumptions and fuzzy definitions that make them either empirically untestable or else mere descriptive labels (Chung 1994; Hofmeister 2007; Hofmeister & Sag 2010; Hofmeister, Jaeger, Arnon, Sag, and Snider in press).

Evaluating the proper role of grammar—particularly syntactic constraints—in islands effects is complicated by the fact that the field presently lacks an explicit, comprehensive theory of grammatical island constraints that also provides a rigorous treatment of the numerous counterexamples that have been observed in the literature. The most recent incarnation of transformational syntax in Chomsky's (1995, 2000) Minimalist Program glosses over island effects, providing no account of how these emerge from grammatical principles. Moreover, accounts that reject processing-based analyses of island effects offer no explicit grammar-based alternative (Phillips in press; Sprouse, Wagers, & Phillips in press). Unfortunately, this makes it effectively impossible to explicitly compare grammatical accounts of the phenomena with processing accounts. This leads us to assess the success of a kind of argument, based on psycholinguistic evidence, contending that island effects must have *some* grammatical foundation without specifying what that foundation is or what is standing on it.

2 Filled gap and plausibility effects (or the lack thereof)

Over the past twenty years, a considerable amount of experimental data has been adduced to support grammar-based accounts of island effects. Among the relevant studies are experiments showing no evidence of dependency formation into island environments. These results have been taken as validation of the claim that grammar forbids attempts to associate items within a syntactic island with items external to that island. In this section, we consider this type of evidence and weigh its relevance to the islands debate.

Evidence of this sort comes from two general paradigms: filled-gap effect and plausibility experiments. The first originates with research by Stowe (1986), which finds elevated reading times at non-empty potential gap positions in sentences involving long-distance dependencies, in contrast to minimally different sentences without a long-distance dependency:

(6) a. The teacher asked what the team laughed about **Greg's older brother** fumbling.
b. The teacher asked if the team laughed about **Greg's older brother** fumbling the ball.

Stowe observed that reading times following the preposition in (6a) shot upward compared to the same region in (6b). These elevated reading times were interpreted as indicating an unsuccessful attempt to integrate the filler (*what*) as the preposition's direct object, a role that is already filled by the overt noun phrase (*Greg's older brother*) in (6a). In short, the parser makes a prediction about where to interpret the *wh*-phrase, but finds counter-evidence that causes it to reconsider the initial analysis. Crucially, these 'filled-gap effects' provide a means for showing that the reader has attempted to create a syntactic dependency. Plausibility studies like Traxler & Pickering (1996) similarly produce signs of dependency creation. Evidence for this comes from signs of difficulty at the lexical head of the dependency when the head and its argument are semantically implausible together:

(7) That's **the pistol/the garage** with which the heartless killer **shot** the hapless man yesterday afternoon.

Since the logic behind plausibility and filled-gap experiments is much the same, we focus here on filled-gap paradigms, but the arguments apply equally to plausibility paradigms.

Through the use of the filled-gap paradigm, it has been shown repeatedly that filled-gap effects do not appear within at least some island environments (Stowe 1986; Bourdages 1992; Pickering, Barton, & Shillcock 1994; McElree & Griffith 1995; Traxler & Pickering 1996; Phillips 2006):

(8) a. The teacher asked what the silly story about Greg's older brother was supposed to mean.
b. The teacher asked if the silly story about Greg's older brother was supposed to mean something.

Similar findings appear in plausibility studies: participants do not show signs of disruption when they reach an implausible lexical head that occurs inside a syntactic island (Traxler & Pickering 1996; Phillips 2006). The absence of these effects in island environments is taken as proof that dependencies into these domains are not attempted, or as Phillips (2006: 813) puts it: "the parser avoids constructing gaps inside islands that cannot be licensed".

However, the existence of grammatically-based island constraints is not the only possible explanation for why filled-gap effects might not occur in island contexts. Wagers and Phillips (2009), for instance, observe that "[t]here is a general concern that many of the environments tested are independently complex to process and, for that reason, dependency construction is difficult or seldom observed in those domains" (p. 423).[2] As noted in Hofmeister and Sag (2010), the absence of filled-gap effects may also follow from expectations regarding the syntactic positions that are more or less likely to host a gap site.

Indeed, much recent work in psycholinguistics suggests that comprehension involves probabilistic, predictive processing (Seidenberg & MacDonald 1999; Crocker & Brants 2000; Hale 2001; Levy 2008). From this perspective, retrieving and integrating filler-phrases involves probabilistic predictions about the appropriate retrieval site, providing a natural account of many contrasts that have been noted in the literature. For example, Pickering and Traxler (2001) suggest that the parser may

[2]Wagers & Phillips (2009) address this issue by examining constructions of putatively comparable complexity: parasitic gap constructions such as *The wines which the gourmets were energetically discussing ___ before slowly sipping ___ during the banquet were rare imports* and across-the-board coordinate extractions like *The wines which the gourmets were energetically discussing ___ and slowly sipping ___ during the banquet were rare imports.* They show that there are no signs of dependency creation in parasitic gaps in adjunct islands using plausibility manipulations. Critically, sentences with across-the-board extractions from verb phrase coordinations do show signs of dependency creation in the second conjunct. Since these sentence types receive similar acceptability scores, this seems to argue that the absence of dependency creation effects in adjunct islands cannot be explained in terms of complexity.

It is conceivable, however, that readers *are* forming a dependency, given that comprehension accuracy did not differ across conditions in their reading time study (Experiment 1) and the overall mean was quite high (88.8%) (Wagers & Phillips, 2009: 409). This suggests that either the method is not sensitive enough to detect dependency formation within the island. Otherwise, there is no explanation for how participants could understand the parasitic gap sentences with such high accuracy. Regardless, there is a more substantial flaw in their argumentation. While comparing acceptability judgments to processing complexity can yield useful insights, they are not perfect images of each other: differences in complexity are not necessarily reflected in acceptability judgments (for evidence of this, see Sprouse (2008)). Consequently, it may well be that sentences containing parasitic gaps in adjuncts differ in complexity from their coordinate counterparts in ways not detectable by acceptability judgments.

simultaneously entertain several candidate gap sites before encountering the true gap in a filler-gap dependency, as in (9) (see also Boland, Tanenhaus, Garnsey, and Carlson (1995)):

(9) a. That's the diver that the coach persuaded a few pupils to watch before the tournament.
b. That's the event that the coach persuaded a few pupils to watch before the tournament.

The possibility that *the diver* in (9a) is the object of *persuade* – plausible on semantic grounds – is at least momentarily considered in real-time, creating a filled-gap effect when the true object *a few pupils* is encountered. In comparison, no filled-gap effect emerges in (9b) at the object phrase, nor does a plausibility effect occur at the verb. Since the direct object analysis becomes implausible in (9b) by the time the verb *persuaded* is processed, the parser appears to forego an attempt to integrate the filler with the direct object position. Instead, the parser turns to the alternative infinitival complement analysis and thus does not experience difficulty at the subsequent NP. In short, the absence of filled-gap effects in (9b) does not license the conclusion that the grammar prohibits making a dependency between the object of the verb and the clefted element. If this is the case, then the absence of evidence for dependency creation in island contexts may similarly follow from extra-grammatical factors.

On the expectation-based view, therefore, filled-gap effects are due to the fact that a particular syntactic environment, e.g. a verb or other subcategorizing element, has a reasonable contextual probability for hosting the gap being processed. Crucially, such effects may be absent when factors (structural or otherwise) conspire to reduce this probability significantly. For instance, if an object gap is strongly predicted in a given sentence context (given local cues and past experience with such constructions), a gap within a preceding complex subject would have a low probability, making signs of a dependency inside the complex subject unlikely. This scenario leaves open the possibility that gaps within complex subjects can be predicted and posited, given sufficient cues that change the parser's expectations.

Our claim that the absence of filled-gap effects may reflect probabilistic biases about where to integrate the filler raises the question of where these biases come from. One possibility is that when cognitive resources are stretched thin or occupied at certain points in sentence processing, memory retrieval that would compound processing difficulty is avoided. In other words, if certain syntactic environments generally impose high processing demands, resource limitations may inhibit linguistic operations such as restoring a filler from memory. This type of explanation leaves open the possibility that the magnitude of the environment's processing difficulty can be reduced, making dependency creation more likely. The idea is simply that the parser is more likely to attempt an operation when it has sufficient time and resources, and less likely to do so when it does not.

As Phillips (in press) acknowledges, the absence of filled-gap effects is ultimately compatible with "reductionist accounts", but "equally compatible with formal grammatical accounts of the constraint." We agree with this assessment. The evidence from the processing of non-islands shows that signs of dependency creation can be absent in syntactic environments even when the parser clearly can make such associations under different circumstances. Thus, the absence of filled-gap effects and their correlates in plausibility paradigms have other viable explanations which do not invoke grammatical island constraints.

3 Effects of Repeated Exposure

Acceptability judgments sometimes rise over the course of the experiment, a phenomenon that has been labeled 'structural facilitation', 'satiation', and 'priming' (Luka & Barsalou 2005; Sprouse 2007a, 2007b, 2009; Francom 2009; Hofmeister et al. in press). Whatever the name, this phenomenon is potentially of importance in the islands debate: if a particular kind of ungrammatical sentences lack a coherent representation, then no matter how many times such a sentence is seen, judgments should remain consistently low (Sprouse 2007a). By contrast, sentences judged to be unacceptable due to processing difficulty ought to become easier to process with familiarity, and consequently receive higher judgments of acceptability (see Wexler and Culicover (1980)). Given these assumptions, identifying the source of island effects theoretically becomes as simple as seeing whether judgments of island-violations rise throughout the course of an experiment.

Along these lines, Sprouse (2007a: 123) reports that acceptability judgments for a variety of island types do not increase significantly with repeated exposure and thus concludes that "there are no syntactic priming effects on acceptability for ungrammatical structures," referring to island-violating structures. Under a theory of categorical grammaticality, these results are expected if islands lack a licit representation because they are ungrammatical. Thus, the absence of repeated exposure effects lends credence to the idea that the unacceptability of islands stems from grammatical constraints. Sprouse (2007a: 123) further suggests that "that asymmetric extra-grammatical effects on acceptability may be a useful tool for identifying the grammatical status of structures that are neither clearly grammatical or clearly ungrammatical."

However, the overall evidence regarding the effects of exposure on judgments for island-violating sentences is rather mixed. Snyder (2000) found that judgments for several types of island-violations rise with exposure, while Sprouse (2007b) discovered that these changes disappear after equalizing the number of sentences that elicit acceptable and unacceptable judgments. But other researchers have replicated the finding that judgments for island-violating sentences increase with repetition (Francom 2009; Hofmeister & Sag 2010) and others have shown that reading times for island-violating structures decrease with exposure (Braze 2002; Hofmeister et al. in press).

These differences with respect to exposure may potentially be explained in terms of different items, different island types, different acceptability scales, different presentation methods, etc. But, besides the conflicting results, there is a more serious problem with interpreting past results on repeated exposure: it has never been established that judgments for difficult grammatical sentences increase with repeated exposure, but those for ungrammatical sentences do not. Before interpreting repeated exposure data with respect to islands, we need to know how exposure affects uncontroversial cases. For starters, the findings from Luka & Barsalou (2005) indicate that sentences of 'moderate grammaticality' do increase with exposure, e.g. *Sam recites poems as well as playing the piano*. Minimally, this implies that at least some sentences with structural abnormalities become better with familiarity. It also raises the possibility that even sentences that are uncontroversially ungrammatical might improve with exposure.

To investigate this, we conducted an acceptability experiment studying uncontroversial sources of unacceptability, using the Thermometer Judgment (TJ) methodology described in Featherston (2008). In TJ studies, participants judge the test items compared to two reference sentences; however, they are not asked to assess magnitudes of acceptability differences as in ME. One of the reference sentences is quite good and the other quite bad, and we follow Featherston (2008) in

assigning these sentences the arbitrary values 20 and 30. For all of the experiments described here, we used the same reference sentences:

(10) a. The way that the project was approaching to the deadline everyone wondered. [= 20]
b. The architect told his assistant to bring the new plans to the foreman's office. [= 30]

Sentences were shown word-by-word at a fixed rate of presentation in the center of the screen (250 ms + (33.33 ms $*$ the number of characters in the word)), so that longer words remained visible for longer. We used word-by-word presentation (not full sentence presentation) in order to prevent excessive introspection about the test sentences; the presentation was auto-paced (rather than self-paced) so that there would be no differences in how long each participant studied a given stimulus.

For the statistical analysis, we first log-transformed all acceptability scores and then computed z-scores for each subject on the basis of their judgments for all experimental items, including fillers. We then excluded data points with z-scores more than 2.5 standard deviations from the subject's mean. Finally, we used linear mixed-effects (LME) models for data analysis. All predictors which provided the basis for higher order variables (interactions) were centered to reduce effects of collinearity. In all the experiments we discuss, we used a maximal random effect structure for the LME model with random intercepts for participants and items, and by-participant and by-item random slopes for any fixed effect factors. Log-likelihood ratio tests, where the deviance of a model containing a fixed effect is compared to that of an otherwise identical model without it, were used to obtain p-values (Pinheiro & Bates 2000; Baayen 2007).

In the first experiment, 28 Stanford University community members (all native English speakers) saw 12 items that had multiple words moved to illicit positions, as in (11) (materials can be found at the first author's website):

(11) Iran has gun-control strict laws that bar private citizens carrying from firearms.

Any adequate grammar would treat sentences like (11) as ungrammatical. Nevertheless, the intended message of such sentences is not difficult to understand.

As shown by Figure 1, the jumbled items received significantly higher ratings, the later they appeared with respect to each other in the experiment ($\beta = .051$, $SE = .015$, $t = 3.38$, $p = .003$). This consequently means that grammaticality is not a precondition for observing the facilitation effects of repetition. So, if ungrammatical strings can in fact prime, then even grammar-based accounts of island phenomena should predict that judgments for island violations ought to improve with increased exposure.

Why do these ungrammatical items get better with repeated exposure? Francom (2009) hypothesizes that the extent of exposure-related effects on acceptability judgments depends upon more than just grammaticality. Drawing on research by Freedman and Forster (1985) and by Maclay and Sleator (1960), Francom notes that subjects provide similar ratings whether they are asked to judge according to grammaticality or meaningfulness. Accordingly, acceptability judgments for items like (11) may increase with exposure because participants rely less and less on word order cues to form meaningful interpretations. Interpretability therefore appears to have a major role in whether some structure becomes more acceptable with repetition.

As to why island violations become more acceptable with exposure in some cases and not others, we can only conjecture based on the available evidence. Some experimental materials may

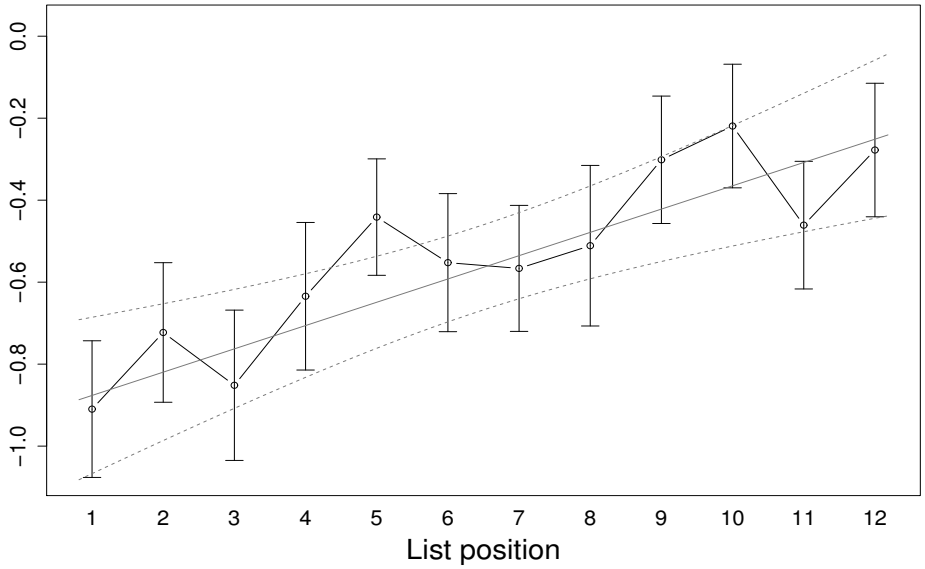

Figure 1: Mean acceptability z-score by list position, according to a linear regression model; dotted lines indicating 95% confidence intervals.

introduce excessive processing difficulty. That is, sentences with massive processing difficulty may never get better over time because whatever facilitating effects may come from increased exposure, they cannot compensate for the overall difficulty of the items. Evidence for this hypothesis comes from an experiment using twelve doubly-nested, center-embedding structures, such as (12), with 24 University of California - San Diego undergraduates as participants:

(12) The cheerleader who the quarterback who was on the team dated snubbed the teammates although this hurt her reputation.

There were no signs that repeated exposure improved judgments for these items ($\beta = .012$, $SE = .020$, $t = 0.60$, $p = .56$). While this is a null result, it suggests that repetition effects may not always be observed in the presence of major hurdles to successful processing, even if the sentences are 'grammatical', as these are standardly assumed to be.

Fundamentally, we are left with the conclusion that the field lacks a firm understanding of why some structures get better with exposure and not others. A full understanding will likely involve considering a complex interplay of sentence interpretability, processing difficulty, and the relative ease with which structural anomalies can be identified and corrected. In any case, the critical assumption that judgments of ungrammatical sentences will not improve with exposure, while judgments of sentences that are merely difficult to process will improve does not seem to be accurate. Hence, data from repeated exposure studies cannot yet inform us of the respective weights of grammar and processing in producing island effects.

4 The relationship between working memory capacity and judgments of acceptability

A third type of argument that has been offered in favor of grammar-based accounts of islands concerns the relationship between acceptability judgments and individual differences in working memory (WM) (Sprouse et al. in press). This line of argumentation proposes that processing-based accounts of island phenomena predict that island sensitivity should vary with processing resource capacity. Consequently, WM measures, such as n-back and serial recall tasks, should correlate with the strength of island effects. Sprouse et al. (in press) tested this prediction by conducting a large-scale study with four different types of island-violations: *wh*-islands, subject islands, adjunct islands, and complex NP islands. According to their findings, individual differences in performance on the memory tasks accounted for a small amount of the overall variance (0%-2%). These findings seem to run counter to the predictions of a processing-based perspective on island effects, lending indirect support to the conclusion that island effects arise from grammatical constraints.

One problem with using correlations with WM to infer whether a phenomenon has a grammatical or processing basis is that there are a variety of means for measuring WM, as Sprouse et al. acknowledge. So a possible explanation for their null result is that the specific choice of working measures—n-back tasks and serial recall—tap memory and cognitive resources of minimal importance to island processing. In defense of their means for measuring WM, Sprouse et al. contend that most measures of WM are highly correlated, citing Conway et al. (2005). The clear suggestion is that the choice of another WM measure would lead to a similar null result. They further claim that it is improbable that a WM measure exists which reflects sensitivity to islands, but which does not correlate with performance on either the serial recall or the n-back task.

However, Conway et al. (2005) never state that most measures of WM are highly correlated. Rather, they provide evidence that WM scores from reading span, counting span, and operation span tasks (not n-back tasks or serial recall tasks) are highly correlated with one another, as well as performance on a range of higher order cognitive tasks, such as reading comprehension. Conway et al. (2005) also explicitly note that tasks like the n-back task "present quite different cognitive demands" from WM span tasks and that "the n-back task may be a more appropriate indicator of the construct measured by STMC [short term memory capacity], rather than by WMC tasks" [pp. 780-781]. This is not to say that we advocate WM span tasks as the best means for predicting performance in sentence processing; we are simply reiterating a point made by Conway et al. (2005): little data currently exists that can help us evaluate how different measures of WM reflect underlying cognitive abilities. Conway et al. (2005) also note that correlations between span tasks range from .40 to .60, "suggesting that they are indeed tapping some common process or ability but also suggesting that they are not identical" [p. 780]. In general, it is inaccurate to assume that because two WM measures are highly correlated, they will necessarily reflect similar sensitivities to islands. So, even if there *were* evidence that all other sensible measures of WM are correlated with at least one of the tasks used by Sprouse et al, this would still not license the conclusion that similar results would obtain with another choice of WM task.

As in studies looking at effects of repeated exposure, however, an even more serious problem prevents us from drawing conclusions about the islands debate from this type of evidence – this work hinges upon on another poorly understood aspect of acceptability judgments. The question of how individual differences in WM relate to acceptability judgments has never been systematically investigated. While it is certainly plausible that individuals who score higher on memory tasks

should exhibit less sensitivity to processing manipulations, this hypothesis requires validation with uncontroversial data before it can be used to interpret contentious issues.

In another experiment, we sought to determine how WM differences predict judgments for sentences that are grammatical but extremely difficult to process. To assess WM capacity, we employed a reading span task along the lines of that described in Daneman and Carpenter (1980). In such a task, participants read a series of sentences (2, 3, 4, or 5) and after reading all the sentences in each series, subjects attempt to recall the last word from each sentence. Following suggestions from Conway et al. (2005), we employed a partial-credit unit scoring method that corrects for the problem that traditional all-or-nothing scoring discards a large amount of potentially useful data.

The materials ($n = 24$) for the experiment varied in two respects: (1) the distance between a *wh*-phrase and its subcategorizing head and (2) the presence of either a subject or object relative clause.

(13) a. [SHORT-SRC] Someone figured out which politician wrote that Robert bribed a reporter that trusted Nancy without thinking about it.

b. [SHORT-ORC] Someone figured out which politician wrote that Robert bribed a reporter that Nancy trusted without thinking about it.

c. [LONG-SRC] Someone figured out which politician a reporter that trusted Nancy wrote that Robert bribed without thinking about it.

d. [LONG-ORC] Someone figured out which politician a reporter that Nancy trusted wrote that Robert bribed without thinking about it.

28 Stanford University undergraduates, naïve to the purposes of the study, provided the judgment data. The LME model for analysis of the results included fixed effect factors for the two processing-related factors, their interaction, reading span score and its interaction with the two manipulations and their interaction. As Figure 2 depicts, higher WM scores are associated with higher acceptability z-scores in the two relatively easy conditions with short dependencies, (13a) & (13b). But in the more difficult conditions with long dependencies, (13c) & (13d), there is no evidence of relationship between WM scores and judgments. This pattern accounts for the significant interaction between WM and dependency length in the LME model of acceptability judgments ($\beta = -.090$, $SE = .046$, $t = -1.98$, $p = .049$). In addition, the R^2 between z-score and WM score in the conditions with long dependencies is just .012. Such statistics would thus not reveal a relationship between judgments and individual cognitive differences, despite the clear fact that it is the processing difficulty of these items that yields the low acceptability ratings.

What these results tell us is that significant processing costs do not guarantee that individuals with high WM capacities will rate the relevant items as more acceptable than individuals with relatively low WM capacities. Perhaps items with long dependencies produce such extreme processing difficulty that individual differences make insufficient difference. Items with short dependencies, in contrast, are not so difficult to process, leaving room for differences due to individual variation to emerge. These findings raise the possibility that the lack of WM effects in some island datasets may follow from the extreme difficulty of such items. True or not, the above data from an uncontroversial case of processing difficulty offers telling evidence that sentences with processing difficulties are not always rated as more acceptable by individuals with high WM scores.

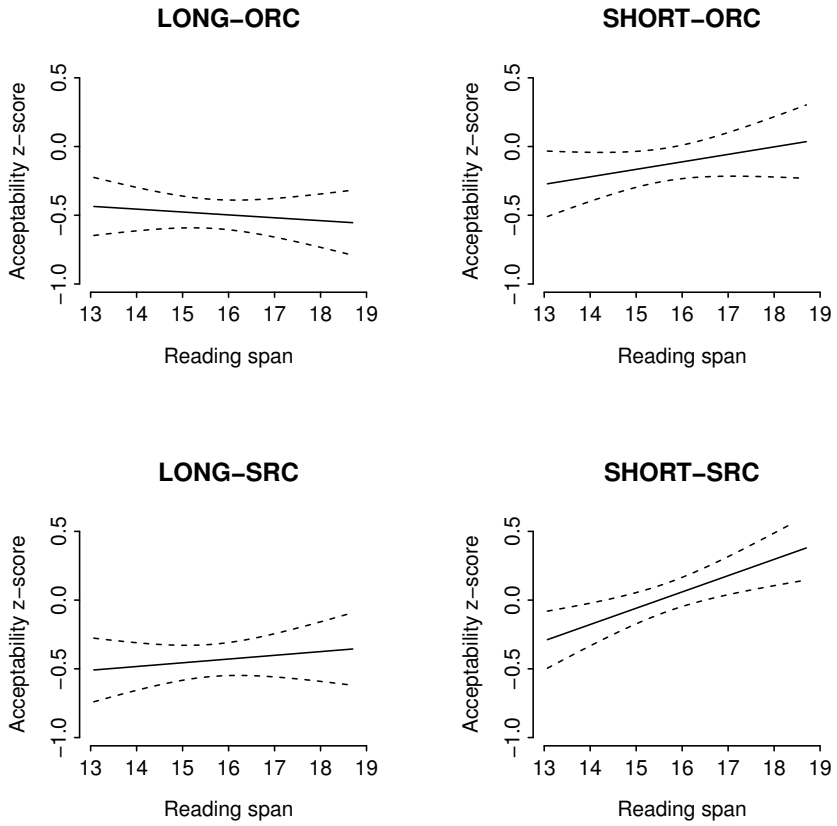

Figure 2: Effects of reading span on acceptability z-score for sentence types with varying degrees of difficulty

In sum, WM estimates and processing difficulty are not straightforwardly related in acceptability judgment tasks. Coupled with the possibility that unsuitable memory estimates may have been used, the Sprouse et al. (in press) findings do not undermine processing-based accounts, nor are they uniquely consistent with grammatical analyses of island effects. However, WM estimates may still have a role to play in functional vs. formal debates of acceptability contrasts. Staum Casasanto et al. (2010) present acceptability data showing that participants with higher reading spans judge ungrammatical sentences as being worse than their low-span counterparts do, yet they judge difficult (but not extremely difficult) sentences as being better than participants with lower reading span scores. These patterns suggest that when positive linear relationships between acceptability and WM measures are observed, functional pressures play a role in the acceptability contrasts. In short, WM measures have the potential to identify functional factors at work in acceptability judgments, but the absence of correlations or linear relationship cannot eliminate the possibility that processing difficulty is the primary explanation for island effects.

5 The argument from overgeneration

Phillips (2006, in press) provides yet another argument against accounts that stress the role of processing in island effects. In essence, the overgeneration argument states that if an individual can demonstrate the capability to process and understand some construction, but nevertheless judges it to be unacceptable, then this argues for the role of grammatical constraints:

> " . . . it is hard in such cases to argue that the overgenerated forms are ruled out by limitations of the human representation-building capacity, since we have evidence that speakers are able to construct exactly those representations. It therefore becomes more likely that the constraint that normally blocks the overgenerated forms is a formal grammatical constraint." [Phillips, in press]

As an example, Phillips points to his work on subject islands and parasitic gap constructions (Phillips, 2006). Acceptability data confirmed long-standing claims about the unacceptability of filler-gap dependencies into subject NPs, as in (14a) below, compared to a minimally different sentence with an object gap (14b):

(14) a. Which parts did [the attempt to repair ___] ultimately damage the car?
b. Which parts did [the attempt to repair the car] ultimately damage ___?
c. Which parts did [the attempt to repair ___] ultimately damage ___?

Critically, this acceptability evidence accompanies reading-time data showing signs of attempted dependency formation inside infinitival subject NPs. Phillips infers that a dependency can be made into infinitival subject NPs because subject-internal gaps are only highly acceptable in the context of infinitival parasitic gap constructions like (14c). In non-finite subject NP contexts like (15), however, parasitic gap constructions elicit low acceptability ratings and signs of dependency formation via plausibility manipulations are absent:

(15) The outspoken environmentalist worked to investigate what the local campaign that preserved ___ had harmed ___.

It is concluded that the parser only selectively violates subject island constraints when it can potentially lead to an acceptable parasitic gap construction. Since subjects rate island-violating sentences like (14a) as relatively unacceptable despite being to able to construct the dependency as evidence by the plausibility effects, Phillips concludes that the unacceptability cannot be due to resource limitations that affect dependency processing.

Key to this line of argumentation is an assumption attributed to processing-based accounts of island effects: "if an island constraint is ultimately a consequence of limits on on-line structure building, then we should not find that the constraint is spontaneously violated during on-line structure building" (Phillips, in press). That is, no signs of dependency formation should be evident in island contexts, if processing accounts are true, because the excessive processing cost and lack of cognitive resources should cut off any opportunity to make the dependency. Indeed, this prediction does follow from some older processing-based accounts of island effects (Berwick & Weinberg 1984; Pritchett 1991), which assume that architectural properties of the parser impose functional limitations that prevent dependencies from crossing certain clausal boundaries.

But this assumption is not a necessary property of a processing-based account of acceptability contrasts and in particular it plays no role in limitation-based models of the sort we advocate. While excessive processing demands may sometimes lead to a processing breakdown or retrieval failure, processing costs will often simply create difficulty, without causing a total breakdown of parsing processes. This point bears repeating: whether a parsing operation fails or succeeds depends on *how* difficult it is, not just *whether* or not it is difficult, and thus predictions of difficulty do not equal predictions of parsing failure. In this vein, processing accounts such as Hofmeister & Sag (2010) emphasize that dependency formation inside islands is difficult, not impossible or prohibited. Indeed, the variable cost of retrieval is a major driver in the overall processing difficulty of an island-violating utterance, according to Hofmeister & Sag (2010), which would be impossible if dependency formation never occurred.

It is worth pointing out yet again that the basic assumptions behind the overgeneration argument would be better supported by examining uncontroversial cases. The argument assumes that wherever it is possible to construct a representation, low judgments of acceptability must be explained in terms of the violation of grammatical principles, rather than processing difficulty. However, it is not at all clear that this conclusion is warranted. Sentences such as *The administrator who the nurse who was from the clinic supervised scolded the medic while a patient was brought into the emergency room* receive low acceptability ratings compared to semantically equivalent, right-branching structures such as *The nurse from the clinic supervised the administrator who scolded the medic while a patient was brought into the emergency room* (Staum Casasanto et al. 2010). The comprehension question accuracies for the different sentences were nearly identical (80.7% vs. 81.2%), however, suggesting participants can ultimately represent and understand the nested sentence despite the low ratings given. Thus, a sentence's processing costs may give rise to low acceptability ratings and yet simultaneously allow a comprehender to construct a coherent representation. In such a case, the overgeneration argument would lead us to inaccurately conclude that grammatical principles produce the low acceptability judgments. In light of such points, it seems critical to test the overgeneration argument's application to various uncontroversial cases before applying it to islands.

Overall, the argument based on overgeneration applies only to a limited set of processing accounts that assume dependencies into island contexts are uniformly prohibited by architectural constraints. For reasons discussed above, processing accounts such as ours (and those of Kluender

(1991, 1992, 1998) and Kluender & Kutas (1993)) do not predict that speakers should categorically attempt to create dependencies or that they should not; instead, they predict that comprehenders are more likely to attempt to create dependencies, the easier they are to create. Even if they did apply, the basic claim of the overgeneration argument has not withstood tests from less contentious data. Hence, while we endorse efforts to find discriminating evidence that separates grammatical from processing-based accounts, the overgeneration argument cannot help to distinguish these positions.

6 Criticisms of processing accounts of island effects

Alongside attempts to find positive evidence for a grammatical source of island effects, proponents of such an account also argue for it indirectly by pointing to challenges for processing accounts, many of which are discussed in this volume. Although we cannot respond to all of these challenges within the space of this chapter, we take up a few of the issues raised by Phillips (this volume) to demonstrate that seeming problems for a processing account of island effects are not as problematic as they are made out to be.

One challenge, related to the data covered in the previous section, is to explain why signs of dependency creation appear in non-finite subject islands but not (some) other island contexts such as finite subject islands.[3] Phillips (this volume) presents this as problematic for a processing account, because "if difficulty does not entail the absence of active gap-creation effects, then we should expect to see active gap-creation effects in other types of islands". But the contrast between non-finite subject islands and finite ones is actually predicted by a processing perspective. Gibson (1998, p. 12) suggests that only tensed verbs contribute to locality costs because only they introduce discourse referents that are tracked in the discourse. Thus, non-finite clauses will be easier on average than finite clauses to process. Accordingly, stronger evidence for dependency formation inside non-finite islands should generally be easier to come by.

The existence of cross-linguistic variation is often raised as a further challenge for processing accounts of island effects. Some languages show remarkably little island sensitivity, allowing structural dependencies that lead to sharp drops in acceptability in comparable sentences of English, and some languages prohibit long-distance dependencies that English permits. Formal grammatical accounts reckon with such variation in the following way: "When faced with a novel island effect, or a novel case of cross-language variation, it is easy to simply add a constraint or parameter to a grammatical account", whereas "[r]esource-based reductionist accounts make strong predictions about cross-language uniformity" (Phillips, this volume). This statement is accurate insofar as speakers of all languages are assumed to have roughly identical cognitive constraints and sentence processing architectures. The statement is inaccurate, however, in that accounts such as ours do not assume that the costs associated with processing comparable structures are equivalent across languages. There is no reason to assume that the cost of, say, forming a dependency between a *wh*-word and an element inside a *wh*-island is the same from language to language. One language

[3]Phillips acknowledges that the parser can form a dependency into islands when forced to do so, e.g. *wh*-islands, but argues that representability and well-formedness are two different things. From this view, some islands are representable but ill-formed. It is not clear, though, how or whether representability and well-formedness can be empirically separated via independent evidence.

may utilize features and cues to aid retrieval that another does not, e.g. resumptives.[4]

Such an analysis of cross-linguistic differences in the case of Superiority violations (e.g. *What did who buy?*) appears in Arnon, Snider, Hofmeister, Jaeger, & Sag (2006) who argue that the strength of Superiority effects varies with the availability of case cues. Arnon et al. observe that Superiority effects are larger in English than in German, while they appear to be entirely absent in Russian. However, these languages also differ in the availability of case marking cues that can aid in the processing of *wh*-dependencies. While English has no case marking on *wh*-phrases, case marking exists for both German and Russian *wh*-phrases. Moreover, three out of seven case markings are ambiguous between nominative and accusative in German, while only three out of ten are ambiguous in Russian. A corpus study verified that unambiguous *wh*-words were three times more frequent in Russian than German. Arnon et al. thus conclude that while there is a universal difficulty connected to dependency processing, "the extent of the difficulty will depend on speakers ability to draw on additional information".

Nonetheless, it remains a challenge for processing accounts to identify what information varies across languages along with island sensitivity. Although we do not claim at present to have a full account of what the relevant variables are for each type of island, the absence of detailed accounts should nevertheless not be taken to mean that processing accounts are incompatible with the existence of cross-linguistic variation. While all languages face the same fundamental processing problems, languages exhibit a diverse array of tools to solve those problems, and so we should not expect a priori that any syntactic operation will incur the same cost across languages.

The last issue we briefly consider concerns learning. The problem, as described by Phillips, is that, if the absence or rarity of island-violating sentences in the child's input is due to resource limitations, the child has to somehow know "that he should avoid drawing conclusions about the target grammar from the gap". This implies that there is a danger of a child positing a grammatical ban for structures that are not represented sufficiently in the input. But this danger then would apply to structures like multiply center-embedded sentences and sentences with *wh*-dependencies that span a thousand words and other structures which are so difficult to process that they are rarely, if ever, attested in natural speech. From the processing perspective, the learning problem for islands is no different than it is for other hard-to-process structures: there is nothing to learn. Recall what is being explained here – that dependencies into islands sound unacceptable. On the analysis that this unacceptability stems from processing constraints, the child does not need any feedback or guidance to learn this.

All this is not to say that processing accounts of islands are without weaknesses. One is that different islands seem to pose different difficulties, preventing a uniform analysis for island effects. Some phenomena like parasitic gaps, and why they improve subject islands, are also currently without a detailed explanation. Furthermore, processing accounts of islands, still in their infancy, have not yet provided a precise model of the relevant factors, their costs and the way these costs combine. Moreover, as noted above, we do not yet know what all the factors are that lead to cross-linguistic variability in island sensitivity. But we contend that many of these weaknesses reflect

[4]Wagers (this volume) expresses skepticism that resumptives facilitate dependency processing, on the basis of acceptability data showing that resumptives are not preferred to gaps in non-island contexts. However, Hofmeister & Norcliffe (to appear) demonstrate in a self-paced reading experiment that the acceptability data by itself is misleading, and that resumptives do signicantly facilitate comprehension in long-distance dependencies that do not violate islands.

the early stages of the research on processing difficulty within islands, rather than some inherent limitation of the approach.

7 Advancing the understanding of island effects

Our efforts here have been largely directed toward addressing the adequacy of accounts favoring grammar-based interpretations of island effects. Along the way, we have observed that the lack of systematic studies based on undisputed instances of processing-induced difficulty and/or violations of grammatical constraints makes it extremely difficult to draw inferences about datasets like those central to the islands debate. We also noted that numerous accounts of islands invoke grammar as the root cause of the effects, without specifying what specific grammatical constraints are being assumed. At present there exists no such account that is both empirically adequate and free from capricious and arbitrary assumptions.

Furthermore, many of the main arguments in favor of grammatical accounts (or against processing-based accounts) are either based on faulty assumptions or else dependent on data amenable to alternative explanations, and the supporting evidence often takes the form of null results that are hard to evaluate. Even if statistical tests and replication verify that the null results are extremely unlikely to be spurious, this does not remove the possibility that some experimental confound lies behind the null result. Given these points, the question arises: what kind of evidence *would* support an account of island effects rooted in formal constraints?

Unfortunately, it remains unclear what kinds of evidence will be helpful in determining whether island effects have a grammatical basis. A common theme across our discussions of the arguments for grammatical accounts is that they rely on measures that have not yet been tested on uncontroversial examples of formal and functional constraints. While we have undertaken a few experiments to remedy this problem, answering how processing and grammatical constraints differ requires a larger body of research comparing specific examples of constraints in order to license generalizations and understand their limits. This is a considerable undertaking, but having a clear idea of how grammar and processing influence the dependent measures we use is a critical prerequisite for using these measures to draw conclusions about the basis of controversial contrasts. One possible outcome is that there will not be clear behavioral differences that distinguish formal from functional constraints. If it does not appear to be possible to operationalize this distinction, the logical conclusion may be that this dichotomy is a construct that does not necessarily exist in the minds of language users.

Another potentially fruitful way of pursuing this debate is to classify the kinds of constraint interactions, without resolving their basis. Whether functional and formal constraints are cleanly separated in the mind or not, constraints that are more closely related to one another should share more properties. These properties may include things like how constraints combine in acceptability judgment tasks, how much variation occurs across individuals, how stable judgments are, how judgments relate to other behavioral measures such as reading times, eye movements, and reaction times, etc. By considering a variety of phenomena and examining how their properties differ from one another, we can hope to classify island effects with respect to other, perhaps better understood effects. For example, if island effects largely reflect formal constraints on hierarchical structure, then they ought to resemble uncontroversial constraints on hierarchical structure in observable ways. This is not meant to imply that constructions related by the source of their (un)acceptability should display identical properties. However, a sufficiently large set of data built from non-controversial

cases should help us to define the range of variation and to notice patterns that characterize effects that are conventionally ascribed to formal and functional constraints.

One might also reasonably ask how a processing account of islands could be disproved. Here, we believe the answer is more straightforward. A processing-based account makes clear predictions that island processing should be accompanied by signs of difficulty. Thus, if it was convincingly shown that island structures lead to *easier* processing than minimally different non-island structures, this would contradict the predictions of a processing-based theory. Additionally, processing-based explanations of island effects would be weakened (although not disproved) by showing that there are no reasons to expect that particular island effects should be linked to general processing principles. And finally, such an account would be further weakened by research demonstrating that behavioral and electrophysiological responses to islands contrast with responses to constructions with high processing difficulty.

We believe the best strategy for moving forward in the islands debate involves developing a better understanding of how a range of construction types pattern, not just with respect to acceptability judgments, but dependent measures of all sorts. This will take time, but it has the potential to help us resolve many questions about the roles of competence and performance in shaping the facts we take as evidence for building syntactic theories. Progress on this issue is not optional; without it, we cannot hope to come, as a field, to an informed understanding of the nature of grammar and how it is embedded in a model of sentence processing.

References

Arnon, Inbal, Snider, Neal, Hofmeistee, Philip, Jaeger, T. Florian, and Sag, Ivan A. 2006. 'Cross-linguistic variation in a processing account: The case of multiple wh-questions', *Proceedings of 32 Annual Meeting of the Berkeley Linguistics Society.*

Baayen, R. Harald 2007. *Analyzing linguistic data: A practical introduction to statistics.* Cambridge, UK: Cambridge University Press.

Berwick, Robert and Weinberg, Amy S. 1984. *The grammatical basis of linguistic performance.* Cambridge, MA: MIT Press.

Boland, Julie E., Tanenhaus, Michael K., Garnsey, Susan M., and Carlson, Greg N. 1995. 'Verb argument structure in parsing and interpretation: Evidence from wh-questions', *Journal of Memory and Language*, *34*, 774–806.

Bourdages, Johanne S. 1992. 'Parsing complex NPs in French', in H. Goodluck and M. Rochemont (eds.), *Island constraints: Theory, acquisition and processing*, pp. 61–87. Dordrecht: Kluwer Academic Publishers.

Braze, Forrest D. 2002. *Grammaticality, acceptability, and sentence processing: A psycholinguistic study.* Unpublished doctoral dissertation, University of Connecticut, Storrs, CT.

Chomsky, Noam 1962. 'The logical basis of linguistic theory', in M. Halle (ed.), *Preprints of the papers for the Ninth International Congress of Linguists, August 27–31, 1962*, pp. 509–574.

Chomsky, Noam 1973. 'Conditions on transformations', in S. Anderson and P. Kiparsky (eds.), *A Festschrift for Morris Halle*, pp. 232-86. New York: Holt, Reinhart & Winston.

Chomsky, Noam 1977. 'On *wh*-movement', in P. Culicover, T. Wasow, and A. Akmajian (eds.), *Formal syntax.* New York: Academic Press.

Chomsky, Noam 1981. *Lectures on Government and Binding.* Dordrecht: Foris.

Chomsky, Noam 1986. *Barriers*. Cambridge: MIT Press.

Chomsky, Noam 1995. *The Minimalist Program*. Cambridge: MIT Press.

Chomsky, Noam 2000. 'Minimalist Inquiries', in R. Martin, D. Michaels, and J. Uriagereka (eds.), *Step by step*, pp. 89–155. Cambridge: MIT Press.

Chomsky, Noam and Miller, George 1963. 'Introduction to the formal analysis of natural languages', in R. Luce, R. Bush, and E. Galanter (eds.), *Handbook of mathematical psychology*, pp. 269-321. New York: Wiley.

Chung, Sandra 1994. '*Wh*-agreement and "referentiality" in Chamorro', *Linguistic Inquiry*, *25*, 1-44.

Cinque, Guglielmo 1990. *Types of A'-dependencies*. Cambridge: MIT Press.

Conway, Andrew, Kane, Michael, Bunting, Micheal, Hambrick, D. Zach, Wilhelm, Oliver, and Engle, Randall 2005. 'Working memory span tasks: A methodological review and user's guide', *Psychonomic Bulletin and Review*, *12*, 769–786.

Cowart, Wayne 1997. *Experimental syntax: Applying objective methods in sentence judgments*. Thousand Oaks, CA: Sage Publications.

Crocker, Matthew and Brants, Thorsten 2000. 'Wide-coverage probabilistic sentence processing', *Journal of Psycholinguistic Research*, *29*, 647–669.

Daneman, Meredyth and Carpenter, Patricia A. 1980. 'Individual differences in working memory and reading', *Journal of Verbal Learning & Verbal Behavior*, *19*, 450-466.

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

Featherston, Sam 2008. 'Thermometer judgments as linguistic evidence', in M. Claudia and A. Rothe (eds.), *Was ist linguistische Evidenz?* Aachen: Shaker Verlag.

Fodor, Janet 1978. 'Parsing strategies and constraints on transformations', *Linguistic Inquiry*, *9*, 427–473.

Francom, Jerid 2009. *Experimental syntax: Exploring the effect of repeated exposure to anomalous syntactic structure–evidence from rating and reading tasks*. Unpublished doctoral dissertation, University of Arizona.

Frazier, Lyn and Clifton, Charles, Jr. 1989. 'Successive cyclicity in the grammar and the parser', *Language and Cognitive Processes*, *4*, 93–126.

Freedman, Susan and Forster, Kenneth 1985. 'The psychological status of overgenerated sentences', *Cognition*, *19*, 101–131.

Georgopoulos, Carol 1985. 'Variables in Palauan syntax', *Natural Language & Linguistic Theory*, *3*, 59–94.

Georgopoulos, Carol 1991. *Syntactic Variables: Resumptive Pronouns and A' Binding in Palauan*. Dordrecht: Kluwer Academic Publishers.

Gibson, E. 1998. 'Linguistic complexity: Locality of syntactic dependencies', *Cognition*, *68*, 1–76.

Hale, John 2001. 'A probablistic Earley parser as a psycholinguistic model', in *Proceedings of the Second Meeting of the North American Chapter of the Association for Computational Linguistics*, pp. 1–8. Morristown, NJ: Association for Computational Linguistics.

Hofmeister, Philip 2007. *Representational complexity and memory retrieval in language comprehension*. Unpublished doctoral dissertation, Stanford University, Stanford, CA.

Hofmeister, Philip 2011. 'Representational complexity and memory retrieval in language comprehension', *Language and Cognitive Processes*, *26*, 376-405.

Hofmeister, Philip, Jaeger, T. Florian, Arnon, Inbal, Sag, Ivan A., and Snider, Neal in press. 'The source ambiguity problem: Distinguishing the effects of grammar and processing on acceptability judgments', *Language and Cognitive Processes*.

Hofmeister, Philip, Jaeger, T. Florian, Sag, Ivan A., Arnon, Inbal, and Snider, Neal 2007. 'Locality and accessibility in *wh*-questions', in S. Featherston and W. Sternefeld (eds.), *Roots: Linguistics in search of its evidential base*, pp. 185–206. Berlin: Mouton de Gruyter.

Hofmeister, Philip and Norcliffe, Elisabeth to appear. 'Does resumption facilitate sentence comprehension?'.

Hofmeister, Philip and Sag, Ivan A. 2010. 'Cognitive constraints and island effects', *Language*, *86*, 366–415.

Kluender, Robert 1991. *Cognitive constraints on variables in syntax*. Unpublished doctoral dissertation, University of California, San Diego, La Jolla, CA.

Kluender, Robert 1992. 'Deriving islands constraints from principles of predication', in H. Goodluck and M. Rochemont (eds.), *Island constraints: Theory, acquisition and processing*, pp. 223-258. Dordrecht: Kluwer.

Kluender, Robert 1998. 'On the distinction between strong and weak islands: a processing perspective', in P. Culicover and L. McNally (eds.), *Syntax and semantics 29: The limits of syntax*, pp. 241-279. San Diego, CA: Academic Press.

Kluender, Robert and Kutas, Marta 1993. 'Subjacency as a processing phenomenon', *Language and Cognitive Processes*, *8*, 573–633.

Levy, Roger 2008. 'Expectation-based syntactic comprehension', *Cognition*, *106*, 1126–1177.

Luka, Barbara and Barsalou, Lawrence 2005. 'Structural facilitation: Mere exposure effects for grammatical acceptability as evidence for syntactic priming in comprehension', *Journal of Memory and Language*, *52*, 436–459.

Maclay, Howard and Sleator, Mary D. 1960. 'Responses to language: Judgements of grammaticalness', *Journal of American Linguistics*, *26*, 275–282.

Manzini, M. Rita 1992. *Locality: A theory and some of its empirical consequences*. Cambridge: MIT Press.

McElree, Brian and Griffith, Teresa 1995. 'Syntactic and thematic processing in sentence comprehension: evidence for a temporal dissociation', *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *21*, 134–157.

Miller, George A. and Chomsky, Noam 1963. 'Finitary models of language users', in R. D. Luce, R. R. Bush, and E. Galanter (eds.), *Handbook of mathematical psychology* Vol. 2, pp. 419-492. New York: Wiley.

Pashler, Hal 1994. 'Dual-task interference in simple tasks: Data and theory', *Psychological Bulletin*, *116*, 220-244.

Pesetsky, David 1987. '*Wh*-in-situ: movement and unselective binding', in E. Reuland and A. ter Meulen (eds.), *The representation of (in)definiteness*, pp. 98–129. Cambridge: MIT Press.

Pesetsky, David 2000. *Phrasal movement and its kin*. Cambridge: MIT Press.

Phillips, Colin 2006. 'The real-time status of island phenomena', *Language*, *82*, 795–823.

Phillips, Colin in press. 'Some arguments and non-arguments for reductionist arguments of syntactic phenomena', *Language and Cognitive Processes*.

Phillips, Colin this volume. 'On the nature of island constraints. I: Language processing and reductionist accounts'.

Pickering, Martin, Barton, Stephen, and Shillcock, Richard 1994. 'Unbounded dependencies, island constraints and processing complexity', in C. Clifton Jr., L. Frazier, and K. Rayner (eds.), *Perspectives on sentence processing*, pp. 199–224. London: Lawrence Erlbaum.

Pickering, Martin and Traxler, Matthew 2001. 'Strategies for processing unbounded dependencies: Lexical information and verb-argument assignment', *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *27*, 1401–1410.

Pinheiro, Jose C. and Bates, Douglas M. 2000. *Mixed-effects models in S and S-PLUS*. New York: Springer.

Pritchett, Bradley 1991. 'Subjacency in a principle-based parser', in R. C. Berwick (ed.), *Principle-based Parsing: Computation and Psycholinguistics*, pp. 301-45. Dordrecht: Kluwer.

Rizzi, Luigi 1990. *Relativized minimality*. Cambridge: MIT Press.

Ross, John R. 1967. *Constraints on variables in syntax*. Unpublished doctoral dissertation, MIT, Cambridge, MA. (Published in 1986 as *Infinite syntax!* by Ablex, Norwood, N. J.)

Saah, Kofi and Goodluck, Helen 1995. 'Island effects in parsing and grammar: evidence from Akan', *Linguistic Review*, *12*, 381–409.

Sag, Ivan A., Hofmeister, Philip, and Snider, Neal 2007. 'Processing complexity in Subjacency violations: the complex noun phrase constraint', in M. Elliott, J. Kirby, O. Sawada, E. Staraki, and S. Yoon (eds.), *Proceedings of the 43 Annual Meeting of the Chicago Linguistic Society*, pp. 215-229. Chicago: University of Chicago.

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

Seidenberg, Mark S. and MacDonald, Maryellen C. 1999. 'A probabilistic constraints approach to language acquisition and processing', *Cognitive Science*, *23*, 569–588.

Snyder, William 2000. 'An experimental investigation of syntactic satiation effects', *Linguistic Inquiry*, *31*, 575–582.

Sorace, Antonella and Keller, Frank 2005. 'Gradience in linguistic data', *Lingua*, *115*(11), 1497–1524.

Sprouse, Jon 2007a. 'Continuous acceptability, categorical grammaticality, and experimental syntax', *Biolinguistics*, *1*, 118–129.

Sprouse, Jon 2007b. *A program for experimental syntax*. Unpublished doctoral dissertation, University of Maryland, College Park, College Park.

Sprouse, Jon 2008. 'The differential sensitivity of acceptability judgments to processing effects', *Linguistic Inquiry*, *39*, 686–694.

Sprouse, Jon 2009. 'Revisiting satiation: evidence for an equalization response strategy', *Linguistic Inquiry*, *40*, 329–341.

Sprouse, Jon, Wagers, Matt, and Phillips, Colin in press. 'A test of the relation between working memory and syntactic island effects', *Language*.

Staum Casasanto, Laura, Hofmeister, Philip, and Sag, Ivan A. 2010. 'Understanding acceptability judgments: Distinguishing the effects of grammar and processing on acceptability judgments', in S. Ohlsson and R. Catrambone (eds.), *Proceedings of the 32nd Annual Conference of the Cognitive Science Society*, pp. 224–229. Austin, TX: Cognitive Science Society.

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

Traxler, Matthew and Pickering, Martin 1996. 'Plausibility and the processing of unbounded dependencies: An eye-tracking study', *Journal of Memory and Language*, *35*, 454–475.

Wagers, Matt this volume. 'Memory mechanisms for wh-dependency formation and their implications for islandhoood'.

Wagers, Matt and Phillips, Colin 2009. 'Multiple dependencies and the role of the grammar in real-time comprehension', *Journal of Linguistics*, *45*, 395–433.

Wexler, Kenneth and Culicover, Peter 1980. *Formal principles of language acquisition*. Cambridge: MIT Press.