Cognitive Constraints and Island Effects

Philip Hofmeister
Center for Research on Language
University of California-San Diego
La Jolla, CA
Telephone: (650)-387-6641
E-mail: phofmeister@ucsd.edu

Ivan A. Sag
Department of Linguistics
Stanford University
Stanford, CA
Telephone: (650)-723-4284
E-mail: sag@stanford.edu

Abstract

Competence-based theories of island effects play a central role in generative grammar, yet the graded nature of many syntactic islands has never been properly accounted for. Categorical syntactic accounts of island effects have persisted in spite of a wealth of data suggesting that island effects are not categorical in nature and that non-structural manipulations that leave island structures intact can radically alter judgments of island violations. We argue here, building on work by Deane, Kluender, and others, that processing factors have the potential to account for this otherwise unexplained variation in acceptability judgments.

We report the results of self-paced reading experiments and controlled acceptability studies which explore the relationship between processing costs and judgments of acceptability. In each of the three self-paced reading studies, the data indicate that the processing cost of different types of island violations can be significantly reduced to a degree comparable to that of non-island filler-gap constructions by manipulating a single non-structural factor. Moreover, this reduction in processing cost is accompanied by significant improvements in acceptability. This evidence favors the hypothesis that island-violating constructions involve numerous processing pressures that aggregate to drive processing difficulty above a threshold so that a perception of unacceptability ensues. We examine the implications of these findings for the grammar of filler-gap dependencies.*

*Many of the ideas and results reported here grew out of collaborative research conducted on related issues with Florian Jaeger, Neal Snider, and Inbal Arnon, who have contributed to our thinking in many ways. In addition, we would like to acknowledge helpful discussions and/or critical interactions we have had with Bill Badecker, Tom Bever, Greg Carlson, Herb Clark, Daniel Casasanto, Ted Gibson, Robert Kluender, Colin Phillips, Perry Rosenstein, Laura Staum Casasanto, and Tom Wasow. This first author gratefully acknowledges research support from NIH Training Grant T32-DC000041 via the Center for Research in Language at UC-San Diego.

1 Introduction

It is generally agreed that grammar imposes no upper bound on the distance spanned by filler-gap dependencies (FGDs) in English *wh*-constructions like (1):[1]

(1) What did Sawyer say he thought everyone knew ... [$_S$ the president had eaten __]?

In addition, as more than a half-century of intense debate has made clear, *wh*-questions and FGDs like those in 2–4 sound unacceptable to most speakers of English:

(2) [Sawyer wondered who committed this atrocity.]
What did Sawyer wonder [$_S$ who committed __] ?

(3) [The journalist asked whether the general planned the assault.]
What did the journalist ask [$_S$ whether the general planned __]?

(4) [We met the mathematician who solved the puzzle.]
This was the puzzle that we met the mathematician who solved __ .

To explain the deviance of these examples and that of other filler-gap constructions, syntacticians have posited the existence of so-called 'island' constraints: grammatically-imposed restrictions preventing any constituents from escaping from certain configurationally-defined environments. Since Chomsky (1962), island constraints on FGDs have played a major role in the development of syntactic theory. Not only have the exegeses surrounding these data occupied a major portion of highly influential transformational studies (e.g. Chomsky 1973, 1977, 1981, 1986), but countless other syntactic analyses depend upon islands as a diagnostic for movement or variable binding.

In spite of their importance in current syntactic theory, island constraints have not enjoyed a stable analysis. In fact, the data used to motivate these constraints have been slippery at best. Syntactic theories of islandhood have had to recognize a heterogeneity in the universality and rigidity of different island constraints (see Szabolcsi 2006 for a useful review). Proposals have been made to treat counterexamples in various ways, e.g. by introducing ad hoc exception principles, by questioning the movement status of certain gap-binding dependencies, or by assigning 'marginal' status to the counterexemplifying data. Most importantly for the current discussion, the acceptability of sentences containing island violations appears to vary systematically with the manipulation of non-structural factors, i.e. with manipulations that leave the phrase structure configuration intact.

The present analysis pursues a competing line of thought – that the variation in acceptability judgments associated with these constructions, both language-internally and cross-linguistically, can be better explained by appealing to cognitive constraints on language processing (Deane 1991; Kluender & Kutas 1993b; Kluender 1991, 1992, 1998, 2005; Alexopoulou & Keller 2007). More generally, this view follows a tradition that analyzes acceptability as the end result of a series of cognitive processes. Accordingly, acceptability reflects the contribution of grammatical principles (competence-based factors), as well as resource limitations (performance-based factors).

Such an approach is articulated by Miller & Chomsky (1963), who argue that the unacceptability of complex center-embedding constructions results from cognitive limitations:

(5) The boy the girl the host knew brought left.

This perspective – that processing difficulty sometimes accounts for the perception of unacceptability – also underlies the standard treatment of 'garden path' sentences, e.g. *The horse raced past the barn fell* (Bever 1970; Gibson 1990; Osterhout & Swinney 1993). Acceptability contrasts for other sentence types have received a similar treatment: unacceptability is attributed to constraints on learnability or processing limitations (Bever 1975, 2009; Culicover 1984, inter alia). The unacceptability of sentences like 2-4, in contrast, has not been widely regarded as stemming from limitations on memory or other cognitive resources. The syntactic literature instead has maintained that examples like these reflect universal principles of grammar, e.g. the Complex NP Constraint, *Wh*-Island Constraint, or the Subjacency Condition.

In this paper, we argue that at least some island phenomena, in particular examples like 2–4, owe their character to the accumulation of performance-related difficulties that rises above a threshold to create the perception of unacceptability. This position is supported by a number of facts that are unexpected if island effects follow from constraints on hierarchical structure, but which follow naturally from an analysis based on processing difficulty. First, many island constructions have features that are known independently to produce processing difficulty. Variation in acceptability judgments for island-violating sentences appears to correlate with the presence/absence of those features that produce processing difficulty both within and across languages. Adding to the arguments offered by Kluender (1991, 1992, 1998, 2005), we present experimental evidence here from controlled acceptability studies and behavioral measures of processing which indicate that the factors that create processing difficulty also lower judgments of acceptability for island-violating sentences. As we will show, controlling these factors that make processing more difficult can also dramatically improve the acceptability of island-violating sentences.

It is also a striking fact that counterexamples have been discovered to nearly every structure-based island constraint proposed in the literature – that is, acceptable examples involving movement out of the putative island structure. This poses a further dilemma for explanations couched in terms of inviolable, structural constraints. By contrast, a theory based on processing has the capability to both explain and predict the existence of such examples, since manipulations that leave the island violation intact may evoke different processing challenges. Of course, the existence of counterexamples to proposed grammatical constraints does not in and of itself motivate a processing-based reinterpretation of the data. However, because the factors contributing to island effects are known to make processing more difficult in non-island environments, the processing-based account acquires considerable independent support.

On the basis of these arguments, we suggest that the grammar of FGDs is far simpler than generally countenanced in the syntactic literature – more 'minimal' even than analyses proffered within the 'Minimalist' Program (Chomsky 1995). In our view, competence grammars overgenerate significantly, leaving it to processing and other factors (including semantic and pragmatic effects) to explain why only a proper subset of the grammatically licensed sentences (including those like 2–4 above) are judged as fully acceptable. As noted above, the possible relationship between cognitive constraints and island effects has been raised before.

Hence, the primary goal of this paper is to identify specific linguistic manipulations that are independently tied to a reduction in processing difficulty and which also serve to increase the acceptability of island constructions. In addition, we argue that the observed variation in acceptability judgments is difficult to explain with standard grammatical machinery. Finally, we provide empirical results demonstrating that these non-structural manipulations do indeed affect both island processing and acceptability judgments.

The claim that competence grammar has no need for a Subjacency Condition or a *Wh*-Island Constraint should not be misconstrued as a claim that there are no constraints (or even universal island constraints) within competence grammar, e.g. there is sparingly little counterevidence to the Coordinate Structure Constraint, as well as behavioral evidence to support it (Wagers & Phillips 2009). However, an analysis of these island effects in cognitive terms opens the way to a more homogenous and transparent set of grammatical constraints on FGDs. That is, once we obtain a deeper understanding of the interacting performance factors, we may obtain a clearer understanding of the grammatical constraints on gap-binding dependencies.

To lay the groundwork for such an analysis, we begin the next section by briefly reviewing a number of influential theories of island phenomena, where the variance associated with these constructions is implicitly acknowledged. Not only are the judgments surrounding the key data often questioned by the authors themselves, but the resulting theories are often forced to reclassify linguistic dependencies in otherwise unmotivated ways in order to preserve the force of the basic generalizations. In the following section, many of the inherent processing difficulties imposed by island constructions are identified and discussed in the context of the discoveries made by psycholinguistic research. This discussion reveals that many island effects are accompanied by a variety of cognitive burdens that are typically not controlled for in syntactic research. While no single processing burden is likely to pose a serious problem for the human sentence processor, the combination of multiple, simultaneous demands and costs can become overwhelming.

After reviewing these factors, we present a series of experiments that evaluate the impact of one variable on the processing and acceptability of island-violating sentences: the syntactic and semantic complexity of the filler-phrase. Each of these experiments shows that increasing the complexity of the filler-phrase significantly facilitates the processing of FGDs, attenuating the island effect. In the remaining sections, we evaluate how both grammar-based theories and processing-based theories fare with respect to the data we discuss. We conclude with a general discussion of the implications of such an approach for grammatical theory.

2 Theories of islandhood

2.1 A brief history of island constraints

The earliest attempt to lay out a general principle to restrict the set of possible FGDs comes from Chomsky (1962) where the A-over-A Condition is introduced:

(6) **The A-over-A Condition (AOAC)**
An element of category A cannot be extracted out of a phrase of category A.

This principle rules out all sentences where an NP has been extracted from an NP, as in 7a-b:

(7) a. *What did he know [$_{NP}$ someone who has __]?
b. *What did you see the man read [$_{NP}$ the book that was on __]?

However, as Ross (1967) details, the AOAC predicts ungrammaticality for tokens considered acceptable [8–9] and grammaticality in cases judged unacceptable [10]:

(8) Who would you approve of [$_{NP}$ my seeing __]?

(9) Which astronaut did you read [$_{NP}$ a book about __]?

(10) *Which dignitaries do you think [[Sandy photographed the castle] and [Chris visited __]?

Counterexamples like these led Ross to introduce a number of distinct island constraints that are still part of the descriptive vocabulary of the modern syntactic literature, such as the Complex Noun Phrase Constraint:

(11) **The Complex NP Constraint (CNPC)**
No element contained in a sentence dominated by a noun phrase with a lexical head noun may be moved out of that noun phrase by a transformation.

These individual constraints cover the empirical ground more accurately than Chomsky's original proposal, allowing for examples like 8 and 9, but ruling out examples like 10 on the basis of a separate constraint (the Coordinate Structure Constraint).[2]

Like the AOAC, however, the individual constraints proposed by Ross were not without exceptions, many of which were observed and discussed by Ross himself. For instance, Ross observed that tense significantly alters the acceptability of some island-violating dependencies, e.g. CNPC violations, as does a shift to light verbs in periphrastic phrases like *make the claim that* or *have hopes that*. Thus, 12a sounds better than examples like 12b (we follow Kluender in writing $A \geq B$ to indicate that A is judged to be at least as acceptable as B):

(12) a. How much money are you making the claim that the company squandered? $\geq$
b. How much money are you stating the fact that the company squandered?

According to Ross, this perceived difference did not refute the validity of the proposed principle. Examples like 12a were instead removed from the domain of relevance because they were assigned a significantly different syntactic structure, one where the *that*-clause functions as a complement of the verb, not the noun.

Chomsky subsequently proposed a more general principle to account for a range of island phenomena that included CNPC violations in his theory of Subjacency (Chomsky 1973). In this theory, nestings of certain phrasal categories are blocked by the following condition on transformational movement:

(13) **The Subjacency Condition**
No rule may move a phrase from position Y to position X (or conversely) in:

$\ldots\ X\ \ldots\ [_{\alpha}\ \ldots\ [_{\beta}\ \ldots\ Y\ \ldots]\ \ldots\ ]\ \ldots\ X\ \ldots$

where α and β are cyclic nodes.

In English, IP and NP (or DP) are cyclic nodes, meaning that complex noun phrases like [$_{NP}$ *the rumor that* [$_{IP}$ *they started a new company*]] and subject noun phrases, as in [$_{IP}$[$_{NP}$ *the attempt to find the fountain of youth*] *ended in failure*]], constitute islands for movement. Hence, the grammar excludes examples like the following:

(14) a. It was a new company that Simon spread [$_{NP}$ the rumor that [$_{IP}$ they started __]].
b. What did [$_{IP}$[$_{NP}$ the attempt to find __] end in failure]?

Subjacency accordingly limits how many cyclic nodes a FGD can legally cross. Because of its formulation in terms of cyclic nodes, it also has the potential to explain constraints on superficially disparate structures, giving it a desirable generality.

However, as Deane (1991) shows, it is possible to construct acceptable examples where three or more cyclic nodes are crossed in English:

(15) a. Nixon was one president that [$_{IP}$ they had [$_{NP}$ no trouble finding [$_{NP}$ votes for [$_{NP}$ the impeachment of __]]]].
b. The chief purpose is [$_{IP}$ to avoid any sentences that [$_{IP}$ our informants report [$_{NP}$ significant variations in [$_{NP}$ their judgments about __]]]].

Moreover, Ross (1967) explicitly discusses similarly complex examples of 'deep extraction' like the following:

(16) Which reports does [$_{IP}$ the government prescribe [$_{NP}$ the height of [$_{NP}$ the lettering on __]]]?

Reasoning much as Ross did in his discussion of *make the claim* examples, Chomsky acknowledged that certain examples ruled out by the Subjacency Condition are fully grammatical (Chomsky 1973). For example, the choice of clause-initial *wh*-element seems to affect the grammaticality of the dependency:

(17) a. What crimes does the FBI know how to solve __ ?
b. *What crimes does the FBI know whether to solve __ ?

Admitting the graded acceptability of extractions out of embedded questions, Chomsky (1973, p. 245) concludes that 'the *know how to* examples such as [17a] are unique in permitting further *wh*-Movement [of the sort illustrated here]'. That is, examples of this type are excluded or regarded as 'unique' because the Subjacency theory fails to account for them, rather than because of any theory-independent considerations. We are not arguing against

the possibility of there being something grammatically special or extraordinary about certain phrases or collocations. Rather, we are simply pointing out a consistent pattern in the literature – the pattern of asserting without argument that counterexamples are 'special' or 'exceptional', thereby rendering them irrelevant for purely grammatical theories of syntactic islandhood.

In another case, where variation across speakers is evident, the reliability or standardness of some judgments is questioned:

> Some speakers seem to accept such forms as *What did he wonder whether John saw? What crimes did he wonder how they solved?* For me, these are unacceptable. It would be possible to add special rules to allow for these examples by a complication of the particular grammar, given the suggested interpretation of the conditions. [Chomsky 1973, p. 244]

Problematic examples and speakers are again treated as 'special'. The grammars of speakers who allow such tokens are nonoptimal, by implication, since a more complex grammar is necessary to explain their perceptions. Effectively, such speakers (and their associated grammars) do not have to be seriously considered or related to speakers with differing judgments.

There is also a very open acknowledgment of 'the gradation of acceptability' for extraction of picture NPs, where manipulating definiteness appears to affect acceptability:

> Some speakers (myself included) find a three-way gradation of acceptability, with (30b) [*Who did you see pictures of?*] better than *Who did you see the pictures of?*, which is in turn preferable to (31b) [*Who did you see John's picture of?*]. [Chomsky 1973, p. 239]

At the time, however, the Subjacency theory had no means of expressing this gradation, nor was it predicted by any aspect of the theory. But in a hint of things to come, Chomsky remarked that some structures might involve a 'double violation' and others only a 'single violation'.

The Barriers account (Chomsky 1986) realizes this notion of gradience in a more explicit fashion. According to this theory, certain XPs act as barriers to movement or extraction, specifically XPs that are not theta-governed (or L-marked, as in the definition below) by a lexical category, i.e. phrases not selected by a governing lexical head.

(18) γ is a BC [blocking category] for β iff γ is not L-marked and γ dominates β
γ is a barrier for β if and only if (a) or (b):

a. γ immediately dominates δ, δ a BC for β;
b. γ is a BC for β, $\gamma \neq$ IP.

Unlike the earlier Subjacency theory, the Barriers treatment explicitly acknowledges different degrees of acceptability, determined by the number of barriers crossed by movement:

(19) β is n-subjacent to α if and only if there are fewer than $n + 1$ barriers for β that exclude α.

Dependencies which cross zero barriers are thus 0-subjacent and should sound perfectly acceptable, '1-subjacency' should translate to marginal acceptability, but anything higher 'should yield a considerable decrement in acceptability' (*ibid.*, p. 30).[3]

As an illustration, examples like 20a-b involve movement across two barriers, according to Chomsky (1986):

(20) a. What did Simon spread [$_{NP}$ the rumor [$_{CP}$ that [$_{IP}$ they started __]]]?
b. What did Harold wonder [$_{CP}$ whether [$_{IP}$ they had ruined __]]?
c. What did the captain give [$_{NP}$ the command [$_{CP}$ [$_{IP}$ to start __]]]?
d. Who did Adele wonder [$_{CP}$ whether [$_{IP}$ to invite __]]?

One barrier is posed by the CP, which acquires this status since it immediately dominates a blocking category – the most embedded IP. The second barrier is actually the IP itself, despite the exception statement in 18 suggesting that IPs cannot be inherent barriers. This is because, in order to account for the apparent difference between extraction out of a tensed island [as in 20a and 20b], Chomsky assumed that the most deeply embedded tensed IP constitutes an inherent barrier to movement. 20c and 20d each present only a single barrier to movement: the CP that again earns its barrier status via inheritance from the IP. In these cases, the IP is only a blocking category and not a barrier, since the IP is not tensed. The latter two examples should thus sound better than the first two, if the predictions of the Barriers account are accurate. This theory accordingly recognizes that the island data are graded, but the gradience is limited to three distinct levels of grammaticality.

Unfortunately, it is not clear that there are only three levels of acceptability ('good', 'marginal', 'bad'), nor that three levels are adequate to explain perceived differences. Examples can be constructed which illustrate five or more levels of grading. Kluender (1992), for instance, offers the following acceptability ordering (from best to worst) of CNPC violations:

(21) a. This is the paper that we really need to find *someone* who understands. $\geq$
b. This is the paper that we really need to find *a linguist* who understands. $\geq$
c. This is the paper that we really need to find *the linguist* who understands. $\geq$
d. This is the paper that we really need to find *his advisor*, who understands. $\geq$
e. This is the paper that we really need to find *John,* who understands.

Further levels of grammaticality could be stipulated in the grammar, of course, but it is not clear that any finite number would actually be adequate. And, as Chomsky noted, if languages do not employ 'counters' for the purpose of tracking violations, it becomes difficult to justify the existence of some large number of discrete grammaticality levels.

Certain island constraints in the Minimalist Program emerge as a consequence of the Minimal Link Condition (Chomsky 1995, 2000). This principle constitutes the only locality constraint in the grammar, replacing both the *Wh*-Island Constraint and Chomsky's (1973) Superiority Condition, which stipulates that one *wh*-expression cannot be fronted over another structurally higher *wh*-expression:

(22) **The Minimal Link Condition** [Chomsky, 1995: 311]
K attracts α only if there is no β, β closer to K than α, such that K attracts β.

However, as Sabel (2002a) notes, the MLC fails to predict acceptability differences among *Wh*-island violations, e.g. object *wh*-movement out of a *wh*-island is better than adjunct movement or subject *wh*-movement (examples from Sabel 2002a):

(23) What do you wonder [how John could fix __]? ≥

(24) a. How do you wonder [what John could fix __]?
b. Who did do you wonder [how __ could fix the car]?

According to Sabel (2002a), these empirical oversights warrant an expansion of grammatical machinery, rather than a reevaluation of the generalizations about islands. The precise details are not critical for the present discussion; the relevant fact is simply that here again a theoretical choice has been made not to reassess the fundamental data that must be accounted for by grammar. Accordingly, even some of the most recent renditions of generative syntactic theory appeal to grammatical constraints to explain the bulk of island effects.

The island exceptions already noted in English are further accompanied by a panoply of counterexamples from other languages. Although the proposed notions of Subjacency, Barriers, and other island constraints purportedly constitute innate universal principles of language, many languages such as Swedish (Allwood 1976; Engdahl 1982; Andersson 1982), Danish (Erteschik-Shir 1973), Icelandic (Maling 1978), Norwegian (Taraldsen 1982), Italian (Rizzi 1982), French (Sportiche 1981; Hirschbühler & Valois 1992), Akan (Saah & Goodluck 1995), Palauan (Georgopoulos 1985, 1991), Malagasy (Sabel 2002b), Chamorro (Chung 1994), Bulgarian (Rudin 1988), Greek (Alexopoulou & Keller 2003), Yucatec Mayan (Elisabeth Norcliffe p.c.), and doubtless many others, exhibit counterexamples to this prediction:[4]

(25) a. Den där gämla skräphögen kanner ja killen som köpte?
That old piece junk know I the guy who bought.
'That old piece of junk, I know the guy who bought (it).'
b. Vilken bok kunde ingen minas vem som skrivit?
Which book could no one remember who that had written?
'Which book could no one remember who had written?'

[Swedish: Andersson 1982]

(26) Petta er lagið, sem enginn vissi hver samdi.
This is song-DEF that no one knew who wrote.
'This is the song that no one knew who wrote.'

[Icelandic: Maling 1978]

(27) Pion anarotihikes an tha apolisoune?
who-ACC wondering-2SG whether/if will fire-3PL
'Who did you wonder whether they will fire?'

[Greek: Alexopoulou & Keller 2003]

(28) Voilà la personne [que vous ne sauriez imaginer [avec quelle sauvagerie la police secrète a essayé de [faire parler]]].
'This is the person that you can't imagine with what brutality the secret police tried to get to talk.'

[French: Hirschbühler & Valois 1992]

(29) Hafa na kareta guäha mayulang ramienta in-isa pära
what? L car AGR.exist broken tools WH[OBJ].AGR-use FUT
infa'maolik
WH[OBJ].AGR-fix
'Which car were there some broken tools that you used in order to fix?'

[Chamorro: Chung 1994]

(30) máax t-u=tukl-ah Juan wáah t-u=ts'u'uts'-ah x-maria-o'
who PERF-A3=think=CMP Juan QPART PERF-A3-kissed-CMP FEM-Maria-D2
'Who did John wonder whether Maria kissed?'[5]

[Yucatec Mayan: Elisabeth Norcliffe p.c.]

But cross-linguistic counterevidence to a given island constraint has not generally been taken to be a refutation of the universality of that constraint as part of Universal Grammar. Perhaps the most common approach to such counterevidence is to assume that the apparent island violation really constitutes a dependency between a base-generated (i.e. non-moved) phrase and a null or overt resumptive (Georgopoulos 1985, 1991; Saah & Goodluck 1995). This strategy is essentially what is proposed by Rizzi (1990) and Cinque (1990), whose ideas we discuss in the next section: apparent violations of movement principles are categorized as cases of non-movement – that is, anaphor-antecedent dependencies. Unfortunately, the reasoning here quickly becomes circular: the apparent violations cannot involve movement, because this would contradict island constraints; and since they do not contradict island constraints, island constraints have not been counterexemplified. Effectively, as long as a case can be made for the existence of at least *some* resumptives in a language, there is a way of reanalyzing apparent island violations of this kind.

What stands out in the history of research on syntactic islands, in our view, is the entrenchment of the generalizations that were formed from early datasets. Despite extensive counterexamples or graded datasets, reassessments of the generalizations embodied in the grammatical constraints that were first proposed to deal with particular phenomena (e.g. the AOAC) have been rare. Instead, problematic examples have been labeled as 'special', 'peripheral' or 'unique', in the absence of independent motivation for these assessments. As we will show, there is a plausible alternative approach to CNPC and *Wh*-Island phenomena that makes sense of both the problematic counterexamples and the graded datasets.

2.2 Factors affecting the acceptability of islands

Beyond the graded nature of acceptability judgments observed in the primary literature on islands, a number of other sources have observed manipulations that affect the 'naturalness'

of island constructions. Notably, these manipulations do not change or eliminate the syntactic configurations that form the basis of the island constraints in question. As we shall see, however, these observations have not been interpreted as evidence against configurationally-defined syntactic islands; instead, they have spawned a series of proposals that proliferate mechanisms within the theory of phrasal movement. These complications of grammar effectively remove more and more data from the scope of island constraints, allowing the latter to remain part of the grammar. The effect is that the theory of islands applies to an ever diminishing subset of the observable data.

Perhaps the most widely discussed factor that affects the acceptability of *wh*-extractions is the nature of the *wh*-filler-phrase. Karttunen (1977) observes that while a bare *wh*-word fronted over a bare subject *wh*-word is judged as unacceptable (ostensibly due to the Superiority Condition), *which*-$\overline{\text{N}}$ phrases improve the acceptability:

(31) a. Which medication did which patient get? $\geq$
b. What did who get?

Maling & Zaenen (1982) notice, in addition, that *Wh*-island violations sound better with a *which*-$\overline{\text{N}}$ phrase compared to a bare *wh*-item:

(32) a. Which article don't you remember who wrote? $\geq$
b. What don't you know who wrote?

Similar sorts of observations regarding properties of the filler-phrase soon followed in Pesetsky (1987, 2000), Comorovski (1989), Rizzi (1990), and Cinque (1990). Pesetsky theorized that *wh*-phrases can be 'D(iscourse)-linked' and that this property voids the normal constraints on Superiority (or some comparable constraint that blocks 31b, e.g. the Minimal Link Condition). On a D-linking account, 31a is better than 31b because the *wh*-phrases are D-linked in the former, but not in the latter example. Unfortunately, a formal definition of D-linking is absent from the literature, making it virtually impossible to evaluate any theory based on this notion. The most common use of the term is intended to mean roughly that the set of possible answers is pre-established or otherwise salient. For numerous reasons, however, such a definition is ultimately inadequate (Hofmeister et al. 2007). First, regarding decontextualized linguistic examples like those in 31, the set of possible answers is no more pre-established in 31a than it is in 31b. Second, nothing stands in the way of associating a pre-established answer set with a bare *wh*-item, as Pesetsky himself acknowledges. Finally, there is no account of how or why general discourse properties like salience (sometimes) interact with the application of grammatical constraints, leaving a black box at the center of the D-linking analysis.

Rizzi (1990) documents similar effects for Italian *wh*-phrases, as in 33, where the complex interrogative phrase *a quale dei tuoi* leads to higher acceptability than a minimally different question with *a chi*:

(33) a. A chi non ti ricordi quanti soldi hai dato?
To whom don't you remember how much money you have give.
'To whom don't you remember how much money you gave?'

b. A quale dei tuoi figli non ti ricordi quanti soldi hai dato?
To which of your kids don't you remember how much money you have give
'To which one of your kids don't you remember how much money you gave?'

To account for such contrasts, Rizzi (1990) divides NPs (or DPs) that are 'intrinsically referential' from those that are inherently nonreferential. Referential NPs are defined as those arguments which 'refer to specific members of a set in the mind of the speaker or pre-established in discourse'. Cinque (1990) also adopts this distinction, in order to deal with extraction phenomena involving a range of quantifier phrases in Italian. Complex *wh*-phrases like those in 33b are thus categorized as referential, while *wh*-phrases like *a chi* are labeled nonreferential.

Dividing arguments into these two classes is explanatory only if there is a fundamental difference between the movement constraints that govern the two classes of expression. Rizzi and Cinque claimed precisely this, arguing that the movement traces of referential arguments do not need to be antecedent-governed. That is, structural considerations do not constrain the relationship between a referential filler-phrase and its trace. The traces of adjuncts and nonreferential arguments, however, must still be antecedent governed. Thus, movement of phrases of the latter type must proceed in a fundamentally different way, i.e. via successive-cyclic movement. Nonreferential phrases like *who* in English and *a chi* in Italian, therefore, are treated as being sensitive to structural constraints on movement, unlike referential arguments.

Rizzi and Cinque justified this division by appeal to the relationship between referentiality and the notion of 'bearing a referential index'. In particular, they assumed that referential arguments carry referential indices, enabling a binding relation to be made via coindexing (as in anaphoric binding). Since binding relations remain unconstrained by syntactic boundaries, 'long movement' (movement that crosses syntactic island boundaries) of referential arguments is licensed. On the further assumption that nonreferential phrases lack such indices, their relationship to their traces must be licensed by some distinct mechanism.

However, as Chung (1994) argues, this proposal is at odds with standard assumptions about bound variable anaphora. Argument phrases falling into the nonreferential class can clearly antecede a bound variable pronoun (e.g. *Who$_i$ loves his$_i$ mother?*). Note too that anaphor-antecedent dependencies may cross sentence boundaries, but 'referential' elements cannot be extracted out of a sentence. The above analysis would thus require a separate indexing system to relate 'referential' elements to their traces, essentially undermining the argument from binding:

> ... Rizzi's referential indices cannot be identified with the indexing mechanism that is a mainstay of current approaches to anaphora ... But if that is so, then we are left wondering whether the use of indices in this theory amounts to more than a diacritic to distinguish the DPs that allow long movement from those that do not. [Chung 1994, p. 33]

Even without these problems, the indexing proposal faces the question of why movement possibilities should vary with, or be determined by, specificity. As Chung (p. 39) succinctly asks:

> Why should long movement be legitimized in just those cases where the trace ranges over a sufficiently restricted set? To put the question differently, what is it about the ability to narrow down the domain of *wh*-quantification 'enough' that makes it possible for strict locality to be violated?

Neither Pesetsky's D-linking account nor the indexing analysis of Rizzi and Cinque clarifies or motivates the relationship between properties of the filler-phrase and the acceptability of FGDs. While they make the valuable contribution of identifying an acceptability effect based on the nature of the extracted element, they see the effect in terms of a sharp bifurcation of elusive grammatical mechanisms.

Consider one additional factor that affects the acceptability of island violations, but which has not been taken as counterevidence to island theory (cf. Kluender 1998). As mentioned in the previous section, islands demonstrate a sensitivity to the definiteness and/or specificity of intervening constituents. In his 1973 Subjacency theory, Chomsky highlighted this peculiarity in his discussion of picture NPs, but did not discuss why the definiteness of an intervening NP should interact with the application of movement transformations. Chomsky (1973) does, however, make the following statement, based on the observation that judgments surrounding picture NP extractions are not uniform:

> A refinement of the condition (26) [*the Specified Subject Condition*] incorporating the feature [definite] as well as the property of lexical specification might be proposed to accommodate these judgments. Specified subjects in NPs are [+ definite]. If (26) is revised to include [+ definite] as well as specified subjects, then (31b) [*Who did you see John's picture of*] will involve a double violation and *Who did you see the pictures of* only a single violation. This might account for the gradation of acceptability. [Chomsky 1973, p. 239]

As with the accounts of Pesetsky, Cinque, and Rizzi, such a proposal offers no way of understanding why the definiteness of intervening NPs should significantly alter the applicability of movement transformations.[6] Again, our concern here is not so much with the particulars or the adequacy of these early accounts, but with the sustained practice of dismissing problematic examples as either inconsequential or as uncorroborated evidence of some secondary grammatical mechanism.

It is interesting to note that the same sensitivity evidenced by English FGDs has also been documented in Swedish. While Swedish seems to demonstrate fewer island effects than English, CNPC violations are appreciably worse when an intervening NP is definite (exx. from Engdahl 1982):

(34) a. Johan känner jag ingen som tycker om.
Johan know I no one that likes
'Johan, I know no one that likes.' ≥

b. Johan känner jag en flicka som tycker om.
Johan know I a girl that likes
'Johan, I know a girl that likes.' ≥

c. Johan känner jag flickan som tycker om.
Johan know I DEF-girl that likes
'Johan, I know the girl that likes'

Relative clauses within nonspecific or generic NPs tend to be easier to extract out of than relatives that are part of definite NPs. However, as Andersson (1982) points out, this is merely a tendency – extractions out of relative clauses with definite heads are indeed possible.

This sensitivity to the referential properties of intervening constituents in both Swedish and English intriguingly matches the contrast observable in center-embeddings, where decreasing the specificity of embedded subjects improves the examples (although the construction remains difficult to process):

(35) a. The boy [someone [I knew __] brought __] left. ≥
b. The boy [the girl [the host knew __] brought] left.

Here, as in the picture NPs and the Swedish CNPC violations, an indefinite or other nonspecific intervener improves acceptability.

In fact, the sensitivity to the properties of intervening constituents does not appear to be limited to nominal arguments. The frequency, specificity or 'semantic richness', verb class, and argument selection properties of verbs on the extraction path also affect acceptability judgments. In Swedish, for instance, extraction out of complex NPs (putative CNPC violations) is preferable when the matrix verb is stative, as in 36a [exx. from Andersson 1982]:

(36) a. De blommorna känner jag en man som säljer.
Those flowers know I a man who sells
'Those flowers, I know a man who sells (them).' ≥
b. De blommorna talar jag med en man som säljer.
Those flowers talk I with a man who sells
'Those flowers, I talk with a man who sells (them)'

In English, as in Swedish, the acceptability of dependencies varies with properties of intervening verbs. Stative verbs, for example, are preferable to activity verbs, 37, and communication verbs are preferred when they do not specify the manner of speaking [38; Erteschik-Shir 1973] (for evidence that frequency and context have a significant role to play in determining the acceptability of examples like 38b, see Kothari 2008):

(37) a. Which book did you see pictures of? ≥
b. Which book did you destroy pictures of?

(38) a. Who did you say that John believes you saw? ≥
b. Who did you lisp that John believes you saw?

In fact, the acceptability of sentences containing FGDs looks to be generally caught up with features tied to intervening verbs – and not merely a sensitivity seen in island-violating constructions. Deane (1991) categorizes the first two examples below as acceptable, but the latter two are marked or worse in comparison (Deane's diacritics):

(39) a. Who did you obtain votes for the impeachment of?
b. Who did you find votes for the impeachment of?
c. *Who did you buy votes for the impeachment of?
d. **Who did you criticize votes for the impeachment of?

Deane notes that the interpretation of the latter two involves less plausible or likely scenarios (criticizing or buying votes) compared to the first two examples (obtaining or finding votes).

Cumulatively, the acceptability of FGD sentences varies in seemingly systematic ways with the nature of material that intervenes along the filler-gap path. This fact is irreconcilable with the syntactic proposals outlined in the previous section, since they have no means for evaluating lexical properties and adjusting the possibilities for movement accordingly. As illustrated by Chomsky's proposal, island constraints could be embellished with additional featural specifications, but these elaborations are fundamentally ad hoc – there is no accompanying theory of *why* these features should impact dependency formation.

The literature on islands has also exhibited a steadily increasing acknowledgment that not all island phenomena can be explained on structural grounds alone. Alongside the attempts to preserve the core syntactic generalizations, semantic and pragmatic explanations for Subjacency effects have been proposed as well, especially for the case of *weak islands* like tenseless *wh*-islands, factive islands, and negative islands (Erteschik-Shir & Lappin 1979; Kroch 1998; Comorovski 1989; Rizzi 2001; Szabolcsi & Zwarts 1993; Szabolcsi 2006; Oshima 2007; Truswell 2007). In general, these semantic and pragmatic accounts operate from the starting point that not all island effects are equally strong, and that syntactic constraints cannot or should not explain these differences. For instance, extraction out of a tenseless *wh*-clause is considered to be mildly degraded or marginally acceptable, compared to extraction out of a tensed *wh*-clause:

(40) This is a topic which John wondered whether to talk about. ≥
This is a topic which John wondered whether she talked about.

In fact, the defining property of so-called weak islands is that they block movement of some but not all phrases. For example, tenseless *wh*-islands have been claimed to allow argument extraction, but not adjunct extraction (Huang 1982). Consequently, they are contrasted with strong islands which purportedly block movement of all phrase types.[7]

Kroch (1998) also rejects a 'pure' syntactic explanation for island phenomena, arguing that some island effects, like the ban on moving adjunct phrases out of *wh*-islands and negative islands, owe their status to pragmatic considerations, rather than purely syntactic constraints. Others have come to a similar conclusion with respect to other island effects, arguing that putative island violations owe their status to pragmatically aberrant, or even paradoxical interpretations (Oshima 2007). That said, these semantic and pragmatic accounts do not dismiss the validity of syntactic island constraints in general. As a consequence, the standard syntactic island constraints are left in place. The semantic and pragmatic factors function so as to 'explain away' problems that can then be ignored by syntactic theories of islands.

In many ways, our approach here closely resembles the semantic and pragmatic accounts. We are reinterpreting some of the island data and suggesting that syntactic principles cannot

account for the observed variation in acceptability. In this sense, our analysis is not as radical as it first appears – it leaves intact the possibility that some island phenomena originate from syntactic constraints. The major difference between our analysis and the semantic-pragmatic theories mentioned above is that we believe that once the processing burdens are properly understood (and explained partly in terms of semantic and pragmatic factors), there remains little work to be done by purely syntactic island constraints.

In sum, the island data we have surveyed are characterized by graded and variable judgments, as well as exceptions within and across languages. Historically, the most influential proposals for analysis of such data have attempted to present a unified syntactic theory, but these have been repeatedly met with clear counterexamples and problematically graded acceptability data. In turn, this has led to numerous accompanying proposals aiming to preserve the core principles of the syntactic theory, while explaining away the 'peripheral' data that counterexemplify them. While there is consensus within the field that the full set of island data cannot be dealt with in purely syntactic terms, there remains a strong bias towards treating as much of the data as possible in purely syntactic terms.

3 Processing islands

Despite the early recognition of the role that processing pressures play in determinations of acceptability, the syntactic literature has by and large avoided invoking performance-related factors to account for either categorically different or gradient perceptions of island phenomena. This situation is especially puzzling, as island phenomena are characterized by so many features that independently contribute to processing difficulty. This section identifies some of the most significant factors that are likely to influence the processing of island constructions.

The processing-based analysis we advocate is consistent with a number of different models of sentence processing. The key property of such models is that they predict that performance on one cognitive operation can be disrupted by intervening tasks. To expend cognitive resources in one part of a sentence is assumed to reduce the availability of resources to process other parts of that sentence (Just & Carpenter 1992; Gibson 1990, 1998, 2000; Warren & Gibson 2002). For instance, Warren & Gibson (2002) propose that the difficulty of completing a linguistic dependency is contingent on processing demands imposed by material between the endpoints of a dependency. In particular, as more working memory resources are dedicated to referential processing inside a dependency, 'fewer working memory resources are available for maintaining syntactic representations and they decay' (Warren & Gibson 2005, p. 753).[8] That is, increased referential processing of dependency-internal constituents reduces the activation of other discourse representations, such as those of filler-phrases being stored in memory. Consequently, the cost of processing a discourse referent (including events) will have a predictable effect on the retrieval of other linguistic representations in memory. According to this theory, filler-gap processing will be less costly as processing demands along the filler-gap path decrease.

Following the ideas expressed by Gibson and colleagues, we assume that the memory representations associated with dislocated filler-phrases have an activation level that fluctuates throughout discourse. While repeated mentions and retrievals can increase the activation of

a mental representation (Vasishth & Lewis 2006), orthogonal processing tasks can dampen the activation of that representation. The factors that we review in the subsequent pages are identified as contributors to this decrease in activation.

3.1 Locality

The defining feature of all island violations is, of course, that they contain gap-binding dependencies. It has been clear for some time that a sentence with a FGD incurs a relatively high degree of processing difficulty, compared to a minimally different sentence without the dependency (Wanner & Maratsos 1978; King & Just 1991; Kluender & Kutas 1993a; Hawkins 1999):

> Filler-gap dependencies are difficult structures to process . . . Identifying the gap is not easy. It is an empty element with no surface manifestation and its presence must be inferred from its immediate environment. At the same time, the filler must be held in working memory, and all other material on the path from filler to gap must be processed simultaneously, and the gap must be correctly identified and filled. [Hawkins 1999, p. 246-7]

All sentences with FGDs are therefore relatively costly to process. But this additional processing cost is not without purpose: it pays for a noncanonical information structure or it otherwise achieves some discourse purpose that a minimally different construction lacking such an FGD would not be able to achieve. In other words, the special discourse functions that are associated with FGDs come with a processing cost.

Psycholinguistic research has confirmed not only that processing load increases inside a dependency, but also that processing difficulty generally increases as the dependency gets longer (Gibson 1998, 2000; Gibson & Pearlmutter 1998; Hawkins 1999; Fiebach *et al.* 2001; Grodner & Gibson 2005).[9] The standard explanation for this effect is that the activation of a mental representation decays over time, making it progressively harder to efficiently reaccess the memory item:

> . . .the greater the distance between an incoming word and the head or dependent to which it attaches, the greater the integration cost; and . . . the longer a predicted category must be kept in memory before being encountered, the greater is the cost for maintaining that prediction. [Gibson & Pearlmutter 1998, p. 265]

This leads to a preference for local attachment in ambiguous structures, and it also supports the claim that there is a preference associated with positing a gap as soon as possible (Stowe 1986; Frazier 1987; Clifton & Frazier 1989). But this memory-based constraint on dependencies also predicts processing differences for unambiguous dependencies where the length of the dependency varies.

Accordingly, prior research has established the existence of processing differences between nested and non-nested structures with meanings that are equivalent, modulo presupposition (Grodner & Gibson 2005):

(41) a. [The scientist collaborated with the professor [who had advised the student [who copied the article]]].

b. [The student [who the professor [who the scientist collaborated with] had advised] copied the article].

In the non-nested sentence 41a, the distance between dependents is minimal, making the sentence relatively easy to process. But the nesting or center-embedding shown in 41b increases the distance between the dependents, which raises the retrieval and integration costs significantly.

Many island-constructions, including CNPC and *Wh*-Island violations, are characterized by relatively long FGDs. This feature makes them particularly susceptible to the influence of other factors that compound comprehension difficulties. To be clear, dependency length alone and the associated processing difficulty cannot account for the unacceptability of islands. Sentences with comparably long dependencies are generally considered acceptable English sentences. The burden of maintaining a dependency over relatively long periods, however, may accompany other on-line tasks that significantly heighten overall processing difficulty.

It is important to understand, moreover, that increasing dependency distance in unquestionably grammatical contexts *does* reduce acceptability, in accord with the assumption that increased processing difficulty lowers acceptability. Arnon et al. (2005), for instance, verify that subject *wh*-interrogatives are judged more acceptable than object *wh*-interrogatives:

(42) a. Which man __ saw the girl in the bar on California Avenue? ≥

b. Which man did the girl in the bar on California Avenue see __ ?

In the subject *wh*-question, the *wh*-phrase can essentially be integrated and assigned a semantic role at the very next word (*saw*). By contrast, the *wh*-phrase in the object question must be stored and maintained in memory while three additional discourse referents are identified, categorized, and integrated. Hence, even in quite commonplace FGDs, the distance between the filler and gap can have a significant bearing on both measurable processing difficulty and judgments of acceptability, as obtained in controlled experimental circumstances.

3.2 Referential processing inside the dependency

Although Chomsky and others have noted that properties of intervening constituents, including specificity and referentiality, appear to interact with the possibilities for movement, there has been no principled account of why such interactions should occur. But from a psycholinguistic perspective, the impacts of specificity and/or referentiality form part of a more general pattern in dependency processing: intervening constituents whose processing consumes more resources reduce the resources available to link the filler with its gap. In other words, FGDs are processed more efficiently when the material between the filler and gap is itself easier to process.

Reasoning along these lines, Deane (1991) suggests that the ability to form a dependency is contingent upon the ability to attend to both the extracted element and the retrieval site simultaneously – in other words, the ability to retrieve an extracted element (or have

it in some sort of attentional focus, in his terms) and integrate it at the correct gap site. These attentional demands are impaired by distractions that occur along the filler-gap path (for related discussion of islands in terms of informational notions, see Goldberg 2006 and Erteschik-Shir 2007):

> Now, in a sentence involving extraction, we have hypothesized that the extraction site and the extracted NP must command attention. Of course, they would elicit attention most readily in the absence of competing elements (distractors). It would follow that the rest of the sentence should be relegated to the background, that is, presupposed or at least given ... Failures of extraction come when these conditions are not simultaneously satisfied. [Deane 1991, p. 36]

Along these lines, the effects tied to the referentiality and specificity of intervening NPs are also illuminated by processing studies. Warren & Gibson (2002) provide reading time evidence that definite NPs and proper names that intervene along a filler-gap path cause slower reading times at the retrieval site than do intervening personal pronouns:

(43) The consultant who (we/Donald Trump/the chairman/a chairman) called advised wealthy companies.

Warren & Gibson attribute this cost difference to the mental accessibility of the nominal referents, noting that first and second person pronouns are old referents in discourse:

> ... the effort spent attempting to access a referent before quitting and instantiating a new one may be less for NP types whose referents must be highly activated than for NP types whose referents are usually less activated. [Warren & Gibson 2005, p. 754]

Ariel (1990), Gundel *et al.* (1993), and others have developed scales of relative 'accessibility' with informationally-light elements like pronouns on the high end of the scale, and informationally-heavy expressions like definite descriptions much lower on the scale. The guiding intuition here is that the less salient (and therefore harder to retrieve from memory) a referent is, the more instructions a listener needs in order to identify the appropriate referent. Hence, when it is already clear what the intended referent of an anaphor is, short and nonspecific forms are licensed – few instructions have to be parsed and interpreted, which means less processing work. In contrast, when the referent is less salient, more instructions are needed to ensure successful comprehension, e.g. a definite description. And in order to parse and comprehend more instructions, more resources must be expended.

Now add to this picture the previously mentioned principle that there are limited cognitive resources available for linguistic processing, and expending those resources *in between* a filler-phrase and its gap increases the difficulty of processing the dependency (Warren & Gibson 2002, 2005). Certain types of NPs engender more processing difficulty than others because they contain more information or instructions for identifying and/or situating referents, e.g. *the girl* vs. *someone*. The process of resolving these references diminishes the

activation of other representations held in working memory, thus making retrieval more difficult. Hence, discourse references inside a filler-gap dependency influence processing effort at two different sites (at least): at the discourse reference itself, but also at the conclusion of the filler-gap dependency.

Accordingly, the acceptability variation in extraction out of picture NPs, complex NPs, and other islands is predictable on the assumption that increased processing effort generally leads to lower acceptability judgments. Definiteness effects can thus be reasonably understood as the byproduct of processing difficulty. Since definites overwhelmingly refer to old or established referents in the discourse, such NPs may automatically trigger a search for a referent. Where there is no contextually established antecedent, such searches will ultimately fail and a new mental representation will be created to accommodate the presuppositional demands of the sentence. Where context does provide an antecedent, the mental search will still have to be conducted. An indefinite, however, would not trigger a search in either situation, and a new mental representation would immediately be created, thus bypassing a potentially costly processing task.[10]

In sum, processing a filler-gap dependency and retrieving the filler at the appropriate site is made easier when there are fewer cognitive demands associated with processing intervening references. It is therefore notable that island violations are demonstrably worse when the FGD crosses references to less accessible entities, as compared to references that are easier to situate. Based on the evidence from sentence processing, it seems reasonable to hypothesize that processing differences contribute to these contrasts in acceptability.

3.3 Clause boundaries

One of the fundamental observations behind syntactic accounts of islands is that certain syntactic configurations impose barriers or boundaries to movement. For instance, *wh*-clauses limit extraction possibilities, purportedly because the grammar specifies restrictions that guarantee this. Psycholinguistic research shows, however, that processing clause boundaries generally lowers acceptability ratings and causes an increase in processing time (Frazier & Clifton 1989; Kluender & Kutas 1993b; Kluender 1998).

Experimental findings from Kluender & Kutas (1993b), for instance, raise the possibility that boundary-based effects ultimately stem from processing considerations and not grammatical constraints. They report that, even in yes-no questions where no filler is carried across the clause boundary, different complementizers elicit significantly different neurophysiological responses, as well as acceptability judgments. The complementizer *that* produces the highest rating of acceptability, while the bare *wh*-word *who* garners the lowest rating, producing contrasts like those shown in 44:

(44) a. Has she forgotten [*that* he dragged her to a movie on Christmas Eve]? ≥
b. Has she forgotten [*if* he dragged her to a movie on Christmas Eve]? ≥
c. Has she forgotten [*who* he dragged __ to a movie on Christmas Eve]?

These findings suggest that there is indeed something special about embedded interrogatives, as in 44b and 44c. Compared to 44a, they elicit neurological responses compatible

with an interpretation of greater processing difficulty, and judgments that reflect degraded acceptability.

This evidence suggests that different types of syntactic boundaries appear to have different cognitive costs, independent of whether or not a FGD crosses that boundary, even in uncontroversially grammatical sentences. Kluender (1998) elaborates on this theme and argues that some of these differences may stem from contrasts in referential processing at the clause boundary. The example in 44c, for instance, requires setting up a new discourse referent at the clause boundary, which does not occur in the minimally different variants. As discussed in the previous section, this additional referential processing may reduce the activation of other discourse representations being held in memory.

In addition, the clause types themselves seem to initiate unique sets of processes that differentiate the clauses from one another in terms of processing load. An interrogative clause, for example, is standardly regarded as being semantically more complex than a declarative clause; a question is frequently analyzed as a set of propositions (Hamblin 1973; Karttunen 1977; Groenendijk & Stokhof 1997) or else as a propositional function in which a set of parameters is abstracted (Ginzburg & Sag 2000). On the computationally lightest end of interrogatives, a *whether*-clause requires consideration of the positive and negative alternatives. In cases like 44c, the *wh*-word also launches a long-distance dependency and the search for a suitable integration site, thereby taxing the processing system even more. Differences among types of embedded clauses may thus be framed in terms of the kinds and number of cognitive processes they initiate. Essentially every semantic theory of interrogatives posits a complexity or abstractness for questions that is absent from propositions. These complications pose likely hindrances for on-line sentence processing, and thus offer an explanation for why carrying a dependency into certain clause types may lead to significantly reduced judgments of acceptability.

3.4 Complexity of the filler-phrase

The difficulty of processing FGDs is contingent not only on the length of the dependency and the number and kind of intervening constituents, but also properties of the filler itself. Hofmeister (2007) provides evidence that the quantity of syntactic and semantic features encoded in a filler affects filler-gap processing in a wide variety of sentential contexts. In particular, syntactically and semantically more complex filler-phrases are shown (via reading time studies) to facilitate processing at downstream retrieval points. For example, definite NPs like the one in 45b, which is syntactically and semantically more complex (by a simple subset-superset relationship) than the definite in 45a, were found to produce faster reading times beginning at the subcategorizing verb (*encouraged*):

(45) a. The diplomat contacted *the dictator* who the activist looking for more contributions *encouraged* to preserve natural habitats and resources.

b. The diplomat contacted *the ruthless military dictator* who the activist looking for more contributions *encouraged* to preserve natural habitats and resources.

Similar complexity effects appear in clefts and relativizations, as well as in non-island *wh*-dependencies. Moreover, these complexity-based effects occur with *wh*-phrases and definites,

as well as indefinite NPs.

To account for this pattern, Hofmeister (2007) argues that linguistic expressions that contain more syntactic and semantic features facilitate downstream retrieval of the corresponding linguistic representations from memory, due to increased activation and resistance to interference. In essence, processing a phrase or word that syntactically depends on or modifies some other representation in memory re-activates (or pre-activates) that representation, as claimed by Vasishth & Lewis (2006). This re-activation process provides a boost to the activation level, making the representation easier to access subsequently. Higher activation further means that activation will stay relatively high, even though dependency-internal processing may reduce activation. And by providing representations with distinguishing semantic features, these representations are less likely to be confused with other discourse representations in memory at the time of retrieval. Such a theory of the complexity effects in sentence processing also accords with related experimental findings. A wealth of memory research shows that elaboration or increased semantic processing during the study phase of recall tasks benefits subsequent retrieval of isolated words and propositions (Reder 1980; Anderson & Reder 1979; Bradshaw & Anderson 1982; Anderson 1983; Wiseman *et al.* 1985; Reder *et al.* 1986; McDaniel *et al.* 1989, inter alia). This memory-based account thus offers one means for explaining why semantically richer, more complex filler phrases promote increased acceptability.

But semantically richer filler-phrases offer other potential advantages in sentence processing. A problem with many of the island violations cited in the literature is that they lead to an initial misparsing of the sentence. That is, given a preference for associating a filler with a head as early as possible, i.e. an active-filler strategy (Clifton & Frazier 1989), the parser will try to integrate a filler-phrase stored in memory at the first possible opportunity. Thus, in the partial string *What did you assert* ..., the *wh*-item *what* could reasonably act as the direct object of the verb *assert*, so the parser is likely to attempt integration at this site. If more material follows, disconfirming this initial analysis, then the original parse has to be undone, creating processing difficulty. In contrast, for a partial string such as *Which article did you assert* ..., the complex *wh*-phrase cannot plausibly be interpreted as the direct object of the verb. Accordingly, any analogous reading is quickly rejected on the grounds of semantic incongruity, leaving no prior analysis to be undone when the actual gap is encountered (Pickering & Traxler 2001). Semantically richer filler-phrases, therefore, reduce the chances of garden-pathing.

Lastly, in the context of direct interrogatives, semantically richer *wh*-phrases potentially narrow down the list of candidate answers (or focus alternatives). To illustrate, consider a context in which there are numerous animate entities, among which are two students but multiple non-students. A question beginning with *Which student* ..., as opposed to *Who* ..., narrows down the range of alternatives that have to be considered, thereby reducing computational effort. This aspect of filler-complexity perhaps bears a resemblance to the notion of D-linking discussed above, insofar as contextual restrictions can improve certain *wh*-dependencies. But whereas D-linking voids the application of a grammatical constraint on the basis of contextualization, the relevant principle here is the simple fact that computing an answer to a question is easier when the scope of the question is narrowed by the discourse context.

3.5 Summary

This list of sentence processing factors is hardly exhaustive. Sentence processing research has also identified critical roles for frequency or predictability in determinations of processing difficulty – less frequent or unpredictable constituents are more difficult to process (Hale 2001; Jurafsky 2003; Levy 2005). Other sentence processing factors that are potentially relevant for the processing of islands include similarity-based interference (Lewis 1996; Gordon *et al.* 2002; Van Dyke 2007, inter alia), collocation frequency, and degree of contextualization difficulty.

An analysis incorporating these processing facts can side-step the need to posit 'special' rules or a difference between normal and 'special' speakers. Our main point with respect to processing costs in islands is that *multiple* processing burdens can interact to create unacceptability as the byproduct of multitask interference. Dual-task and multitask interference is a well-known and well-investigated phenomenon in psychology quite generally, e.g. Pashler (1994) among many others. Moreover, this interaction is potentially nonlinear in nature. Factors that consume the same or conflicting cognitive resources can interact in a highly nonlinear (or 'superadditive') fashion (for one example of such effects in sentence processing, see Fedorenko *et al.* (2007)). Definiteness by itself, therefore, may contribute to an overall perception of reduced acceptability. But definiteness along with additional resource-consumptive factors may pose a serious obstacle to sentence processing, and may hence dramatically reduce acceptability.[11]

Some of these features tied to processing difficulty can be altered or eliminated in island-violating sentences. For instance, dependencies can be made shorter, references to less salient entities can be replaced with references to more salient entities, filler-phrases of greater complexity can be used, etc. However, other features appear to be inherent properties of the islands themselves. A dependency into a *wh*-island, for instance, requires processing an embedded interrogative (as opposed to an embedded proposition), while holding a filler-phrase in working memory and actively searching for a suitable subcategorizer. Following Kluender (1998), this combination of processing tasks is what we claim to be responsible for the 'island' effect.

CNPC violations, too, have a number of inherent features that make them difficult to process. To begin with, a dependency into a complex noun phrase in English requires processing at least three nominal references inside the dependency and crossing a clause boundary:

(46) Which politician did you$_1$ read reports$_2$ that we$_3$ had impeached?

Added to this, a syntactic ambiguity arises after processing *reports that*. At this point, both a sentential complement and a relative clause parse (as in *Which politician did you read reports* $[_{RC}$ *that we had written] in front of?*) are theoretically possible. Even if one parse is more likely from a top-down perspective, these parses may nevertheless compete with one another (Gibson 2006). Not until the embedded verb is there strong evidence for which parse is correct (even here, the strength of this evidence depends on the compatibility of the verb and island-forming NP). But, as we believe the above example demonstrates, CNPC violations can be made quite acceptable by using semantically rich fillers, choosing intermediate verbs

that cannot plausibly take the filler-phrase as a direct argument, and intervening NPs of 'high accessibility.' By mitigating the chances of misparsing and processing overload along the filler-gap path, these CNPC violations approach the acceptability of sentences whose dependency length and number of intervening discourse referents is comparable.

Given these strong indications that the variation in islands may reflect processing preferences, we now turn our attention to a series of experimental investigations that explore the role of processing in island effects. These studies are intended to evaluate the extent to which acceptability differences are tied to processing differences. A theory that explains island effects in terms of processing differences must verify that the processing differences actually exist. If controlled experiments find that acceptability differences are not accompanied by processing differences in islands, then there would be reason to look for an explanation elsewhere. Similarly, if the results turn out to be the opposite of what is predicted by our knowledge of cognitive constraints, then the processing hypothesis would be similarly challenged. On the other hand, if processing differences within islands are systematically matched by contrasts in acceptability judgments, then this would be compatible with the view that limitations on cognitive resources are responsible for island effects. But there are other possible interpretations of such findings, e.g. that grammatical differences underlie the processing differences. In section 8 we present a number of arguments supporting the conclusion that cognitive constraints are playing a causal role in the island phenomena we investigate.

4 Experimental investigations

In the following sections, we present the results of three experiments that utilize the self-paced reading methodology. In these comprehension experiments, subjects read sentences at their own pace on a computer screen (Just *et al.* 1982). Participants are presented with a screen of dashes separated by spaces, representing the words for that experimental item. With each press of a predefined key, a new word appears on the screen and the previous word disappears. An example of a sequence of such states is given below:

(47) —— call — ——— ———- —- ——— – — ———
—— —- to ——- ———- —- ——— – — ———
—— —- — confirm ——- —- ——— – — ———

The amount of time between each key-press is automatically recorded and tagged with relevant information. Longer reading times at a particular word or region are interpreted as an indication of processing difficulty.

Between six and eight practice examples preceded the real experimental items in order to acquaint participants with this manner of reading and to reduce the magnitude of order effects on items presented early in the experiment. During the experiments proper, each participant saw each item in exactly one condition (Latin-square design). Blocking of items into lists and randomization within lists was automatically managed by the reading-time software, LINGER v. 2.94, developed by Doug Rohde (available at http://tedlab.mit.edu/ dr/Linger/).

Each experimental item was followed by a comprehension question. After reading the question, subjects responded either by selecting the correct answer from a set of possible

choices or by responding 'yes' or 'no' in the case of polar interrogatives. Response times for answering the question, as well as question-answer accuracy, were also studied. Response times that were 2.5 standard deviations from each subject's mean were removed prior to analysis. Only reading times from correctly answered stimuli are considered in the statistical analyses presented here.

If the mean question-answer accuracy for a subject was below 67%, then the entire data set for that individual was dropped from the analysis. If a subject's global reading time average differed from the sample's global average by more than 2.5 standard deviations, that subject's results were also excluded.

Residual reading times were derived for each word on the basis of a linear regression equation that computed reading time for each individual as a function of word length. A subject's residual reading time for a given word of length n expresses how fast that word was read compared to the subject's predicted reading time for words of length n. This practice of using residual reading times reduces variability due to word-length differences and participants' reading rates (Ferreira & Clifton 1986). For each experiment, however, the data patterns evident in the residual reading times appear in the raw reading times as well. Statistical outliers were removed by deleting residual reading times that were more than 2.5 standard deviations from the mean at each word region. This process affected less than 3% of the total data in Experiment I, less than 2% in Experiment II, and less than 4% in Experiment III.

All experimental data were analyzed with repeated measures ANOVAs or t-tests. In experiments where more than two treatment or condition levels are compared with one another, the results of planned t-tests are also reported. As repeated hypothesis testing with the same sample increases the likelihood of false positives (Type I error), the family-wise error rate must be corrected in such situations. Within the text itself, we report only the uncorrected p-values, but since it is never the case that more than three treatment levels are compared with one another, the Bonferroni corrected level of significance can be safely estimated at .0167 for such cases.

5 Experiment I: complex noun phrase constructions

The first experiment looks at the role of two independent processing factors in the processing of CNPC violations: the complexity of the displaced *wh*-phrase and the type of island-forming NP. Both factors have been identified as determinants in the overall acceptability of CNPC violations, as outlined above. In particular, increasing the complexity of the filler-phrase appears to raise acceptability judgments, while extraction out of a complex definite NP intuitively seems harder than extraction out of an indefinite.

If these acceptability effects are connected to processing effort, then manipulating these factors should produce observable processing differences. More specifically, relatively complex fillers like *which*-$\overline{\text{N}}$ phrases should lead to facilitated processing, as compared to simple fillers like bare *wh*-words. Similarly, if the observed effects of definiteness are tied to processing, then definite islands should be accompanied by greater processing difficulty than indefinite ones.

5.1 Materials & participants

The stimuli in this experiment consisted of 36 embedded CNPC violations, which varied with respect to the content of the filler-phrase and the nature of the island-forming NP. For each item, each participant saw one of seven experimental conditions ((2 x 3) + 1). Across conditions, the fronted *wh*-phrase was either a bare *wh*-item (= BARE condition), such as *who* or *what*, or a comparatively more complex *which*-$\overline{\text{N}}$ phrase (= WHICH condition). The other factor considered in this experiment was the effect of NP type on subsequent sentence processing. Subjects read one of three kinds of island-forming NPs: a definite NP (DEF), an indefinite plural (PL), or an indefinite singular (INDEF). Additionally, a baseline for each item was included that lacked the island-forming NP. A sample experimental item with all seven conditions is shown in 48:

(48) BARE-DEF: I saw **who** Emma doubted the report that we had captured in the nationwide FBI manhunt.
BARE-PL: I saw **who** Emma doubted reports that we had captured in the nationwide FBI manhunt.
BARE-INDEF: I saw **who** Emma doubted a report that we had captured in the nationwide FBI manhunt.
WHICH-DEF: I saw **which convict** Emma doubted the report that we had captured in the nationwide FBI manhunt.
WHICH-PL: I saw **which convict** Emma doubted reports that we had captured in the nationwide FBI manhunt.
WHICH-INDEF: I saw **which convict** Emma doubted a report that we had captured in the nationwide FBI manhunt.
BASELINE: I saw **which convict** Emma doubted that we had captured in the nationwide FBI manhunt.

Eighty filler items of comparable sentence length accompanied these experimental items. Comprehension questions followed every experimental stimulus. These questions were polar interrogatives that probed subjects' understanding of the sentence, e.g. *Was Emma skeptical that we had captured someone?* Participants were provided negative feedback if they answered a question incorrectly. A full list of materials is provided in Appendix A.

Twenty-nine Stanford University undergraduates participated in this study. All participants were native speakers of English, naive to the purpose of the study. The results of four subjects were dropped due to question-answer accuracies below 67%. Participants were paid $15 to complete this experiment and an unrelated off-line survey. Due to poor question-answer accuracy for some items in certain conditions (2.77% of the item analysis data set), the missing data for those cells were imputed via the linear trend for that point using the *Replace Missing Values* command in SPSS 15.0. This method of data imputation inserts predicted values of empty data cells by regressing against the existing data points (see also note 12 for further statistical analysis without data imputation).

5.2 Results

Reading times

We evaluate here the reading times for the word regions following the *wh*-phrases to word regions inside the embedded complement clause. Of particular interest are the reading times inside the complement clause, since this region constitutes a proposed island for extraction in the island-violating conditions. Table 1 lists the mean residual and raw reading times from the complementizer to two words after the embedded verb.

[TABLE 1 ABOUT HERE]

Immediately after the *wh*-phrases (e.g. the word *Emma* in 48), there is a main effect for *wh*-phrase complexity, such that the *less complex* BARE conditions lead to significantly faster reading times (BARE = −42.65 ms, $SE = 6.86$, BASELINE = 6.67 ms, $SE = 20.97$, WHICH = −18.24 ms, $SE = 7.77$) (F1(1,24) = 4.725 $p < .05$, F2(1,35) = 2.901, $p < .1$). Similar trends show up at the following word, but they do not reach significance by subjects or items (BARE = −38.40 ms, $SE = 9.30$; BASELINE = −36.32 ms, $SE = 17.79$; WHICH = −14.44 ms, $SE = 8.52$) (F1(1,24) = 1.597, $p = .219$, F2(1,35) = 3.685, $p = .063$). NP type is, of course, irrelevant at both these sites as the NPs have not been reached yet.

Starting at the complementizer *that*, the WHICH conditions generate significantly faster reading times than the BARE conditions, as detailed in Table 1. In all three NP type conditions, as shown in Figure 1, the WHICH version is processed faster than the corresponding BARE version. This processing advantage for the WHICH conditions extends beyond the complementizer to include the subsequent pronominal subject, the embedded auxiliary and verb (see Figure 2), as well as the regions after the verb where the presence of the gap is confirmed. In other words, from the complementizer until several words after the subcategorizer, the WHICH conditions are read faster than the corresponding BARE conditions. Averaging over all the word regions from the complementizer to the spillover regions for the verb, there is a highly significant main effect of *wh*-phrase complexity (F1(1,24) = 18.365, $p < .001$; F2(1,35) = 22.723, $p < .001$).[12]

[FIGURE 1 ABOUT HERE]

[FIGURE 2 ABOUT HERE]

In addition, throughout the embedded clause, the region-by-region reading times for the WHICH conditions are not significantly different from those of the non-island baseline, as Table 1 evidences. This contrasts with the BARE condition, which remains slower than the baseline throughout the embedded clause.

At only two word regions are there significant effects of NP type: at the complementizer and at the embedded verb (see Table 1 for ANOVA results). At the complementizer, the singular indefinite conditions are read faster than both the definite and plural indefinite conditions, which creates a main effect of NP type. Since the difference between the indefinites and the other two NP types is far greater in the BARE conditions (BARE-DEF: 150.43 ms, $SE = 48.29$; BARE-INDEF: −12.47 ms, $SE = 15.63$; BARE-PL: 127.18 ms, $SE = 33.16$; WHICH-DEF: 21.28 ms, $SE = 22.84$; WHICH-INDEF:−18.94 ms, $SE = 13.00$; WHICH-PL: 28.44 ms, $SE = 16.77$), there is also a marginal interaction of *wh*-phrase type and island NP type. At the embedded verb, we also observe an effect of NP type in the by-items analysis, but

not in the by-subjects analysis.[13] In this case, the effect appears to be driven by the fact that the DEF conditions are slower than the INDEF conditions (BARE-DEF: −11.79 ms, *SE* = 17.93; BARE-INDEF: −62.25 ms, *SE* = 10.79; BARE-PL: −51.40 ms, *SE* = 13.57; WHICH-DEF: −71.68 ms, *SE* = 11.42; WHICH-INDEF:−81.77 ms, *SE* = 10.26; WHICH-PL: −62.10 ms, *SE* = 13.84). This contrast is much more pronounced in the context of a bare *wh*-word, which yields the marginal interaction shown in the ANOVA results in Table 1. In other words, the combination of the bare *wh*-word and the definite island-forming NP resembles a super-additive effect, although these effects are marginal in significance. No other word regions evidence a significant interaction.

In sum, the reading results identify a brief slowdown after processing a complex *wh*-phrase, but beginning with the complementizer, this disadvantage reverses and becomes a highly significant processing advantage. This advantage persists until several words after the subcategorizing verb. Effects due to the type of the island-forming NP are much more localized (and relatively weaker) and appear only at the complementizer and verb. Nevertheless, these effects do suggest that definite NPs lead to slower reading than indefinite NPs.

Comprehension questions
Reaction times to the comprehension questions in the island-violating conditions show no main effects or interactions of complexity and NP type (BARE-DEF: 2502.77 ms, *SE* = 105.11; BARE-INDEF: 2703.46 ms, *SE* = 122.88; BARE-PL: 2454.99 ms, *SE* = 113.17; WHICH-INDEF: 2570.18 ms, *SE* = 120.10; WHICH-DEF: 2549.41 ms, *SE* = 103.92; WHICH-PL: 2515.45 ms, *SE* = 97.07). The BASELINE, however, produces faster reaction times than the island-violating conditions (BASELINE: 2280.27 ms, *SE* = 95.42). This effect is significant by subjects, but only marginal by items after corrections for multiple comparisons (BASELINE-BARE: $t_1(24) = -3.713$, $p = .001$; $t_2(35) = 1.609$, $p = .117$; BASELINE-WHICH: $t_1(24) = -3.179$, $p < .01$; $t_2(35) = 2.425$, $p = .021$). Question-answer accuracies do not evidence main effects of either NP type or *wh*-phrase complexity (BARE-DEF: 81%, *SE* = 3.5; BARE-INDEF: 80%, *SE* = 3.5; BARE-PL: 75%, *SE* = 3.8, BASELINE: 77%, *SE* = 3.7; WHICH-INDEF: 75%, *SE* = 3.8; WHICH-DEF: 81%, *SE* = 3.5; WHICH-PL: 84%, *SE* = 3.3).

5.3 Discussion

Reading times in this study show a strong influence of the complexity of the filler-phrase. While reading times immediately after the dislocated *wh*-phrase reflect slowed reading after a complex *which*-$\overline{\text{N}}$ phrase, reading times throughout the most embedded clause are significantly faster in the WHICH condition. In fact, reading times within the most deeply embedded clause in the WHICH condition are not significantly different from reading times in the non-island baseline condition.

An early effect of *wh*-complexity is observed after the head noun of the sentential complement: bare *wh*-items lead to elevated reading times, compared to complex *wh*-phrases. This is consistent with the idea that, in the conditions with bare *wh*-items, reanalysis occurs after the presence of the noun forces the parser to revisit any attempts to associate the filler with that syntactic position, i.e. a filled-gap effect (Stowe 1986). But the data also point to continued reading time differences at the embedded auxiliary, verb, and spillover regions.

These differences may well reflect on-going reanalysis effects, but they are also compatible with retrieval-based differences, such that the memory representation associated with the more complex *wh*-phrase is easier to retrieve (Hofmeister 2007). The results are thus explicable on the analysis that semantically richer *wh*-phrases reduce the likelihood of syntactic misanalysis and/or that these richer *wh*-phrases facilitate retrieval operations. On either analysis, though, the faster downstream reading times for complex *wh*-phrases are predicted by independent principles of language processing.

Although the effect of filler complexity is powerful, the definiteness of the island-forming NP appears to have a somewhat weaker effect. Nevertheless, at two word regions, there is an observable effect of NP type. The reading time differences at these sites line up with the prediction that singular indefinites lead to faster processing than definites. Moreover, the effect of NP type at the embedded verb parallels the findings of Warren & Gibson (2002, 2005), which establish that dependency-internal referential processing impacts reading times at retrieval sites. The interaction at this site between NP type and *wh*-phrase complexity also suggests a type of super-additive effect: the effect of definiteness is more pronounced in contexts where processing is already difficult. The results for the plural condition, however, are harder to interpret, because this condition sometimes patterns with the definite NPs, and sometimes with the singular indefinites.

To complement the self-paced reading experiment and to understand the relationship between these comprehension results and judgments of acceptability, we also conducted a controlled acceptability study using identical stimuli, including the same fillers. As with the self-paced reading task, items were randomized and distributed across lists such that each subject saw only one condition per item. Twenty-one subjects (none of whom had participated in the previous reading time study) were asked to rate the sentences for naturalness on a scale of 1 to 8 and were specifically instructed not to rate the sentences according to prescriptive grammar rules. The subjects in this study were given course credit for their participation. A mean acceptability rating was derived for each subject, based on all items and all fillers. The scores for individual tokens were divided by this overall mean to produce a judgment ratio that takes into account individual variation in usage of the scale (using standard z-scores or the raw scores produces similar results):

[FIGURE 3 ABOUT HERE]

As shown in Figure 3, the acceptability judgments yield the same main effect of *wh*-phrase complexity as in the reading time study ($F1(1,20) = 48.741$, $p < .0001$; $F2(1,35) = 39.494$, $p < .0001$), due to the higher ratings for sentences with complex *wh*-phrases. NP type and the interaction of *wh*-phrase complexity and NP type, in contrast, are not significant, despite a non-significant trend for higher judgments for plural indefinites in the WHICH conditions, compared to singular indefinites and definite NPs ($F1(2,40) = 2.146$, $p = .130$, $F2(2,70) = 2.947$, $p = .059$). Since this trend does not reach significance, we do not discuss it further, except to note that these findings are not incompatible with the processing results, as NP type had relatively weak and transient effects on processing.

As is evident from Figure 3, the baseline in the acceptability study receives significantly higher ratings than the WHICH conditions (ps $< .001$). This contrasts with the reading time

study, where the baseline did not consistently produce significantly faster reading times inside the critical embedded clause; however, reading times may not reflect the effects of semantic, pragmatic, and/or discourse processing that extend beyond the first-pass reading of the sentence. Indeed, we find that the comprehension question response times for the baseline are about 270 milliseconds faster than those of the WHICH conditions. Since acceptability judgments constitute the end-result (or involve the input) of a number of cognitive processes, cognitive processes after the sentence has been read are also likely to have an effect on perceptions of well-formedness.

The acceptability data pair with the reading data to show that significantly reduced processing difficulty accompanies higher judgments of acceptability for island-violating examples. These data are consistent with the interpretation that CNPC violations receive low acceptability ratings due to multiple sources of processing difficulty. An alternative perspective, however, is that processing difficulty varies as a function of grammaticality. For now, we set this ambiguity aside, but we return to it in section 8 to discuss why the available evidence favors a causal story that begins with processing.

6 Experiment II: Wh-islands

In the second experiment, we consider whether *wh*-islands demonstrate the same sensitivity to filler complexity as reflected in the study of CNPC violations. As in the case of CNPCs, *Wh*-island violations are typically judged to be better when the displaced element is a *which*-$\overline{\text{N}}$ phrase, instead of a bare *wh*-item. To understand whether these differences are related to processing differences, we test the effect of such manipulations on comprehension. As in the previous experiment, a processing account of the previously noted acceptability differences predicts accompanying processing contrasts. Specifically, *which*-$\overline{\text{N}}$ phrases, as opposed to bare *wh*-words are predicted to lead to more efficient processing in island violations.

6.1 Materials & participants

All experimental items contained main clause interrogatives, as in 49. Subjects initially read a declarative background sentence. Afterwards, a basic comprehension question asked about the referent of the dislocated object NP. The targets of interest were the comprehension questions themselves; the initial context sentences justified the presence of these questions, which would be unnatural without any background.

Stimuli varied in terms of whether the sentence-initial *wh*-phrase was a bare *wh*-item *who* (BARE) or a more complex *which*-phrase (WHICH). An additional condition was included to serve as a baseline against which the results could be compared. This non-island baseline condition always began with *who*, but instead of a *wh*-phrase in the complementizer position, these baseline items had the complementizer *that*:

(49) Albert learned that the managers dismissed the employee with poor sales after the annual performance review.

BARE: Who did Albert learn whether they dismissed after the annual performance

review?
WHICH: Which employee did Albert learn whether they dismissed after the annual performance review?
BASELINE: Who did Albert learn that they dismissed after the annual performance review?

Twenty-four experimental items and forty-eight fillers constituted the materials for this study. Twelve of the fillers were *wh*-islands with varying lexical and structural properties to act as distractors. Hence, half of the overall items were *wh*-islands. Across the entire item set, an equal number of questions began with *who*, *what*, and *which*-phrases. After reading the question, subjects selected an answer from a set of alternatives provided. Of the three possible answers presented to them, one was correct (*the employee with poor sales*), another was lexically and syntactically similar (*the employee with poor hygiene*), and the third option differed drastically (*the cashier who stole money*). The materials for this experiment are included in Appendix A.

Twenty subjects participated in this second study. All participants were Stanford University undergraduates who received course credit for their participation and were naive to the purpose of the study.

6.2 Results

Reading times

The regions of interest for this experiment lie within the embedded *wh*-clause, as the left edge of this constituent marks the edge of the proposed *wh*-island. We focus in particular on reading times around the embedded verb where retrieval of the filler-phrase takes place, e.g. *dismissed* in 49.

The results verify that *which*-$\overline{\text{N}}$ phrases lead to faster reading times inside the embedded *wh*-clause. At the subject of the embedded clause, there is a slight difference between the BARE condition and the WHICH condition, but this effect is not significant either by subjects or by items, as shown in Table 2.[14] Similarly, at the embedded verb, there is only a nonsignificant trend for faster reading times in the WHICH condition. But at each of the three words immediately following the subcategorizing verb (the spillover regions), reading times are significantly faster in the WHICH condition, compared to the BARE condition, as pictured in Figure 4. Hence, when the spillover regions and the verb are considered together, the effect of *wh*-phrase complexity is observable and statistically significant ($t_1(19) = 3.339$, $p = .01$; $t_2(23) = 3.279$, $p = .01$).

[FIGURE 4 ABOUT HERE]

According to planned comparisons, the processing facilitation associated with the complex *which*-$\overline{\text{N}}$ phrases also eliminates any statistically significant difference (t-scores < 1) across these regions between the island-violating WHICH condition and the BASELINE, which does not violate any putative constraints on extraction. In contrast, the BASELINE is processed faster than the BARE condition across these regions ($t_1(19) = 2.266$, $p < .05$; $t_2(23) = 2.304$, $p < .05$).

[TABLE 2 ABOUT HERE]

Comprehension questions
The question-answer response times also reflect an advantage for the WHICH condition, compared to the BARE condition ($t_1(19) = 3.644$, $p < .01$, $t_2(23) = 2.560$, $p = .018$) (BARE = 2999.82 ms, *SE* = 104.73; WHICH = 2636.72 ms, *SE* = 88.22; BASELINE = 2789.28 ms, *SE* = 133.65). Given a *which*-$\overline{\text{N}}$ phrase in these contexts, as opposed to a bare *wh*-item, subjects answer a question with fewer possible answers, offering one reasonable explanation of this difference. The likelihood of lexical priming is also quite strong, given the layout and presentation of the stimuli. The presence of the head noun in the WHICH conditions and in the answers potentially facilitates question answering—participants may pay less attention to the distractor. Hence, while the question-answer response times also suggest more efficient processing in the WHICH condition, this conclusion is confounded by lexical repetition in the question and answer. Question-answer accuracies did not vary significantly across conditions within the experimental items (BARE = 89.6%, *SE* = 2.21; WHICH = 88.5%, *SE* = 2.31; BASELINE = 90.6%, *SE* = 2.99).

6.3 Discussion

Filler-gap processing, according to the evidence, improves with the syntactic and semantic complexity of the filler-phrase. The processing advantage, however, does not begin immediately after the filler-phrase. In fact, the more complex *wh*-form initially causes slower reading, as illustrated in Figure 4, likely due to the added task of integrating the information from the head noun of the *which*-$\overline{\text{N}}$ phrase – essentially, the cost of building a more complex representation.

The reading time differences found at the post-verbal spillover regions are predictable if one assumes either that more complex *wh*-phrases facilitate retrieval or that they can reduce the computational effort of answering the question. Since both hypotheses are compatible with processing effects emerging at or around the subcategorizing verbs, the present results do not uniquely support one analysis over the other. What is important for the current discussion is that both hypotheses predict facilitated processing for complex *wh*-phrases at the critical regions.

The data also show that processing improves so much that the overall difference between the non-island BASELINE and the WHICH condition disappears. This finding aligns with the intuitions expressed in the theoretical syntax literature that extraction out of island contexts improves with the specificity of the extracted element. The absence of a difference between the BASELINE and WHICH conditions is notable for another reason: some small but significant decrease in processing difficulty may be viewed as immaterial if processing difficulty remains at an extremely high level. Given the fact that reading times in the WHICH condition are essentially equivalent to those of the baseline, however, the evidence argues for a meaningful interpretation of the facilitation.

Because the BASELINE and WHICH conditions produce similar reading times, both significantly faster than the BARE condition, the perception of unacceptability for *Wh*-island violations involving vague or nonspecific filler-phrases may reasonably have its origins in

processing-related difficulties. The evidence at hand, therefore, suggests an alternative to the putative grammatical mechanisms used to explain the previously unsubstantiated contrasts. Specifically, to the extent that acceptability is derived from considerations of processing difficulty, the observed processing contrast stemming from the complexity of the extracted *wh*-phrase may be playing a large role in the perception of the acceptability of the entire sentence.

To test the hypothesis that acceptability improves where processing difficulty decreases, a separate acceptability study was run using twenty of the island-violating items from the reading-time task, with one alteration. To remove the pragmatic oddity of decontextualized questions, the *wh*-islands were presented as embedded questions, as in the following modified version of 49:

(50) a. Only a few individuals repeated who Albert learned whether we dismissed after the annual performance evaluations.

b. Only a few individuals repeated which employee Albert learned whether we dismissed after the annual performance evaluations.

[FIGURE 5 ABOUT HERE]

All stimuli began with vague, quantified NP subjects, such as *some people* or *no one*, where no particular individual is named or evoked, under the assumption that such NPs would incur fewer processing costs than NPs naming a specific individual or type of individual. Sixteen Stanford University participants were instructed to rate how natural the examples sounded as sentences of English on a scale of 1-7 (7 being perfectly natural). As depicted in Figure 5, the more complex *wh*-phrases significantly improve judgments of acceptability ($F1(1,15) = 15.964$, $p = .001$; $F2(1,19) = 14.428$, $p = .001$).

This study therefore verifies that acceptability judgments fall where processing difficulty significantly increases. The parallel between acceptability and processing difficulty in *wh*-islands echoes the results of the CNPC study. Of course, this interpretation of the data could be turned on its head: the evidence could be interpreted as showing that ungrammaticality predicts processing difficulty. The issue of how to interpret the relationship between acceptability and processing results is fundamental, but before we address it we consider a final issue in the processing of islands.

7 Experiment III: Adjunct extraction

The third experiment we discuss here addresses the referentiality of the displaced element. This island study supplements the previous investigations by examining whether effects based on complexity are restricted to referential arguments. Cinque, Rizzi, and others have interpreted differential acceptability in island contexts as a function of referentiality. According to such theories, nonreferential adjuncts that differ in syntactic and semantic complexity should not produce the same effects that have been observed to distinguish putatively referential (*which*-$\overline{\text{N}}$ phrases) and nonreferential (*who*) phrases.

In contrast, a processing-based explanation for these acceptability differences avoids placing boundaries on the types of syntactic entities for which the cognitive constraints are

relevant. To the extent that these processing differences derive from general resource considerations, e.g. working memory limitations, referentiality should not be a necessary precondition for the observation of these complexity-based contrasts. Hence, we predict that a similar sort of processing facilitation should occur for more complex filler-phrases, even when these phrases constitute nonreferential adjuncts. To test this hypothesis, we consider here dependencies involving adjunct phrases extracted out of *wh*-islands.

7.1 Materials & participants

Twenty-four adjunct extractions from *wh*-islands were constructed for this study. Stimuli were systematically varied in terms of the syntactic and semantic complexity of the extracted temporal adjunct phrase. In the SIMPLE condition, the temporal adjuncts contained only two words, either *how long* or *how often*. The COMPLEX condition had longer temporal adjuncts of at least three words and as many as eight, as illustrated in 51 below. As in the previous experiment, a baseline condition containing a different lexical complementizer, *that*, was also included in the study as a means of evaluating differences between the two main conditions of interest. The baseline condition always contained the shorter adjunct phrase that appeared in the SIMPLE condition.

(51) Julie discerned that the survivor had managed to stay alive for eight days after the crash in the harsh conditions.

SIMPLE: How long did Julie observe whether the passenger had survived in the unbelievably harsh conditions?
COMPLEX: For what period of time after the crash did Julie observe whether the passenger had survived in the unbelievably harsh conditions?
BASELINE: How long did Julie observe that the passenger had survived in the unbelievably harsh conditions?

(52) Andrew overheard the daycare staff discussing how they wanted to get away from the children for a few hours.

SIMPLE: How long did Andrew hear whether the children had played during the daycare's afternoon recess?
COMPLEX: How many hours did Andrew hear whether the children had played during the daycare's afternoon recess?
BASELINE: How long did Andrew hear that the children had played during the daycare's afternoon recess?

In contrast with the previous two experiments, subjects were instructed not to answer the comprehension question, but to indicate whether the question *could be* answered, given the information stated in the context sentence. Half of the total items appearing in this experiment, including half of the twenty-four adjunct extractions, presented questions which could not be reasonably answered given the preceding text. For instance, in the item shown in 52, none of the three versions of the question can be answered, since the preceding text

makes no mention of how long the children played during the afternoon recess. Thus, after each question, the subjects saw the prompt, *Is it possible to answer the question?*, and were instructed to provide a negative response when the preceding text did not contain the necessary information to answer the question. Subjects were informed during the training session of this experiment that each item did have a correct answer, and consequently received negative feedback if they answered the question incorrectly. Prior to the presentation of the actual experimental stimuli, subjects became familiar with this task via a practice session with eight items.

This methodology encourages participants to read the comprehension questions carefully. Generally speaking, subjects read comprehension questions faster than the preceding text, partly due to a relatively high degree of lexical overlap and the predictability of upcoming constituents. The methodology employed in this experiment consequently dissuades participants from relying on predictability and repetition in reading and answering the comprehension question itself. Twenty-eight Stanford University students were paid $10 for their participation in this study.

7.2 Results

Reading times

As in the previous experiment, we concentrate here on the reading times within the embedded clause. At the clause boundary, the complementizer *whether*, appearing in both the COMPLEX and SIMPLE conditions, results in faster residual reading times than the complementizer *that* in the BASELINE condition (SIMPLE: -116.05 ms, SE = 10.66; COMPLEX: -138.40, SE = 8.58; BASELINE: -80.13, SE = 6.33); however, this slowdown does not appear in the raw reading times (SIMPLE: 339.93 ms, SE = 10.50; COMPLEX: 322.10 ms, SE = 8.73; BASELINE: 310.11 ms, SE = 7.30) – hence, it implies only that the complementizer *that* was read more slowly than predicted for a word that is four characters long.

[FIGURE 6 ABOUT HERE]

The data identify only one particular word region where the complexity manipulation creates significantly different reading times – at the word after the complementizer, which was always the determiner *the*, as shown in Figure 6. Here, the SIMPLE condition produces significantly slower reading times than the COMPLEX condition by roughly 40 milliseconds ($t_1(27) = 3.484$, $p < .01$; $t_2(23) = 3.513$, $p < .001$). The BASELINE patterns with the COMPLEX condition – reading times for the baseline at the first word of the embedded clause are substantially faster on average than those for the SIMPLE condition (mean residual reading times = SIMPLE: -12.37 ms, $SE = 10.09$; COMPLEX: -53.83 ms, $SE = 6.11$; BASELINE: -61.54 ms, $SE = 5.33$; mean raw reading times = SIMPLE: 350.83 ms, $SE = 10.65$; COMPLEX: 311.44 ms, $SE = 7.45$; BASELINE: 300.01 ms, $SE = 6.79$.) Notice that the reading time difference cannot be simply ascribed to the number of words read previously, i.e. string position, as the BASELINE and SIMPLE conditions contain the same number of words.

Unlike the previous two experiments, no effect of complexity is observed at the embedded verb or subsequent regions. In fact, after the first word of the embedded clause, the COMPLEX condition is not processed significantly faster than the SIMPLE condition at any one

particular word inside the embedded clause. Additionally, while reading times for the entire embedded clause are faster for the BASELINE, as compared to the SIMPLE condition ($t_1(27) = 3.218$, $p < .01$; $t_2(23) = 2.518$, $p = .019$), the BASELINE did not produce reading times in the embedded clause that were significantly faster than the COMPLEX condition (each $t < 1$).

Comprehension questions

Question-answer response times did not differ significantly by condition (SIMPLE = 871.37, $SE = 38.19$; BASELINE = 860.84, $SE = 44.23$; COMPLEX = 806.81, $SE = 32.72$). As in the previous studies, there were no significant comprehension accuracy differences across conditions (SIMPLE = 84.82%; $SE = 2.40$; COMPLEX: 86.16%; $SE = 2.31$; BASELINE: 84.82%, $SE = 2.40$).

7.3 Discussion

As in the experiments discussed previously, greater syntactic and semantic complexity in dislocated adjunct phrases significantly facilitates subsequent processing. At the embedded clause boundary, the reading time measures indicate a highly significant advantage for the complex adjuncts. The effect of complexity on processing, according to this evidence, operates independently of referentiality and also appears to be generally insensitive to the argument-adjunct distinction. If temporal adjunct phrases refer to nonreferential entities, then varying the complexity of said phrases should not affect the ability to extract them out of islands, according to syntactic theories of relativized movement. The results, however, support the conclusion that adjunct phrases are subject to some of the same principles of sentence processing as argument phrases.

Unlike the previous experiments, the complexity-based effect in this experiment emerges just after the clause boundary, rather than at the embedded verb. Where we expect to find a complexity-based effect, however, partly depends upon where retrieval is theorized to occur. If dislocated adjuncts are retrieved at clause boundaries, then the results are unsurprising. In fact, the theoretical syntactic and semantic literatures suggest that adjuncts, including temporal adjuncts, may modify entire CPs or IPs (Hitzeman 1997; Ernst 2002). Such research supports the idea that a dislocated adjunct could be retrieved at the boundary of a clause to modulate its entire meaning. Indeed, if something akin to an active-filler strategy applies to adjuncts as it does for arguments, then the processor would retrieve and attempt to integrate the relevant adjunct information at the first *possible* site. This would mean that, even if retrieval and integration at some downstream point (say, at an embedded verb) was also possible, retrieval would preferentially take place at any plausible, earlier attachment site, e.g. a clause boundary. Nevertheless, additional research is clearly needed to determine when stored information related to adjuncts is generally retrieved, particularly when the adjunct modifies an embedded constituent.

While showing that adjunct dependencies and argument dependencies behave alike, this study does not address why displacement of adjuncts out of *wh*-islands is less acceptable than displacement of arguments. A complete answer to this question goes beyond the scope of this paper. There are, however, a number of reasons to suspect that this difference is attributable to processing-based differences as well: (1) retrieval cues carried by verbs with

missing arguments (e.g. subcategorization and thematic role information) may be better suited for recovering arguments than adjuncts; (2) displaced arguments can provide indirect evidence for the retrieval site in the form of a missing obligatory constituent, but displaced adjuncts do not provide such evidence since they are optional; (3) fronted adjuncts will often be able to modify intervening verb phrases along the filler-gap path, thus raising the probability of a subsequent reanalysis. Beyond these issues, the mental representations associated with adjuncts may be generally harder to retrieve. Hence, it is reasonable to assume that adjunct dependencies present an assortment of processing difficulties that are absent in argument dependencies.

There are a number of open areas of investigation for the processing of adjunct dependencies, but this third experiment clearly confirms that adjunct dependencies are also sensitive to the complexity of the dislocated phrase. As in the first two experiments, more complex filler-phrases produce faster processing inside the syntactic island. These results are not easily explicable under accounts that state movement constraints on the basis of referentiality or other categorical divisions among phrase types.

8 Performance vs. competence

These empirical investigations into the processing of island constraint violations converge on the conclusion that where processing difficulty increases, acceptability decreases. Crucially, the results suggest that these variations in processing difficulty are not insubstantial fluctuations – they can effectively eliminate the processing difference between islands and non-islands. One interpretation of these facts – the one that we adopt here – is that processing differences are influencing perceptions of acceptability. A competing alternative, however, is that it is the grammar itself (competence-based factors) which decides the degree of processing difficulty in these FGDs.[15] On such an account, less acceptable or grammatical constructions would lead to greater processing difficulty. In this section, we consider a number of points that favor the conclusion that processing effort underlies the acceptability differences, rather than the other way around. Taken together, these arguments strongly favor a causal explanation that begins with processing, while a number of theoretical difficulties arise if it starts with grammar.

First among these points is this: a processing-based explanation appeals to the existence of independently-motivated factors, some of which may not even be specific to language. Locality or distance-based effects, referential processing load, and other relevant processing factors have been identified and substantiated outside the domain of island constructions as reliable predictors of processing load. They are, in other words, a necessary part of any general theory of sentence processing. In this sense, the explanation is 'cost-free' – no new, arbitrary, and language-specific constraints have to be introduced.

As noted earlier, the complexity-based effects observed here are found in non-island contexts. For instance, even in fully grammatical interrogatives like those in 53, reading time measures are faster at the retrieval site (e.g. *record* below) when the sentence begins with the more complex *which*-$\overline{\text{N}}$ phrase (Hofmeister 2007):

(53) a. Which album did the musician that Robert saw record with two popular blues guitarists?

b. What did the musician that Robert saw record with two popular blues guitarists?

The existence of these complexity effects in non-island contexts poses a problem for any grammar-based account. If grammaticality differences are to explain the processing differences in islands, then some secondary explanation must be invoked to account for the same processing differences in syntactic contexts which do not contain violations of any known grammatical constraint. Perhaps nothing beyond considerations of elegance and good taste stands in the way of a grammarian who chooses to postulate intricate, noncategorical, and relativized rules that state that dependencies with more complex fillers are more acceptable than dependencies with less complex fillers. For that matter, a grammar of long-distance dependencies might also stipulate that shorter dependencies are more acceptable than longer dependencies.

In contrast to an explanation couched in terms of processing, a grammar-based characterization of the amply documented graded acceptability of sentences with FGDs would therefore require some highly specialized linguistic machinery. This may include FGD-specific constraints like Subjacency, as well as a system for calculating fine-grained acceptability differences. Notably, such a calculator would in all likelihood serve no function other than to tally the result of aggregating grammaticality violations. Additionally, such a system must also provide for the possibility that the interaction of grammatical constraints can lead to interactions in sentence processing.

Judgments concerning islands are also known to vary widely across and even within individuals (Braze 2002; but see Sprouse 2009 for an alternative perspective). Looking more closely at the acceptability judgments for CNPC violations discussed in section 5, for instance, judgments of CNPC violations systematically rise throughout the course of the experiment, as shown in Figure 7.

[FIGURE 7 ABOUT HERE]

This graph illustrates that a CNPC violation is more likely to be judged higher, the later it is presented. Linear mixed effects modeling, where subjects and items can be simultaneously treated as random factors and which allows for a principled way of incorporating longitudinal effects and covariates into the analysis (Pinheiro & Bates 2000; Baayen 2004, 2008), demonstrates that this effect is highly significant ($\beta = .007$, HPD95lower = .004, HPD95upper = .010, p$MCMC < .001$) (Markov chain Monte Carlo (MCMC) sampling (n = 10,000) was used to estimate p-values; HPD (Highest Posterior Density) values indicate upper and lower confidence intervals). Later list positions also increase the probability that other CNPC violations have already been seen by the same individual. Hence, one way of interpreting this relationship is to say that increased exposure to CNPC violations causes acceptability judgments for CNPC violations to rise. Supporting evidence for this interpretation comes from examining the effect of list position on acceptability judgments for each individual subject. As is evident from the subject-by-subject plots in Figure 8, most participants judge CNPC violations increasingly better as they see more and more such violations.

[FIGURE 8 ABOUT HERE]

Notably, some individuals seem fairly accepting of island violations, while others reject the same tokens. This type of variation in acceptability judgments, both within and across subjects emerges naturally on the processing account of islands. Individuals are known to differ significantly from one another in terms of working memory capacity (Daneman & Carpenter 1980; King & Just 1991; Just & Carpenter 1992), and the same individual may have more or fewer resources available, depending upon factors such as fatigue, distractions, or other concurrent tasks. Moreover, exposure to a certain type of linguistic stimulus can make it easier to process the next time a similar stimulus is encountered. This can theoretically account for why the same individual can perceive islands differently over time: island violations become easier to process as familiarity increases, resulting in judgments of higher acceptability. Unlike a performance-based analysis, a (nonstochastic) competence account appears fundamentally incapable of modeling such changes.

Finally, if one of the goals of linguistic inquiry is to explain why we have the linguistic constraints that we do, the performance-based view accounts for islands as the byproduct of general principles of cognition. In contrast, the competence-based theory offers no insight into *why* islands exist, at least at this time. One might reasonably argue, of course, that it is outside the domain of syntactic theory to posit a reasonable account of how a given syntactic constraint might have evolved or the functional motivations for its existence. However, a perspective on islands that depends on the findings of psycholinguistic research links the acceptability of islands to factors that generally affect sentence processing and acceptability. In other words, a performance-based approach not only accounts for the variation, it explains why it exists in the first place.

Although the advantages of the approach we suggest are considerable, various researchers have presented acceptability and parsing data that have been interpreted as being in conflict with a processing-based analysis of island phenomena (Stowe 1986; Pickering *et al.* 1994; Traxler & Pickering 1996; McElree & Griffith 1998; Phillips 2007). In general, it has been argued that these studies demonstrate that 'island constraints are immediately effective in parsing' (Phillips 2007, p. 800), often on the basis of evidence showing that the parser behaves differently in island contexts as opposed to non-island contexts. For instance, Stowe (1986) and others have shown that reading times increase upon evidence of a 'filled-gap' site in non-island contexts, i.e. a syntactic position with which an element stored in memory could be associated, but which is filled with an overt element, as in 54a, compared to a minimally different variant without a *wh*-dependency 54b:

(54) 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.

Crucially, there are no significant, relative increases in reading times at these filled-gap sites in island environments, which is taken as evidence that 'no gap site is posited' (Phillips 2007, p. 798).

It should first of all be noted that these claims are often made with respect to subject or relative clause islands, which we have said little about (cf. Kluender 2005). Hence, it

may well be the case, as already noted, that certain island phenomena are best explained via syntactic constraints, while others may be better and more easily reconciled with processing-based considerations. In addition, the absence of a filled-gap effect in island environments does not uniquely support the conclusion that a dependency *cannot* be made into such an environment (Pickering *et al.* 1994). Such effects are also compatible with the possibility that dependencies into certain environments are strongly dispreferred when other alternatives are simultaneously available and preferred. Along these lines, Pickering & Traxler (2001) examine sentences like 55, where the filler could initially be associated with the direct object position of the verb *persuaded*, although it ultimately acts as the object of the infinitival complement:

(55) 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.

Encountering the overt NP, *a few pupils*, provides direct evidence against the first analysis. But while the filler could plausibly be interpreted as the object of the first verb in 55a, such a reading is implausible in 55b. Participants in Pickering and Traxler's study exhibited greater reading difficulty at *a few pupils* in 55a, compared to 55b, but did not show signs of a plausibility effect at the preceding verb. These results imply that readers considered the object analysis in 55a, but not (as much) in 55b. More importantly for the present discussion, the absence of significant processing difficulty in 55b is consistent with the idea that the processor considers multiple filler assignments in parallel. Since the direct object analysis becomes implausible in 55b by the time the verb *persuaded* is processed, the parser appears to abandon this option in favor of the infinitival complement analysis and thus does not experience difficulty at the subsequent NP. Notably, the absence of a filled-gap effect in 55b does not warrant the conclusion that syntactic restrictions rule out the formation of a dependency between the filler and the direct object position, since the same hierarchical structure in 55a shows evidence of such a dependency.

A similar type of reasoning can be applied to the case of (the absence of) filled-gap effects in subject islands: signs of creating a dependency may not be apparent because a preferred, alternative analysis associates the filler with a position outside the island. For instance, the parser may prefer associating a filler-phrase with a syntactic position at the same level of embedding wherever possible. Hence, positing gaps inside of a subject NP would be dispreferred when a gap at the same layer of embedding was still possible, e.g. an object gap. This would not rule out the possibility that dependencies could be made into these islands environments, as in parasitic gap constructions or attested examples like 56, but that such a parse is not generally preferred in most circumstances:

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

[*Butler's Lives of the Saints* (p. 426), Alban Butler and Michael J. Walsh (1991)]

Thus, the question of how to interpret the data for subject islands and filled-gap effects remains open in our view and awaits further investigation. However, the data for CNPC and

Wh-island violations strongly support the conclusion that filler-phrases are being retrieved even in difficult-to-process contexts, since question-answer accuracies are comparable to those of the baseline non-island violations. If the parser completely disallowed the creation of dependencies into CNPCs and *wh*-islands, we would at a minimum expect some significant decrement in comprehension accuracy for the island-violating items, as compared to the baselines.

While these arguments favor the conclusion that processing factors are responsible for much of the acceptability variation tied to the island violations considered here, they do not rule out the possibility that the grammar itself is responsible for some of the variation. Indeed, it is impossible to prove, given the current state of our knowledge of both grammar and processing, that grammar has no part in creating the observed processing differences. But the simplification of the grammar that the processing perspective allows (e.g. the elimination of the Subjacency Condition), taken together with its reliance on independently motivated properties of language processing, make it by far the more attractive hypothesis.

9 Conclusion

The grammatical constraints that have been proposed to account for syntactic islands are almost uniformly complex, arbitrary, and ultimately either too strong, too weak, or both. They express intricate and highly specific limitations on just a subset of the linguistic dependencies possible in natural language. They are arbitrary in the sense that they bear no relationship to other constraints, emanate from no general principles of language, and have no relevance or parallel outside language. In short, syntactic island constraints mark islands as special within the domain of language, and even more particularly, within the domain of linguistic dependencies. Consequently, island constraints offer little insight into anything about language or cognition, except islands themselves.

At the same time, the island constraints that have been proposed by linguists (at least those we have examined here: the Subjacency Condition, the *Wh*-Island Constraint, and/or the Complex Noun Phrase Constraint) face serious empirical difficulties. Stated as categorical constraints on dependencies, syntactic formulations of island constraints have been counterexemplified on countless occasions over the past half-century. Within and across languages, nearly every proposed structural island constraint has been shown to be violable in at least some circumstances. Judgments have also been shown to vary on the basis of factors that do not alter the structure upon which the island constraint is based. Historically, much of the variation surrounding judgments of critical island-related data has been set aside as 'exceptional' or irrelevant. Without such uncritical and unmotivated data triage, which we have surveyed in some detail, most syntactic theories of islands would have been falsified long ago.

Our approach here has been to legitimize this variation and to find a more adequate means for analyzing it. In so doing, we have proposed a way of eliminating theoretically central island constraints from grammar. This welcome simplification of the theory of grammar interacts with independently-motivated cognitive constraints to predict observed contrasts in reading time data and acceptability judgments. This approach, which acknowledges the role of cognitive constraints on perceptions of acceptability, has been generally underrepresented

in the syntactic literature. Despite the early recognition that performance-related factors can influence acceptability, very few syntactic analyses (since Miller & Chomsky 1963) have argued for theoretically relevant, performance-based distinctions. In essence, a major explanatory tool for linguistic behavior has been seriously underutilized. Given the prospect of accounting for a substantial amount of island variation using this tool, it is quite possible that a similar strategy can be applied to other cases of judgment variation. Particularly when the data surrounding a grammatical construction exhibit a large amount of variation and when non-structural choices have major effects on acceptability and other behavioral measures, appeals to constraints on sentence processing may ultimately offer the most elucidating and economical explanation.

Notes

[1]Not all gap-binding dependencies involve filler-phrases (e.g. bare relative clauses, *tough*-constructions, etc.); we retain the 'filler-gap' terminology as a matter of expository convenience. For a recent comprehensive account of filler-gap constructions and a discussion of their relation to other means of gap binding, see Sag to appear.

[2]Although Ross is often criticized for proposing ad hoc constraints of insufficient generality, Ross 1967 includes an attempt to unite many of his proposed constraints via an 'A *directly* over A' principle.

[3]Chomsky is suggesting that Subjacency violations beyond 1-subjacency should be indiscriminable in terms of acceptability, because 'to specify n-subjacency for higher values of n requires counters'. In other words, a dependency that cross two barriers should incur the same degradation in acceptability as a dependency that crosses three or four barriers.

[4]These languages are by no means uniform with respect to island constraint violations. For example, some readily permit dependencies into *wh*-islands, but seem to exhibit familiar acceptability degredation in other island constructions.

[5]PERF = perfective; A3= cross reference set A, third person; CMP = completive, D2 = distal deixis (clause final particle); FEM = feminine; QPART = interrogative/disjunction/conditional protasis

[6]Manzini (1992) attempts to unify the effects of both tense and definiteness on islands. Her suggestion is that both tense and definiteness block movement out of islands, because they interpose a conflicting case address. However, despite the evidence that definite interveners and tensed clauses make island-violating dependencies *worse*, we note that finiteness and definiteness do not seem to be categorical predictors of grammaticality, even in island contexts, as long as other processing burdens are removed.

[7]However, as noted by Szabolcsi (2006), this 'dichotomy is not particularly straightforward; moreover, the borderline between strong and weak islands is not very firm, there being a number of intriguing empirical parallels between the two.'

[8]Warren and Gibson (2005) acknowledge that the similarity of discourse references also plays a critical role in determining the cost of memory retrieval, as demonstrated by Gordon *et al.* (2002). Similarly, although we regard retrieval difficulty mainly in terms of activation decay here, similarity-based interference can account for many of the same effects.

[9]There are, however, some important exceptions to this principle. Generally, these exceptions involve material interceding between filler and gap that increases the predictability of the subcategorizing head or else improves the retrievability of the dependent phrase (see, for instance, Vasishth & Lewis 2006).

[10]Other discourse considerations also factor into the cost of definites versus indefinites (and other phrasal types). For instance, old or given information is more likely to appear early in the sentence, while new information typically shows up later in the sentence. Hence, indefinite subjects can be a source of processing difficulty under certain discourse conditions.

[11]In this vein, Sprouse (2007) shows that sentences with *wh*-dependencies into embedded clauses yield lower judgments than those with *wh*-dependencies at the matrix level; similarly, sentences with embedded *wh*-interrogatives produce lower judgments than sentences with embedded propositions. A *wh*-dependency into an embedded clause that is interrogative (a *Wh*-island violation), however, lowers judgments more than one would expect based on the effect of each factor in isolation. This is taken as evidence that something above and beyond processing costs (i.e. grammar) contributes to judgments of island violations, but this conclusion crucially assumes that processing costs are strictly additive. In fact, this evidence of an interaction aligns naturally with the position that processing costs can combine superadditively.

[12]Linear mixed effects modeling, which is robust in spite of missing data and does not require any imputation (Pinheiro & Bates 2000; Baayen 2004, 2008), shows highly similar results. For the same region – the complementizer to the verbal spillover regions – *wh*-complexity is a highly significant predictor of reading times ($\beta = -36.056$, HPD95lower $= -35.952$, HPD95upper $= -21.052$, p$MCMC = .0001$). The conservative p-value is based on Markov chain Monte Carlo sampling ($n = 10{,}000$), and the HPD (Highest Posterior Density) values list the upper and lower confidence intervals for the estimate of the parameter. The negative co-efficient (β) here indicates that the WHICH conditions were read faster than the BARE conditions. The effect of NP type, in contrast, is not significant, nor is the interaction of NP type and *wh*-complexity. Additional modeling indicates that reading times do not vary significantly as a function of experimental list, when treating list as a random effect factor. This holds true even when including by-list random slopes for the effects of complexity and NP type. Similarly, for Experiments 2 & 3, comparisons of linear models demonstrates that including list as a random effect (with or without random slopes for complexity) does not significantly contribute any explanatory value.

[13]Greenhouse-Geisser corrections are used for the NP type statistics from the verb and complementizer, due to violations of sphericity.

[14]Reading times in the critical regions are relatively fast, but this is unsurprising since the questions have considerable lexical overlap with the preceding sentences.

[15]Yet another possibility is that the parallel results are merely coincidental and do not reflect a relationship between the two response types. Given that these parallels have been replicated in numerous contexts, including islands and non-island environments, as well as the well-motivated belief that acceptability judgments are influenced by processing, we consider it unlikely that these parallels are accidental.

A Appendix

A.1 Experiment I Stimuli

1. He knew which country/who Emily heard (a/the rumor(s)) that we had invaded due to increased political instability.
2. She discovered which passage/what Jacob read (a/the allegation(s)) that they had copied into the final written report.
3. I saw which convict/who Emma doubted (a/the report(s)) that we had captured in the nationwide FBI manhunt.
4. She remembered which article/what Michael denied (a/the suggestion(s)) that they had plagiarized in order to sound intelligent.
5. He forgot which song/what Jessica reiterated (a/the contention(s)) that we had stolen from the original German composer.
6. I verified which patient/who Chris held (a/the conviction(s)) that they had cured with the new experimental treatment.
7. He realized which prisoner/who Ashley countered (a/the belief(s)) that we had interrogated without regard to international law.
8. She wondered which company/who Matthew confirmed (a/the suspicion(s)) that they had sued for its unethical accounting practices.
9. I researched which student/who Amanda conveyed (a/the threat(s)) that we would reject despite an outstanding academic record.
10. She guessed which client/who Joshua disputed (a/the notion(s)) that they had defended before the federal appeals court.
11. He insinuated which actor/who Jennifer overheard (a/the comments(s)) that we had arrested for drunk driving last night.
12. I acknowledged which novel/what David expressed (a/the worr(y)/(ies)) that they would ban due to its racy content.
13. He indicated which concert/what Sarah answered (a/the objection(s)) that we had canceled unnecessarily because of a disagreement.
14. She understood which intern/who Daniel envisioned (a/the prospect(s)) that they would hire for the emergency room position.
15. I surmised which agency/who Erin echoed (a/the complaint(s)) that we had overcharged for a routine financial report.
16. She testified which supervisor/who James established (a/the expectation(s)) that they would fire because of his lewd behavior.
17. He determined which project/what Nicole shared (a/the intuition(s)) that we would complete with extra time to spare.
18. I learned which route/what Andrew proposed (a/the hypothes(is)/(es)) that they had used to cross the mountain range.
19. He published which structure/what Brittany announced (a/the plan(s)) that we would build to replace the condemned building.
20. She perceived which river/what Robert addressed (a/the fear(s)) that they had polluted with dangerous levels of toxins.

21. I recorded which dictator/who Heather contested (an/the assertion(s)) that we had supported in his rise to power.
22. She uncovered which quality/what John debated (a/the perception(s)) that they had lacked in the home loan application.
23. He confirmed which student/who Elizabeth believed (a/the charge(s)) that we had suspended due to their poor grades.
24. I specified which car/what Ryan considered (a/the demand(s)) that they should recall because of failed safety tests.
25. He divulged which base/what Megan recommended (a/the proposal(s)) that we should abandon in order to minimize casualties.
26. She recalled which leak/what Joseph appreciated (a/the comment(s)) that they would repair within a week from now.
27. I proved which mineral/what Melissa relayed (a/the message(s)) that we had identified on the surface of Mars.
28. She admitted which community/who Brandon printed (a/the warning(s)) that they would forget in the disaster relief effort.
29. He confided which species/what Amber verified (a/the theor(y)/(ies)) that we would discover with enough time and energy.
30. I asked which project/what Justin recognized (a/the concern(s)) that they should coordinate after a series of disasters.
31. He investigated which election/what Lauren repeated (a/the claim(s)) that we had rigged in favor of the Democrats.
32. She decided which company/who William concealed (a/the sign(s)) that they had ruined with numerous illegal takeover attempts.
33. I resolved which spacecraft/what Rachel noted (a/the signal(s)) that we had lost due to an insulation problem.
34. She perceived which tax/what John protested (a/the request(s)) that they should pay for the next ten years.
35. He studied which resource/what Danielle confessed (a/the feeling(s)) that we had depleted over many years of mismanagement.
36. I noticed which player/who Nick disregarded (a/the comment(s)) that they would lose because of a serious injury.

A.2 Experiment II Stimuli

1. Kathy wondered if her friends consulted the doctor from New Madrid at the hospital last night.
Who/Which doctor did Kathy wonder whether they consulted at the hospital last night?
2. Nathan mentioned that the generals promoted the sergeant from western Tennessee during the war with Iraq.
Who/Which sergeant did Nathan mention whether they promoted during the war with Iraq?
3. Winston stated that the prosecutors trained the assistant who needed practice shortly before the trial began.

Who/Which assistant did Winston state whether they trained shortly before the trial began?
4. Sarah wondered if the voters elected the senator from Portland, Oregon in spite of the scandal.
Who/Which senator did Sarah wonder whether they elected in spite of the scandal?
5. Stephen confirmed that the radicals released the prisoner from New Zealand just after the demands were met.
Who/Which prisoner did Stephen confirm whether they released after the demands were met?
6. Brandy pondered if the investigators identified the suspect in Wednesday's robbery at the station near Scranton.
Who/Which suspect did Brandy ponder whether they identified at the station near Scranton?
7. Jason learned that the doctors examined the patient with back pain very quickly at the clinic.
Who/Which patient did Jason learn whether they examined very quickly at the clinic?
8. Charlotte pondered if the delegates nominated the candidate who was liberal at the recent Republican convention.
Who/Which candidate did Charlotte ponder whether they nominated at the recent Republican convention?
9. Marvin mentioned that the violinists accompanied the pianist who played Beethoven at the concert on Sunday.
Who/Which pianist did Marvin mention whether they accompanied at the concert on Sunday?
10. Maureen speculated that the TAs tutored the student who was failing before the final physics exam.
Who/Which student did Maureen speculate whether they tutored before the final physics exam?
11. Albert learned that the managers dismissed the employee with poor sales after the annual performance evaluations.
Who/Which employee did Albert learn whether they dismissed after the annual performance evaluations?
12. Anna said that the members rejected the applicant from Yale University at the meeting about admissions.
Who/Which applicant did Anna say whether they rejected at the meeting about admissions?
13. Oscar stated that the detectives dispatched the officer with little experience after the tragic murder yesterday.
Who/Which officer did Oscar state whether they dispatched after the tragic murder yesterday?
14. Vera indicated that the Yankees retired the batter with two homers after the previous hitter doubled.
Who/Which batter did Vera indicate whether they retired after the previous hitter doubled?
‘ 15. Victor announced that the defendants intimidated the witness who was twenty with numerous threats of violence.
Who/Which witness did Victor announce whether they intimidated with numerous threats of violence?

16. Matthew verified that the landlords evicted the tenant who wasn't quiet to satisfy the other residents.
Who/Which tenant did Matthew verify whether they evicted to satisfy the other residents?
17. Susan disclosed that the parents adopted the child who was sick after making several failed attempts.
Who/Which child did Susan disclose whether they adopted after making several failed attempts?
18. Eric announced that the agents arrested the criminal who escaped yesterday at a motel in Ohio.
Who/Which criminal did Eric announce whether they arrested at a motel in Ohio?
19. Erin disclosed that the terrorists killed the hostage from the U.S. in a moment of panic.
Who/Which hostage did Erin disclose whether they killed in a moment of panic?
20. Thomas indicated that the professors taught the graduate who studied psychology for at least two years.
Who/Which graduate did Thomas indicate whether they taught for at least two years?
21. Thelma speculated that the jurors acquitted the defendant charged with arson after discussing all the evidence.
Who/Which defendant did Thelma speculate whether they acquitted after discussing all the evidence?
22. Ryan verified that the zealots followed the leader of the cult to the islands without forethought.
Who/Which leader did Ryan verify whether they followed to the islands without forethought?
23. Rachel said that the cornerbacks sacked the quarterback for the Seahawks multiple times during the game.
Who/Which quarterback did Rachel say whether they sacked multiple times during the game?
24. Crystal confirmed that the teams rescued the survivor with severe injuries from the wreckage without difficulty.
Who/Which survivor did Crystal confirm whether they rescued from the wreckage without difficulty?

A.3 Experiment III Stimuli

1. Katherine was informed that the salesman often only got four hours of rest per night because of his insomnia.
(How long/How many hours per night) did Katherine find out whether the salesman had slept on account of his schedule?
2. Hilary appreciated the fact that the foreman was going to keep talking about his personal life until the work day was over.
(How long/For how much of the day) did Hilary understand whether the foreman would work before finally leaving for home?
3. Jack was aware that, twice a day, the sheriff had stopped by to gather additional information about the burglary.

(How often/How many times a day) did Jack know whether the sheriff had come to inquire about the burglary?
4. Julie discerned that the survivor had managed to stay alive for eight days after the crash in the harsh conditions.
(How long/For what period of time after the crash) did Julie observe whether the passenger had survived in the unbelievably harsh conditions?
5. Stephen was informed that the representative had gone missing for almost three hours prior to the committee meeting.
(How long/How many hours) did Stephen learn whether the representative had talked at the executive committee meeting?
6. Meghan received information that the patient was working out five days a week without losing any weight.
(How long/How many days per week) did Meghan discover whether the patient had exercised without seeing any weight loss?
7. In her graduate research, Lily looked into the unusual phenomenon of a group of Democrats voting Republican for 3 or more consecutive elections.
(How long/For how many consecutive elections) did Lily ascertain whether the Democrats had voted Republican while writing her dissertation?
8. Mark claimed that his students had spent half of the quarter editing haikus instead of discussing modern poetry.
(How long/What part of the academic quarter in total) did Mark state whether his students had been reading through the novels he assigned?
9. Mark claimed that his students had spent half of the quarter editing haikus instead of discussing modern poetry.
(How long/For how many months) did Hilary doubt whether some people would fish during the official fishing season?
10. Erica's article determined that her school's athletes had consumed the right amount of protein only once a week.
(How often/How many days per week) did Erica establish whether the athletes had eaten the recommended amount of protein?
11. Ethan concluded that one official had information about the modified bus routes for at least a few weeks.
(How long/For what length of time) did Ethan assess whether the official had known about the recent controversial firings?
12. Zack informed the others that the guard had not been coming to work punctually on Sunday mornings during the last 6 months.
(How long/For how many months) did Zack report whether the guard had noticed some suspicious activity taking place?
13. Kevin allowed one of his employees to spend two and a half weeks in the Bahamas for her outstanding performance.
(How long/For how many weeks) did Kevin decide whether the employee could go to vacation in the Bahamas?
14. Morgan mentioned that the customer had lingered for five hours over the decision to buy the computer.

(How long/For how many hours) did Morgan indicate whether the customer had got into an argument with her?
15. Dana left her car with the mechanic for a week in July, before she left for a vacation on Cape Cod.
(How long/For how many days in the summer) did Dana speculate whether the mechanic had worked on fixing the broken engine?
16. Leslie noted that it was only for the last three hours before daylight that the lioness prowled without making a sound.
(How long/For how many hours before daylight) did Leslie perceive whether the lioness had moved without making a single sound?
17. Renee did not believe that the reporter would be interested throughout the entire show, since it went on for so long.
(How long/For how much of the event) did Renee confirm whether the reporter would stay to speak with the celebrities?
18. Alex confessed that, for at least the first week of the school year, the tenant was happy with the living arrangement.
(How long/For how many weeks in the school year) did Alex say whether the tenant had felt happy with the living arrangement?
19. Andrew overheard the daycare staff discussing how they wanted to get away from the children for a few hours.
(How long/How many hours) did Andrew hear whether the children had played during the daycare's afternoon recess?
20. Bill couldn't help but see that his officemate was in a meeting with the boss for several hours after the news of the merger.
(How long/How many hours) did Bill notice whether his colleague had argued about the wisdom of merging?
21. Peter found in his old notes that the garden had blossomed for two weeks less than their normal three months because of a drought last year.
(How long/How many weeks less than normal) did Peter note whether the flowers had bloomed during last spring's dry spell?
22. Jennifer's father told her that in 1990 he took five years off to learn some new skills before starting a different career.
(How long/How many years) did Jennifer verify whether her father had traveled before returning to new work?
23. One day, Mark was informed that his favorite painting at his great uncle's house had been there for a hundred years.
(How long/How many years) did Mark determine whether the painting had hung at his great uncle's house?
24. Jane knew that her nephew had been crashing at her parents for a couple of weeks, judging from the mess in the living room.
(How long/How many weeks) did Jane realize whether her nephew had stayed at her parents' now-messy house?

References

Alexopoulou, Theodora, & Frank Keller. 2003. Linguistic complexity, locality, and resumption. *Proceedings of the 22nd West Coast Conference on Formal Linguistics*, ed. by Gina Garding & Mimu Tsujimura, 15–28, Somerville, MA. Cascadilla Press.

Alexopoulou, Theodora, & Frank Keller. 2007. Locality, cyclicity and resumption: At the interface between the grammar and the human sentence processor. *Language* 83.110–160.

Allwood, Jens. 1976. The complex NP constraint as a non-universal rule and some semantic factors influencing the acceptability of Swedish sentences which violate the CNPC. *University of Massachusetts Occasional Papers in Linguistics, Volume II*, ed. by Justine Stillings, 1–20. Amherst, MA: University of Massachusetts at Amherst.

Anderson, John R. 1983. *The Architecture of Cognition*. Harvard, MA: Harvard University Press. Reprinted in 1996 by Lawrence Erlbaum Publishers, Mahwah, New Jersey.

Anderson, John R., & Lynne Reder. 1979. An elaborative processing explanation of depth of processing. *Levels of Processing in Human Memory*, ed. by Laird Cermak & Fergus Craik, 385–404. Hillsdale, NJ: Lawrence Erlbaum Associates.

Andersson, Lars-Gunnar. 1982. What is Swedish an exception to? Extractions and island constraints. *Readings on Unbounded Dependencies in Scandinavian Languages*, ed. by Elisabet Engdahl & Eva Ejerhed, 33–46. Stockholm: Almqvist & Wiksell International.

Ariel, Mira. 1990. *Accessing Noun-Phrase Antecedents*. London: Routledge.

Baayen, R. Harald. 2004. Statistics in psycholinguistics: A critique of some current gold standards. *Mental Lexicon Working Papers* 1.1–45. The Mental Lexicon Project, http://www.mental.lexicon.ling.ualberta.ca/.

Baayen, R. Harald. 2008. *Analyzing Linguistic Data: A Practical Introduction to Statistics*. Cambridge, UK: Cambridge University Press.

Bever, Thomas. 1970. The cognitive basis for linguistic structures. *Cognition and the Development of Language*, ed. by John R. Hayes, 279–362. New York: John Wiley & Sons.

Bever, Thomas. 1975. Functionalist explanations presuppose independently motived theories of behavior. *Papers from the Parasession on Functionalism*, ed. by Robin Grossman, L. James San, & Timothy J. Vance, 580–635, Chicago. Chicago Linguistics Society.

Bever, Thomas. 2009. Remarks on the individual basis for linguistic structures. *Of Minds and Language: A Dialogue with Noam Chomsky in the Basque Country*, ed. by Massimo Piattelli-Palmarini, Juan Uriagereka, & Pello Salaburu, 279–295. Oxford: Oxford University Press.

Bradshaw, Gary, & John R. Anderson. 1982. Elaborative encoding as an explanation of levels of processing. *Journal of Verbal Learning and Verbal Behavior* 21.165–174.

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

Butler, Alban, & Michael J. Walsh. 1991. *Butler's Lives of the Saints*. San Francisco: HarperCollins.

Chomsky, Noam. 1962. The logical basis of linguistic theory. *Preprints of the Papers for the Ninth International Congress of Linguists, August 27–31, 1962*, ed. by Morris Halle, 509–574.

Chomsky, Noam. 1973. Conditions on transformations. *A Festschrift for Morris Halle*, ed. by Stephen Anderson & Paul Kiparsky, 232–86. New York: Holt, Reinhart & Winston.

Chomsky, Noam. 1977. On *wh*-movement. *Formal Syntax*, ed. by Peter Culicover, Thomas Wasow, & Adrian Akmajian. 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. *Step by Step*, ed. by Roger Martin, David Michaels, & Juan Uriagereka, 89–155. Cambridge: MIT Press.

Chung, Sandra. 1994. *Wh*-agreement and "referentiality" in Chamorro. *Linguistic Inquiry* 25.1–44.

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

Clifton, Chuck, & Lyn Frazier. 1989. Comprehending sentences with long-distance dependencies. *Linguistic Structure in Language Processing*, ed. by Michael Tanenhaus & Greg Carlson, 273–317. Dordrecht: Kluwer Academic Publishers.

Comorovski, Ileana, 1989. *Discourse and the Syntax of Multiple Constituent Questions*. Ithaca, NY: Cornell University dissertation.

Culicover, Peter. 1984. Learnability explanations and processing explanations. *Natural Language and Linguistic Theory* 2.77–104.

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

Deane, Paul. 1991. Limits to attention: a cognitive theory of island phenomena. *Cognitive Linguistics* 2.1–63.

Engdahl, Elisabet. 1982. Restrictions on unbounded dependencies in Swedish. *Readings on Unbounded Dependencies in Scandinavian Languages*, ed. by Elisabet Engdahl & Eva Ejerhed, 151–174. Stockholm: Almqvist & Wiksell International.

Ernst, Thomas B. 2002. *The Syntax of Adjuncts*. Cambridge: Cambridge University Press.

Erteschik-Shir, Nomi, 1973. *On the Nature of Island Constraints*. Cambridge, MA: MIT dissertation.

Erteschik-Shir, Nomi. 2007. *Information Structure: The Syntax-Discourse Interface*. Oxford: Oxford University Press.

Erteschik-Shir, Nomi, & Shalom Lappin. 1979. Dominance and the functional explanation of island constraints. *Theoretical Linguistics* 6.43–84.

Fedorenko, Evelina, Edward Gibson, & Doug Rohde. 2007. The nature of working memory in linguistic, arithmetic and spatial integration processes. *Journal of Memory and Language* 56.246–269.

Ferreira, Fernanda, & Chuck Clifton. 1986. The independence of syntactic processing. *Journal of Memory and Language* 25.348–368.

Fiebach, Christian, Matthias Schlesewsky, & Angela D. Friederici. 2001. Syntactic working memory and the establishment of filler-gap dependencies: Insights from ERPs and fMRI. *Journal of Psycholinguistic Research* 30.321–338.

Frazier, Lyn. 1987. Syntactic processing: Evidence from Dutch. *Natural Language and Linguistic Theory* 5.519–559.

Frazier, Lyn, & Charles Clifton. 1989. Successive cyclicity in the grammar and the parser. *Language and Cognitive Processes* 4.93–126.

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, Edward. 1990. A computational theory of processing overload and garden-path effects. *Proceedings of the 13th Conference on Computational Linguistics*, 114–119, Morristown, NJ. Association for Computational Linguistics.

Gibson, Edward. 1998. Linguistic complexity: locality of syntactic dependencies. *Cognition* 68.1–76.

Gibson, Edward. 2000. The dependency locality theory: a distance-based theory of linguistic complexity. *Image, Language, Brain*, ed. by Alec Marantz, Yasushi Miyashita, & Wayne O'Neil, 95–126. Cambridge: MIT Press.

Gibson, Edward. 2006. The interaction of top–down and bottom–up statistics in the resolution of syntactic category ambiguity. *Journal of Memory and Language* 54.363–388.

Gibson, Edward, & Neal Pearlmutter. 1998. Constraints on sentence comprehension. *Trends in Cognitive Science* 2.262–268.

Ginzburg, Jonathan, & Ivan A. Sag. 2000. *Interrogative Investigations*. Stanford, CA: CSLI publications.

Goldberg, Adele. 2006. *Constructions at Work: The Nature of Generalization in Language*. Oxford: Oxford University Press.

Gordon, Peter, Randall Hendrick, & William H. Levine. 2002. Memory-load interference in syntactic processing. *Psychological Science* 13.425–430.

Grodner, Daniel, & Edward Gibson. 2005. Consequences of the serial nature of linguistic input for sentential complexity. *Cognitive Science* 29.261–290.

Groenendijk, Jeroen, & Martin Stokhof. 1997. Questions. *Handbook of Logic and Language*, ed. by Johan van Benthem & Alice ter Meulen, 1055–1124. Cambridge, MA and Amsterdam: MIT Press and North Holland.

Gundel, Jeanette, Nancy Hedberg, & Ron Zacharski. 1993. Cognitive Status and the Form of Referring Expressions in Discourse. *Language* 69.274–307.

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

Hamblin, Charles L. 1973. Questions in Montague English. *Foundations of Language* 10.41–53.

Hawkins, John. 1999. Processing complexity and filler-gap dependencies across grammars. *Language* 75.244–285.

Hirschbühler, Paul, & Daniel Valois. 1992. Argument extraction out of indirect questions in French. *Island Constraints: Theory, Acquisition and Processing*, ed. by Helen Goodluck & Michael Rochemont, 195–222. Dordrecht: Kluwer Academic Publishers.

Hitzeman, Janet. 1997. Semantic partition and the ambiguity of sentences containing temporal adverbials. *Natural Language Semantics* 5.87–100.

Hofmeister, Philip, 2007. *Representational Complexity and Memory Retrieval in Language Comprehension*. Stanford, CA: Stanford University dissertation.

Hofmeister, Philip, T. Florian Jaeger, Ivan A. Sag, Inbal Arnon, & Neal Snider. 2007. Locality and accessibility in *wh*-questions. *Roots: Linguistics in Search of its Evidential Base*, ed. by Sam Featherston & Wolfgang Sternefeld, 185–206. Berlin: Mouton de Gruyter.

Huang, James, 1982. *Logical Relations in Chinese and the Theory of Grammar*. Cambridge, MA: MIT dissertation.

Jurafsky, Dan. 2003. Probabilistic modeling in psycholinguistics: linguistic comprehension and production. *Probablistic Linguistics*, ed. by Rens Bod, Jennifer Hay, & Stefanie Jannedy, 39–95. Boston: MIT Press.

Just, Marcel, & Patricia Carpenter. 1992. A capacity theory of comprehension: Individual differences in working memory. *Psychological Review* 122–149.

Just, Marcel, Patricia A. Carpenter, & J.D. Woolley. 1982. Paradigms and processes in reading comprehension. *Journal of Experimental Psychology: General* 111.228–238.

Karttunen, Lauri. 1977. Syntax and semantics of questions. *Linguistics & Philosophy* 1.3–44.

King, Jonathan, & Marcel Just. 1991. Individual differences in syntactic processing: the role of working memory. *Journal of Memory and Language* 30.580–602.

Kluender, Robert, 1991. *Cognitive Constraints on Variables in Syntax*. La Jolla, CA: University of California, San Diego dissertation.

Kluender, Robert. 1992. Deriving islands constraints from principles of predication. *Island Constraints: Theory, Acquisition and Processing*, ed. by Helen Goodluck & Michael Rochemont, 223–258. Dordrecht: Kluwer.

Kluender, Robert. 1998. On the distinction between strong and weak islands: a processing perspective. *Syntax and Semantics 29: The Limits of Syntax*, ed. by Peter Culicover & Louise McNally, 241–279. San Diego, CA: Academic Press.

Kluender, Robert. 2005. Are subject islands subject to a processing account? *Proceedings of the 23rd West Coast Conference on Formal Linguistics*, ed. by Vineeta Chand, Ann Kelleher, Angelo J. Rodríguez, & Benjamin Schmeiser, 475–499. Somverville, MA: Cascadilla Press.

Kluender, Robert, & Marta Kutas. 1993a. Bridging the gap: evidence from ERPs on the processing of unbounded dependencies. *Journal of Cognitive Neuroscience* 5.196–214.

Kluender, Robert, & Marta Kutas. 1993b. Subjacency as a processing phenomenon. *Language and Cognitive Processes* 8.573–633.

Kothari, Anubha. 2008. Frequency-based expectations and context influence bridge quality. *Proceedings of WECOL 2008*, ed. by Michael Grosvald & Dionne Soares. UC Davis Department of Linguistics. Available online at http://www.stanford.edu/∼anubha/publications.html.

Kroch, Anthony. 1998. Amount quantification, referentiality, and long *wh*-movement. *Penn Working Papers in Linguistics, Vol. 5.2*, ed. by Alexis Dimitriadis, Hikyoung Lee, Christine Moisset, & Alexander Williams. Philadelphia: University of Pennsylvania.

Levy, Roger, 2005. *Probalistic Models of Word Order and Syntactic Discontinuity*. Stanford University dissertation.

Lewis, Richard. 1996. Interference in short-term memory: The magical number two (or three) in sentence processing. *Journal of Psycholinguistic Research* 25.93–115.

Maling, Joan. 1978. An asymmetry with respect to *wh*-islands. *Linguistic Inquiry* 9.75–89.

Maling, Joan, & Annie Zaenen. 1982. A phrase structure account of Scandinavian extraction phenomena. *The Nature of Syntactic Representation*, ed. by Pauline Jacobson & Geoff Pullum, 229–282. Dordrecht: Reidel.

Manzini, M. Rita. 1992. *Locality: A Theory and Some of Its Empirical Consequences*. Cambridge: MIT Press.

McDaniel, Mark, Paul Dunay, Brian Lyman, & Mary Louis Kerwin. 1989. Effects of elaboration and relational distinctiveness on sentence memory. *The American Journal of Psychology* 101.357–369.

McElree, Brian, & Teresa Griffith. 1998. Structural and lexical constraints on filling gaps during sentence comprehension: A time-course analysis. *Journal of Experimental Psychology. Learning, Memory, and Cognition* 24.432–460.

Miller, George A., & Noam Chomsky. 1963. Finitary models of language users. *Handbook of Mathematical Psychology*, ed. by R. Duncan Luce, Robert R. Bush, & Eugene Galanter, volume 2, 419–492. New York: Wiley.

Oshima, David. 2007. On factive islands: pragmatic anomaly vs. pragmatic infelicity. *New Frontiers in Artificial Intelligence: Joint JSAI 2006 Workshop Post-Proceedings*, ed. by Takashi Washio, Ken Satoh, Hideaki Terada, & Akihiro Inokuchi, 147–161. Dordrecht: Springer-Verlag.

Osterhout, Lee, & David Swinney. 1993. On the temporal course of gap-filling during comprehension of verbal passives. *Journal of Psycholinguistic Research* 22.273–286.

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. *The Representation of (In)Definiteness*, ed. by Eric Reuland & Alice ter Meulen, 98–129. Cambridge: MIT Press.

Pesetsky, David. 2000. *Phrasal Movement and its Kin*. Cambridge: MIT Press.

Phillips, Colin. 2007. The real-time status of island phenomena. *Language* 82.795–823.

Pickering, Martin, Stephen Barton, & Richard Shillcock. 1994. Unbounded dependencies, island constraints and processing complexity. *Perspectives on Sentence Processing*, ed. by Chuck Clifton Jr., Lyn Frazier, & Keith Rayner, 199–224. London: Lawrence Erlbaum.

Pickering, Martin, & Matthew Traxler. 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., & Douglas M. Bates. 2000. *Mixed-effects Models in S and S-PLUS*. New York: Springer.

Reder, Lynne. 1980. The role of elaboration in the comprehension and retention of prose: A critical review. *Review of Educational Research* 50.5–53.

Reder, Lynne, Davida Charney, & Kim I. Morgan. 1986. The role of elaborations in learning a skill from an instructional text. *Memory and Cognition* 14.64–78.

Rizzi, Luigi. 1982. Violations of the *wh*-island constraint and the Subjacency Condition. *Issues in Italian Syntax* 49–76.

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

Rizzi, Luigi. 2001. Reconstruction, weak island sensitivity, and agreement. *Semantic Interfaces: Reference, Anaphora and Aspect*, ed. by Carlo Cecchetto, Gennaro Chierchia, & Maria Teresa Guasti, 145–176. Stanford: CSLI Publications.

Ross, John R., 1967. *Constraints on Variables in Syntax*. Cambridge, MA: MIT dissertation. Published in 1986 as *Infinite Syntax!* by Ablex, Norwood, N. J.

Rudin, Catherine. 1988. On multiple questions and multiple *wh*-fronting. *Natural Language & Linguistic Theory* 6.445–501.

Saah, Kofi, & Helen Goodluck. 1995. Island effects in parsing and grammar: evidence from Akan. *Linguistic Review* 12.381–409.

Sabel, Joachim. 2002a. A minimalist analysis of syntactic islands. *The Linguistic Review* 19.271–315.

Sabel, Joachim. 2002b. *Wh*-questions and extraction asymmetries in Malagasy. *MIT Working Papers in Linguistics*, ed. by Andrea Rackowski & Norvin Richards, volume 44, 309–323.

Sag, Ivan A. to appear. English filler-gap constructions. *Language* .

Sportiche, Dominique. 1981. Bounding nodes in French. *Linguistic Review* 1.219–246.

Sprouse, Jon, 2007. *A Program for Experimental Syntax*. College Park: University of Maryland, College Park dissertation.

Sprouse, Jon. 2009. Revisiting satiation: evidence for an equalization response strategy. *Linguistic Inquiry* 40.329–341.

Stowe, Laurie A. 1986. Parsing *wh*-constructions: Evidence for on-line gap location. *Language and Cognitive Processes* 1.227–245.

Szabolcsi, Anna. 2006. Strong vs. weak islands. *The Blackwell Companion to Syntax*, ed. by Martin Everaert & Henk van Riemsdijk, volume 4, 479–531. Malden, MA: Blackwell Publishing.

Szabolcsi, Anna, & Frank Zwarts. 1993. Weak islands and an algebraic semantics of scope taking. *Natural Language Semantics* 1.235–284.

Taraldsen, Knut T. 1982. Extraction from relative clauses in Norwegian. *Readings on Unbounded Dependencies in Scandinavian Languages*, ed. by Elisabet Engdahl & Eva Ejerhed, 205–221. Stockholm: Almqvist & Wiksell International.

Traxler, Matthew, & Martin Pickering. 1996. Plausibility and the processing of unbounded dependencies: An eye-tracking study. *Journal of Memory and Language* 35.454–475.

Truswell, Robert, 2007. *Locality of* Wh*-movement and the Individuation of Events*. London: University College London dissertation.

Van Dyke, Julie. 2007. Interference effects from grammatically unavailable constituents during sentence processing. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 33.407–430.

Vasishth, Shravan, & Richard Lewis. 2006. Argument-head distance and processing complexity: Explaining both locality and anti-locality effects. *Language* 82.767–794.

Wagers, Matt, & Colin Phillips. 2009. Multiple dependencies and the role of the grammar in real-time comprehension. *Journal of Linguistics* 45.395–433.

Wanner, Eric, & Michael Maratsos. 1978. An ATN approach to comprehension. *Linguistic Theory and Psychological Reality*, ed. by Morris Halle, Joan Bresnan, & George A. Miller, 119–161. Cambridge, MA: MIT Press.

Warren, Tessa, & Edward Gibson. 2002. The influence of referential processing on sentence complexity. *Cognition* 85.79–112.

Warren, Tessa, & Edward Gibson. 2005. Effects of NP type in reading cleft sentences in English. *Language and Cognitive Processes* 20.751–767.

Wiseman, Sandor, Colin M. MacLeod, & P.J. Lootsteen. 1985. Picture recognition improves with subsequent verbal information. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 11.588–95.

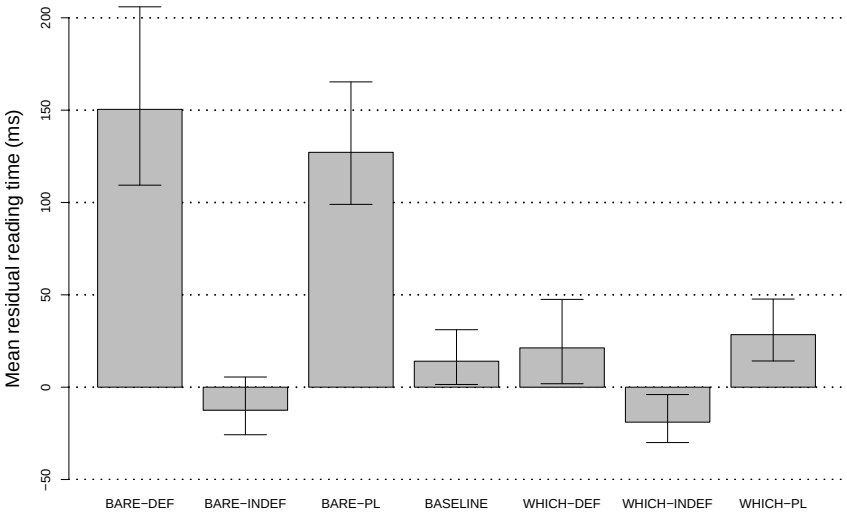

Figure 1: Mean residual reading times at complementizer (*that*) in Experiment 1. Error bars show (+/−) one standard error.

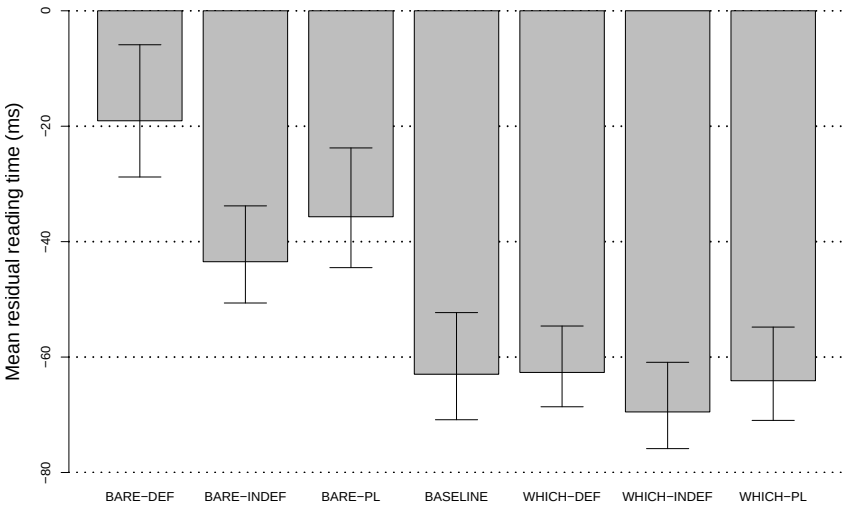

Figure 2: Mean residual reading times at embedded auxiliary and verb in Experiment 1. Error bars show (+/−) one standard error.

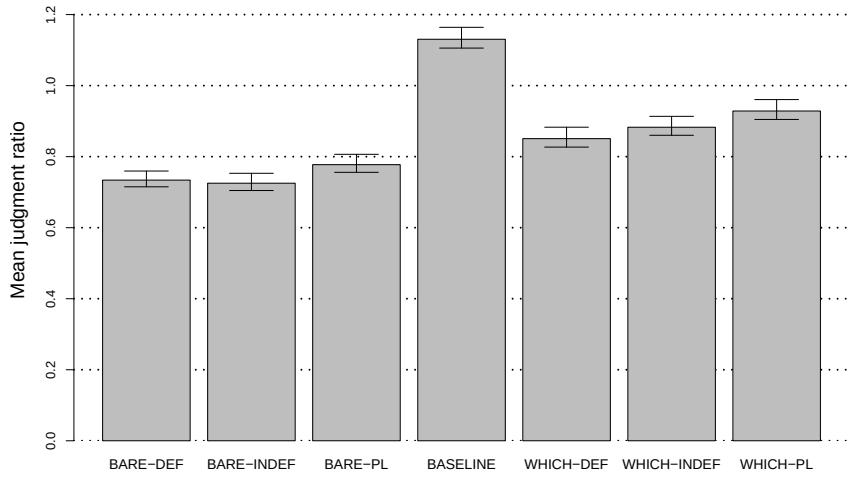

Figure 3: Mean judgment ratios of CNPC violations and baseline. Error bars show (+/−) one standard error.

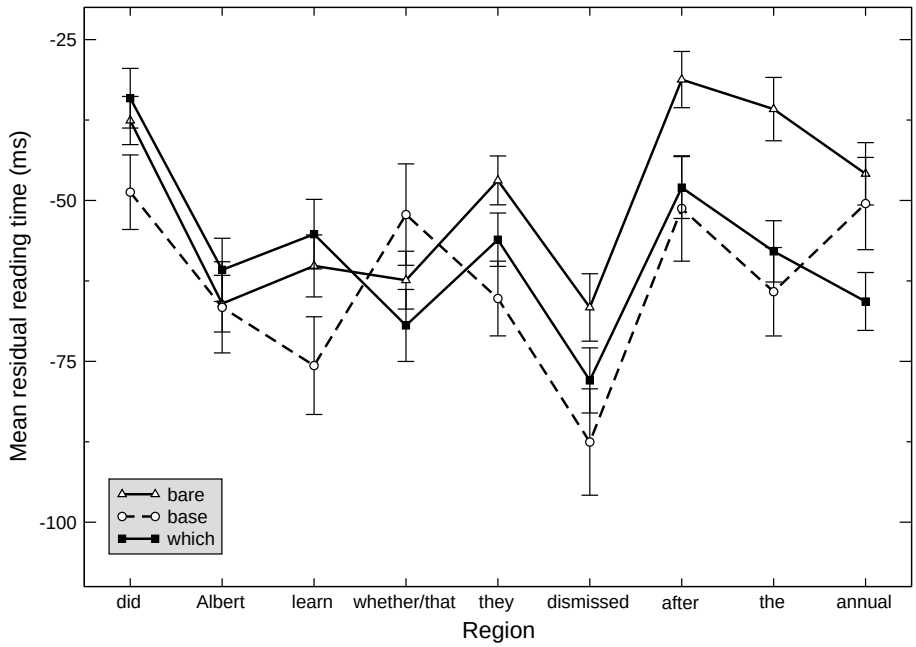

Figure 4: Mean residual reading times in Experiment 2, ranging from first word after *wh*-phrase to three words after the subcategorizing verb. Error bars show (+/−) one standard error.

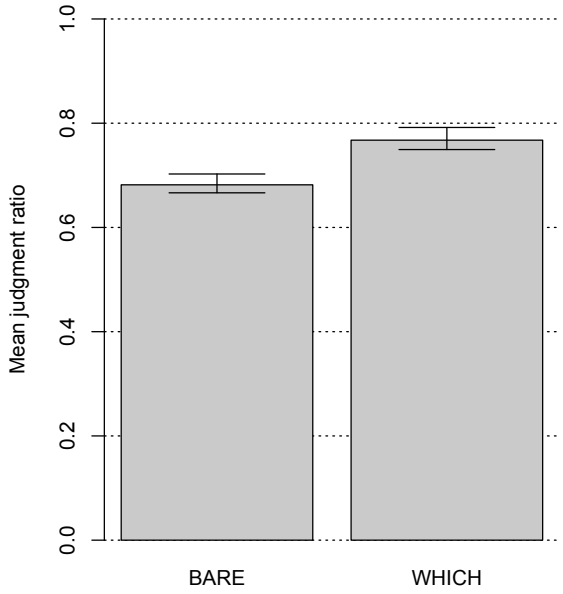

Figure 5: Mean judgment ratios of embedded *Wh*-island violations: BARE = bare *wh*-phrase; WHICH = *which*-$\overline{\text{N}}$ phrase. Error bars show (+/−) one standard error.

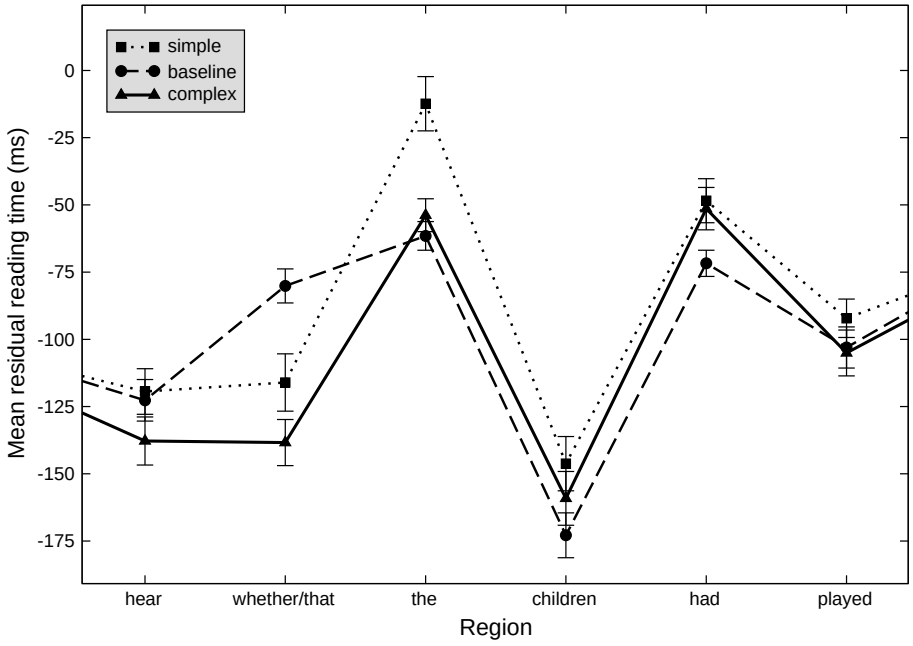

Figure 6: Mean residual reading times from matrix verb to embedded verb in Experiment 3. Error bars show (+/−) one standard error.

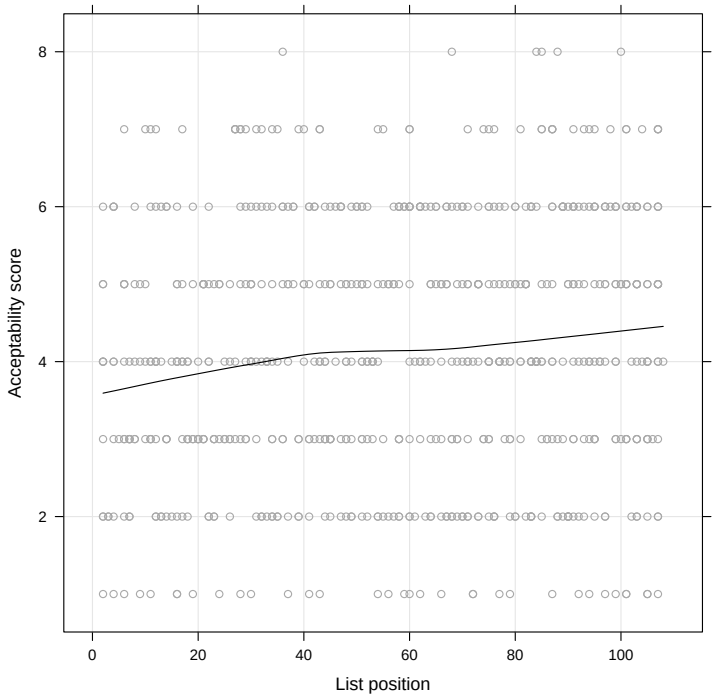

Figure 7: Acceptability judgments by list position of CNPC violations for all participants. Scatterplots fitted with a locally weighted scatterplot smoother (Lowess curve).

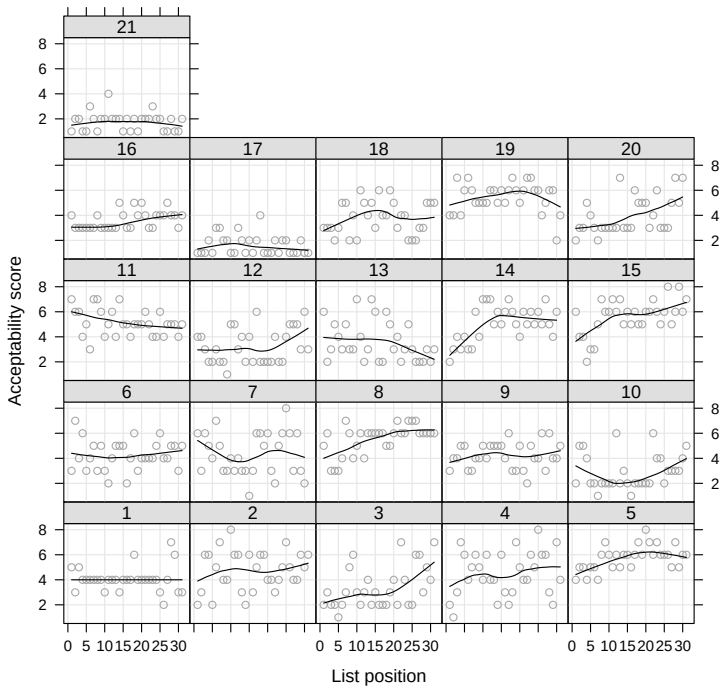

Figure 8: Acceptability judgments by list position of CNPC violations for each participant. Each tile shows the judgments for one subject, according to the list position of each CNPC violation relative to other CNPC violations. Scatterplots fitted with a locally weighted scatterplot smoother (Lowess curve).

		$Region_1$	$Region_2$	$Region_3$	$Region_4$	$Region_5$	$Region_6$
Word		that	they	had	captured	in	the
Residual RT	BARE	87.93 (20.69)	48.88 (14.73)	−23.65 (8.17)	−41.66 (8.40)	−33.68 (7.22)	−36.91 (6.98)
	WHICH	11.27 (10.54)	−6.34 (9.76)	−59.09 (5.23)	−71.53 (6.96)	−48.44 (5.41)	−64.01 (4.90)
	BASELINE	14.07 (14.83)	−29.10 (11.09)	−57.22 (11.84)	−68.72 (14.34)	−22.05 (16.22)	−52.27 (12.41)
Raw RT	BARE	467.51 (21.75)	420.36 (15.48)	352.42 (8.06)	361.41 (9.02)	349.05 (8.39)	338.74 (7.29)
	WHICH	384.02 (10.50)	360.34 (10.26)	318.04 (5.67)	328.64 (6.95)	329.95 (6.39)	309.79 (5.11)
	BASELINE	395.94 (17.67)	343.92 (11.36)	324.54 (11.97)	330.66 (12.80)	358.22 (17.67)	329.35 (13.33)
Complexity	F_1, $df = (1,24)$	14.813***	5.415*	15.083***	8.044**	3.597 (.)	6.575*
	F_2, $df = (1,35)$	11.532**	10.386**	11.572**	6.858*	1.913	13.580***
NP type	F_1, $df = (2,48)$	6.034**	< 1	< 1	5.173*	< 1	< 1
	F_2, $df = (2,70)$	7.506**	< 1	< 1	3.746*	< 1	< 1
Complexity × *NP type*	F_1, $df = (2,48)$	3.073 (.)	1.093	< 1	3.171 (.)	< 1	< 1
	F_2, $df = (2,70)$	2.420	1.902	< 1	2.555 (.)	< 1	< 1

Table 1: Mean residual and raw reading times for word regions inside embedded clause clause in experiment I (standard errors in parentheses) + analysis of variance results.

		$Region_1$	$Region_2$	$Region_3$	$Region_4$	$Region_5$
Word		they	dismissed	after	the	annual
Residual RT	BARE	−46.88 (3.80)	−66.62 (5.24)	−31.21 (4.37)	−35.80 (4.92)	−45.86 (5.28)
	WHICH	−56.10 (4.15)	−77.96 (5.06)	−47.99 (4.81)	−57.91 (4.77)	−65.70 (4.50)
	BASELINE	−65.24 (5.81)	−87.54 (8.27)	−51.24 (8.19)	−64.18 (6.87)	−50.49 (7.17)
Raw RT	BARE	252.45 (5.10)	255.96 (5.68)	265.27 (5.41)	261.48 (5.75)	261.31 (4.94)
	WHICH	245.03 (4.84)	243.74 (4.67)	250.52 (5.26)	244.44 (4.97)	239.44 (4.35)
	BASELINE	238.95 (6.23)	231.79 (6.78)	247.54 (7.94)	239.17 (6.17)	255.40 (8.16)
	$t_1(19)$	1.299	1.283	2.346*	3.180**	4.101***
	$t_2(23)$	1.528	1.441	2.459*	3.053**	3.190**

Table 2: Mean residual and raw reading times for word regions inside embedded clause in experiment II (standard errors in parentheses) + paired t-test results (BARE vs. WHICH) by region