Tel-Aviv University

Lester and Sally Entin Faculty of the Humanities

Department of Linguistics

Processing Complex NP Islands in Hebrew

MA thesis submitted by

Bruno Nicenboim

Prepared under the guidance of

Prof. Julia Horvath

Prof. Tali Siloni

October 2010

Contents

List of Figures

List of Tables

Abstract

Since Ross's (1967) work on island constraints until many present day works, the treatment of island phenomena is mostly based on universal constraints on the grammar. However, it has been claimed that the unacceptability of certain island violations can be explained by extragrammatical factors, such as processing difficulties. The advantage of this approach is that the unacceptability is explained by factors that exist independently of the island phenomena.

The aim of this work is to investigate whether processing factors known to affect filler-gap dependencies (or wh-movement) can account for the unacceptability of Complex NP island violations in Hebrew, including both extractions from complement clauses of NPs (CNPCC) and relative clauses of NPs (CNPRC).

The effect of D-linking, a manipulation over the wh-element that is independently associated with the processing of filler-gap dependencies, is examined in a series of experiments. Two acceptability-rating experiments examine the effect of D-linking on CNPCC and CNPRC violations, and a self-paced reading experiment complements the results of the acceptability study on CNPCC.

Experiment 1 shows that the D-linking of the wh-phrase improves the acceptability judgments of questions with CNPCC violations in a statistically significant manner, although this manipulation does not make the sentences fully acceptable. I suggest that questions with CNPCC violations and D-linked wh-phrases receive better judgments because the D-linked wh-phrases are more easily integrated into the sentence representation. However, these findings cannot base the view that the low acceptability of *CNPCC violations as a whole* is due to processing limitations.

Experiment 2 shows no significant effect for D-linking in CNPRC islands, suggesting that CNPRC violations are not susceptible to processing factors.

Experiment 3 reveals that D-linking facilitates the processing of the filler-gap dependencies in sentences with CNPCC violations. However, processing differences within CNPCC violations and a grammatical baseline are not matched by contrasts in acceptability judgments, which is not compatible with the view that limitations on the cognitive resources related to the filler-gap dependencies process are responsible for CNPCC violation effects.

The experimental results strongly imply that CNPRC and CNPCC are two very distinct phenomena and I argue that the unacceptability of CNPRC violations is due to their ungrammaticality.

Regarding CNPCC violations, part of their unacceptability is due to the processing difficulties involved in parsing a sentence with a filler-gap dependency. These processing difficulties seem not to be the reason for the unacceptability of the violation as a whole, and they seem to explain the same kind of acceptability variation that is found between regular yes-no questions and wh-questions.

However, for CNPCCs, there is a major slowdown in processing at the head of the island. This suggests either that the parser "discovers" the grammatical violation at this region, which provokes both the processing slowdown and the unacceptability; or that the unacceptability is a consequence of the processing efforts of building the derivation of the island, that is, sentences with CNPCC violations are generated by the grammar but are hard to process.

1 Introduction

Discussing the general background assumptions and goals that underlie the work in generative grammar, Chomsky (1978) makes a distinction between what the speaker of a language knows implicitly, his competence, and what he does, his performance. Even though a grammar is an account of competence, it is performance that provides the evidence for its investigation. Thus, Chomsky (1978) claims that "competence must be distinguished from performance if either is to be seriously studied". It is well established that there are sentences which conform to the rules of the grammar, but are judged to be unacceptable because the actual performance of the speaker cannot "deal" with them. Chomsky and Miller (1963) agree that in certain constructions like triple center-embedded sentences (1) memory limitations can explain why certain linguistic data are unacceptable.

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

Regarding Island effects, since Chomsky's assumption of the 'A-over-A Condition' and Ross's (1967) work on island constraints until many present day proposals, much of the treatment of this topic is based on universal constraints on the grammar. However, there is an ongoing research that attempts to explain island phenomena in terms of processing factors, that is performance limitations (Kluender, 1992; Kluender and Kutas, 1993a,b; Kluender, 1998; Hawkins, 1999; Goodluck, 1997; Hofmeister and Sag, 2010). Furthermore, Hofmeister and Sag (2010) show experimentally that the manipulation over factors that ease the processing of filler gap dependencies also improves the acceptability judgment and reading times of sentences with Complex NP constraint violations. Sentence (2 a), which has a wh-element referring to a limited set of referents and semantically richer, and a less specified head of the noun phrase, is perceived as more acceptable than (2 b).

(2) (a) I saw which convict Emma doubted [a report that we had captured _ in the nationwide FBI manhunt].

(b) I saw who Emma doubted [the report that we had captured _ in the nationwide FBI manhunt].

Even though Hofmeister and Sag only examined extractions from complement clauses of NPs, Kluender (1992, 2004) also states that similar improvements occur in extractions from relative clauses of NPs as head nouns become less specified, since they are "less costly in terms of their discourse processing requirements".

Whereas manipulations on the head nouns governing sentences with a gap (indicated by _) are not relevant to sentences with filler-gap dependencies as a whole, manipulations on the wh-phrase have been shown to affect judgments and reading times of sentences with filler-gap dependencies both with and without island violations.

The aim of my thesis is to investigate whether processing factors known to affect filler-gap dependencies (or wh-movement) can account for certain island effects in Hebrew, in particular, for Complex NP island violations, including both extraction from relative clauses (CNPRC) and complement clauses (CNPCC) of NPs. The methodology of this work consists in eliciting acceptability judgments and reading times of questions with Complex NP island violations while their wh-phrase is manipulated.

The experimental results suggest that processing factors that affect filler-gap dependencies do have an effect on sentences with CNPCC violations, but they are not the source of their low acceptability. Questions with CNPCC violations do behave similarly to the sentences of Hofmeister and Sag's study showing a significant improvement in acceptability judgments and faster reading times after the manipulation. However, processing differences within islands and non-islands are not matched by contrasts in acceptability judgments. The results show that, under certain conditions, filler-gap dependencies can be processed faster in structures with CNPCC violations than in grammatical control sentences.

Regarding CNPRC violations, the experimental results show that they are immune to factors that usually affect processing; this suggests that their low acceptability is due to their ungrammaticality.

The following section is intended to draw a line between acceptability and grammaticality, then in section 3 previous investigations about the relation between processing and acceptability are examined. Section 4 gives a brief review on Complex NP constraint, the previous investigation that related it to processing and in particular the effect of the manipulation on the referentiality or informativity of wh-phrases. Section 5 describes and discusses the results of three experiments: an acceptability judgment task on questions in Hebrew with CNPCC violations (5.2) and with CNPRC violations (5.3), and a self-paced reading task on the same stimuli as the first experiment (5.4). A general discussion in section 5.5 follows the experiments, and section 6 summarizes the conclusions of the present work. The results of the statistic analysis of the three experiments are condensed in several tables in Appendix A, and the stimuli used are presented in Appendix B and C.

2 Acceptability and Grammaticality

In order to investigate processing factors and their effect on acceptability judgments, the meaning of speaker's judgments and their implications to the linguistic theory must be examined. It is generally assumed an ideal speaker-listener who is unaffected by extra-grammatical conditions (Chomsky, 1965); in this case, his or her judgment about the acceptability of a sentence would coincide with its grammaticality. However, in most cases, grammaticality and acceptability must be distinguished.

Following Chomsky (1965), I will use the term acceptable to refer to utterances that are perfectly natural and immediately comprehensible. Whereas acceptability is a concept which belongs to the study of performance, grammaticality belongs to the study of competence (Chomsky, 1965). A sentence is grammatical in some language when it does not violate the rules and principles of the grammar of that language or the rules of the universal grammar. Grammaticality is, according to Chomsky (1965), only one of the many factors that interact to determine acceptability. A sentence may be unacceptable because it is ungrammatical but not necessarily, since a grammatical sentence may be unacceptable due to memory limitations, intonational and stylistic factors, pragmatic factors, etc.

Since speakers do not have direct access to their competence, speakers' judgments are presumably an attempt to assess what their reaction to a sentence would be across a range of situations (Schütze, 1996), and then they can only be interpreted as *acceptability* judgments (Cowart, 1997). Nonetheless, it is implicitly assumed that linguists are capable of giving pure grammaticality judgments, and while sometimes it may be the case, this can also be problematic (see Featherston, 2007; Grewendorf, 2007 and Schütze, 1996). Dąbrowska (2010) shows that while linguist judgments differ from naïve judgments, also generative and functionalist linguists may differ significantly in their judgments (which she attributes to either theoretical commitments or differences in exposure). Furthermore, Fanselow and Frisch's (2006) German study shows that both linguists and non-linguists' acceptability judgments are influenced by processing factors.

Acceptability is uncontroversially a gradient concept varying from completely unacceptable to perfectly acceptable. Grammaticality, on the other hand, can be assumed a gradient concept or a dichotomous one.

The view that also grammaticality occurs on a continuum is based on the fact that speakers actually give gradient judgments (Lakoff, 1973 as cited in Schütze, 1996). Such view is represented in, for example, Linear Optimality Theory (LOT). In LOT, the gradience of the data is explained by soft and hard constraints which are ranked and are cumulative (Sorace and Keller, 2005).

On the other hand, grammars that are based on a dichotomous concept of grammaticality explain the gradience of the judgments by the interaction of extragrammatical factors (pragmatic, semantic, memory constraints, etc) (Bever and Carroll, 1981). This view assumes a grammar with rules that either generate a sentence or they do not. Under this approach, while the *grammatical sentences* of a language are a well-defined set, the *actual sentences of a language* belong to a fuzzy subset of the grammatical sentences, which range from fully acceptable to somewhat acceptable.

A full spectrum of grammars are possible between the gradient-free and the one that accepts gradient grammaticality, and they depend on theory internal assumptions, compliance with Occam's Razor (since a discrete system is simpler than continuous one) and where the line between extragrammatical and grammatical factors is drawn.

The possibility of a grammar with some levels of grammaticality but with a clear distinction between grammatical and ungrammatical is represented by Chomsky's (1965) *Aspects of the Theory of Syntax*. Although Chomsky assumes that there are degrees of *un*grammaticality, Chomsky's theory assumes the existence of absolute grammaticality. Sentences that do not violate any constraint are uniformly grammatical whereas if a sentence violate some constraint it will be ungrammatical; the degree of ungrammaticality of a given sentence will depend on the "importance" of the constraint.

Under any view, if a sentence has a low acceptability, there is no way to know *a priori* whether this is because it is ungrammatical, because extragrammatical factors or both. Nonetheless, if by easing the extragrammatical factors, a previously unacceptable sentence becomes acceptable, it can be safely assumed that the reason of the low acceptability is the extragrammatical factor that was eased.

3 Processing and Acceptability

Even though many of the phenomena investigated in the generative linguistic literature are competence-related, many aspects of filler-gap dependencies clearly belong to the realm of processing and have being investigated since Fodor's (1978) work on parsing strategies. Since island effects are a special case of filler-gap dependencies, processing factors that affect the latter should be understood in order to investigate Complex NP island effects properly.

In regular questions (in languages that do not permit wh-elements in situ), a wh-phrase is displaced to the left periphery of a clause (**What** did you say _ to John). Since potentially the wh-phrase can be related to any part of the sentence, the only basis for determining the role of the question word, i.e. the "filler", is the existence of a "gap" (indicated in the example by a blank line). Filler and gap are mutually dependent on each other since they share syntactic and semantic information needed for the comprehension of the sentence and the filler must be held in working memory until filler-gap assignment can take place. Since it is assumed that there may be empty resumptive pronouns (Cinque, 1990) in addition to the overt ones that can fill the place of a "missing argument", I will use gap as a term that covers both traces and empty pronouns.

Although there are many different accounts, it is well established from the processing literature that a sentence with a filler-gap dependency incurs a relatively higher degree of processing difficulty than a minimally different sentence without the dependency[1]. Chen, Gibson, and Wolf's (2005) comparison of reading times of complement clauses (3 a), which lack of filler-gap dependencies, with relative clauses (3 b), which have a filler-gap dependency, shows that people read the critical region in which a filler is pending slower than if no filler is pending.

(3) (a) The claim alleging [that the cop who the mobster attacked ignored the informant] might have affected the jury. (9.a in Chen et al., 2005)

(b) The claim [which the cop who the mobster attacked ignored _] might have affected the jury. (9.b in Chen et al., 2005)

[1] Boston (2010) identifies 7 possible computational models that take into account the memory constraints of the parser in filler-gap dependency processing.

Kluender and Kutas's (1993a) psycholinguistic research compared yes/no question with wh-questions and also shows that the processing costs of holding a filler in working memory and associating it with a gap correlate with the appearance of an Event-Related Potential (ERP) component known as left anterior negativity (LAN), which is associated with working memory.

Furthermore, it has also been shown that a higher degree of processing difficulty is usually associated with a lower acceptability. Chomsky and Miller (1963) claim that triple center embedded sentences are (many times) completely unacceptable due to memory constraints while they are generated by grammatical rules of recursion. In fact their acceptability can be improved dramatically by, for example, decreasing the specificity of their embedded subjects (Warren and Gibson, 2002) (Compare 1 repeated here as 4 a with 4 b).

(4) (a) The girl the boy the host knew brought left.

(b) The girl someone I knew brought left.

The acceptability of "regular" grammatical sentences has also been shown to be sensitive to processing factors. Firstly, Hofmeister et al. (2007) show in indisputable grammatical and acceptable questions with filler-gap dependencies that the greater the distance between the filler and its gap, the less acceptable the sentence. Although distance can be measured in different ways and may be modulated by several factors, it has been shown that increasing the distance between the filler and its gap increases the processing complexity of a sentence (by, among others, Gibson, 2000; Hawkins, 1999, however as Vasishth and Lewis, 2006 notice there are exceptions). Secondly, Kluender (1998) shows that the differences in acceptability across grammatical question types (especially the lower acceptability of wh-questions compared to yes-no questions) coincide with processing complexity and changes of the ERP components (LAN and N400), which index working memory and processing costs.

Thirdly, it has been shown that the parser prefers to complete long distance dependencies as quickly as possible, which has come to be known as the active filling strategy (Frazier and Flores, 1989). Since the quickest possible completion site is not always the correct one, the active filling strategy entails the construction of temporary incorrect representations, which have been shown to lower the acceptability of a sentence (Sprouse, 2008). The completion of the dependency in the quickest way in sentence (5) entails the construction of an incorrect representation indicated by ***_**.

(5) My brother wanted to know **who** Ruth will bring ***_** *us* home to **_** at Christmas. (1.b in Sprouse, 2008)

While the active filling strategy shows that *grammatical* structures may appear less acceptable when their processing temporarily involves a stage in which a constraint, such as θ-Criterion, seems violated (Sprouse, 2008), Fanselow and Frisch (2006) show that *ungrammatical* sentences may appear more acceptable if their parsing involves a temporary construction in which a violated constraint seems fulfilled.

Regarding the relationship between grammatical violations and processing, the situation is more complex. While psycholinguistic research shows that greater processing difficulty results in lower acceptability judgments, ungrammaticality does not necessarily slow processing. Vasishth et al. (2010) look at sentences that lack of a middle verb in a double center embedding in English and in German. Whereas in both languages this entails a grammatical violation, the ungrammatical sentences are processed faster than the grammatical baselines in English while they are processed slower in German.

4 Complex Noun Phrase Constraint in Hebrew

4.1 Review of Complex Noun Phrase Constraint

Islands, in general, and Complex NP, in particular, were first defined by Ross (1967) as a universal grammatical constraint. Ross's first work on Island constraints investigated the fact that seemingly small manipulations in a transformation dramatically affected its acceptability (6 a)(6 b). These Island configurations include domains like complex noun phrases, adjoined clauses, coordinate structures, left branches, sentential subjects, and embedded interrogative clauses. (For an overview of Islands' treatment see Boeckx, 2008; Szabolcsi, 2006)

(6) (a) What did you hear that John's dog did _?

(b) * What did you hear the rumor that John's dog did _?

The Complex NP Constraint as stated by Ross claims that "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". This constraint prevents both the extraction from a sentential complement of a noun as in (6 b) and the extraction from the relative clause of a noun.

Although sentences with Complex NP Constraint violations are unacceptable in many languages including Hebrew, there are two main criticisms to its status as a universal grammatical constraint (which in part can be extended to other islands, but it is beyond the scope of the paper): that the constraint is not crosslinguistically valid and that the unacceptability of these violations can be explained by extragrammatical factors.

Some languages, including Scandinavian languages (Engdahl, 1997; Allwood, 1982, and others), were shown to allow violations of the CNP constraint (both with complement clauses (CNPCC) and relative clauses (CNPRC)) in regular speech.

(7)	*De*	*blommorna*	*känner*	*jag*	*en*	*man*	*som*	*säljer*	
	those	flowers	know	I	a	man	who	sells	(30 in Allwood, 1982)

Since Grosu (1982) and Kuno's (1976) works, which addressed the semantic and pragmatic aspects of Island constraints, many investigations, including processing accounts, have attributed extragrammatical explanations to Island effects. The advantage of most extragrammatical explanations is that they do not assume ad hoc constraints in order to rule out the island violations. Under these different frameworks the unacceptability of island violations is explained using extragrammatical factors which are known to exist independently of the particular island violation and affect information structure, the use of topic/comment properties, processing of filler-gap dependencies, etc.

However, from a processing perspective, the differences between languages are not less problematic. Since a reasonable assumption is that all humans share the same cognitive capacities, it is implausible that speakers of Swedish have more memory than, say, English speakers. Regarding semantic and pragmatic approaches, the question whether languages where island violations are acceptable have different semantic or pragmatic machinery than languages that do not allow islands is left open.

Hawkins (1999) ties the crosslinguistic differences to conventions of grammars, since he argues that grammars define sets of permissible structures for language use and processing. His claim is that gaps in sentences dominated by an NP are not allowed in many languages because they are difficult to process and in those languages that permit them, only gaps in the simplest Complex NPs (in processing terms) are acceptable. Since he considers harder to process structures as more marked and less common crosslinguistically, he predicts that most languages will not grammaticize structures that allow Complex NP violations.

Regarding the Complex NP constraint in Hebrew, it is uncontroversially judged as not acceptable. However, speakers report that Complex NP with relative clause constraint (CNPRC) violations (8 b) are generally worse than Complex NP with complement clause constraint (CNPCC) violations (8 a). Moreover, a preliminary survey that I conducted corroborates that the difference between the acceptability judgments is significant (p = 0.0002 (2-tails), t(134) = 3.78).

(8) (a) * *Ma Itay šama et ha-šmua še-Roy maxar _ ?*
what Itay heard ACC the rumor that Roy sold _?
'What did Itay hear the rumor that Roy sold?' (CNPCC)

(b) * *Ma Itamar pagaš et ha-iš še-maxar _ ?*
what Itamar met ACC the man that sold _?
'What did Itamar meet the man that sold?' (CNPRC)

4.2 Complex NP constraint and Processing Constraints

The reason for tying the Complex NP constraint to extragrammatical factors and in particular to processing constraints is that some of the island constructions discussed in the linguistic literature have characteristics that are shared with other hard-to-process filler-gap dependencies. Kluender and Kutas's (1993a) claim is that many islands seem to arise at "processing bottlenecks" when the processing demands of a filler-gap dependency add up on the critical processing demands of crossing a clause boundary. Thus, at least part of their unacceptability may be explained as the difficulty or failure of the parser in repositing the extracted element within its clause. However, there is some experimental evidence pointing out that processing factors related to filler-gap dependencies are relevant to the extractions from wh-islands (Kluender, 1998; Hofmeister and Sag, 2010 and Boston (2010) from a computational perspective), subjects (Kluender, 2004) and temporal adjunct islands (Goodluck, 1997).

Regarding Complex NP constructions, Kluender (1998) and Hofmeister and Sag (2010) claim that maintaining a filler-gap dependency across a clause inside an NP requires near or already more than the maximum power used by the parser in acceptable sentences. The relatively long filler-gap dependencies and, in many cases, an initial misparsing of the sentence because of the active filling strategy (see section 3) may aggravate the parser's difficulties to cope with these constructions.

The picture that emerges from these studies is that a manipulation that affects the processing difficulties linked to filler-gap dependencies should also affect the reading times (RTs) and the acceptability of these sentences. Since both measures are linked (processing difficulties are characterized by higher RTs and lower acceptability), variation in acceptability should match the variation on RTs. The next section discusses D-linking as a manipulation on the wh-phrase that affects both RTs and acceptability. However, in section 5, experiments 1 and 3 show that such correspondence between acceptability and RTs is not found; this suggests that processing difficulties associated with filler-gap dependencies are not the source of the low acceptability of CNPCC violations.

4.2.1 D-linking

Wh-phrases can be manipulated and distinguished in terms of their referential or informativity properties. The so-called D-linked which-N' phrases are interpreted as referring to a limited set of referents that has already been established (Cinque, 1990; Pesetsky, 1987) and as being more informative or semantically richer than non-D-linked wh-phrases (Hofmeister, 2007).

(9) (a) What have you seen? (without D-linking)

(b) Which movie have you seen? (with D-linking)

D-linked fillers form easier to process dependencies than interrogative pronouns; while such fillers show longer reading times than interrogative pronouns when encountered (Hofmeister, 2007; De Vincenzi, 1996), they are shown to facilitate processing at retrieval points (Hofmeister, 2007) and to speed global RTs in sentences with embedded questions (Frazier and Clifton, 2002). Furthermore, it has been shown that D-linked wh-phrases increase acceptability and reduce reading time at the retrieval site in sentences with wh-islands (Hofmeister and Sag, 2010), sentences with superiority condition violations and even in non-superiority and non-island structures (Hofmeister et al., 2007; Sprouse, 2007).

(10) (a) I wish I knew who read what. (non-D-linked, 21.i in Sprouse, 2007)

(b) I wish I knew which student read which book. (D-linked, 22.i Sprouse, 2007)

Very similar results are obtained when CNPCC violations with D-linked fillers (11 a) and with non-D-linked ones (11 b) are compared (Hofmeister and Sag, 2010); D-linked fillers get better judgments and shorter RTs inside the embedded clause than non-D-linked ones. Moreover, RTs for D-linked CNPCC conditions are comparable to that of the baseline condition (11 c) despite the fact that the acceptability judgments for D-linked CNPCCs are still significantly lower than that of the baseline.

(11) (a) I saw **which convict** Emma doubted [the report that we had captured _ in the nationwide FBI manhunt]. (CNPCC and D-linking)

(b) I saw **who** Emma doubted [the report that we had captured _ in the nationwide FBI manhunt]. (CNPCC and no D-linking)

(c) I saw which convict Emma doubted [that we had captured _ in the nationwide FBI manhunt]. (Baseline)

4.2.2 Specificity of head nouns

While only extractions from complement clauses of NPs were examined experimentally, Kluender (1992, 2004) claims that similar improvements occur in extractions from relative clauses of NPs as head nouns become less specified. Although his findings are not experimental he notes the following contrast (from more to less acceptable):

(12) (a) ?This is the paper that we really need to find [**someone** who understands _].

(b) ??This is the paper that we really need to find [**a linguist** who understands _].

(c) ???This is the paper that we really need to find [**the linguist** who understands _].

Despite the fact that it is claimed that as head nouns become less specified the acceptability judgments of CNP violations in general improves (Kluender, 1992; Hawkins, 1999, among others), Hofmeister and Sag (2010) show no significant improvement in judgments and a weak improvement in reading times for this manipulation in sentences with CNPCC.

4.2.3 Implications of the results

The results of Hofmeister and Sag's study do not straightforwardly show that processing difficulty plays a causal role or that ungrammaticality can be dismissed for CNPCC violations. There are two problematic issues that are partially addressed: D-linking does not fully repair the unacceptability of CNPCCs, and there is no perfect correlation between reading times and acceptability ratings.

That the D-linked sentences with CNPCC are still less acceptable than "regular" sentences, can be explained by the fact that the hard-to-process structure is not removed from the sentences: the filler depends on a gap that is still in a complement clause dominated by a noun. However, even if D-linking have been independently observed to cause grammatical sentences to be judged more acceptable, the possibility that the unacceptability of CNPCCs is due to *both* a grammatical violation (or, in fact, any other extragrammatical factor not related to processing) and processing difficulties cannot be dismissed. Even though the ungrammaticality cannot be dismissed purely on experimental grounds, processing difficulties which are operationalized as reading times aligning themselves with acceptability judgments provide evidence in favor of a processing account. Such correlation is not fully found in Hofmeister and Sag's study. While D-linking does not improve the acceptability of the sentences with CNPCC violations to the level of the acceptable baseline sentences, it does speed reading times to their level. The lack of correspondence is explained in terms of "cognitive processes [which occur] after the sentence has been read". However, this lack of correspondence is more notorious in the current investigation and raises questions concerning a pure processing account.

5 Experiments

As mentioned before, the present study aims to investigate whether processing factors can account for Complex NP island violations, including both extraction from relative clauses and complement clauses of nouns in Hebrew.

In order to test the effect of processing factors on Complex NP island violations, both CNPRC and CNPCC in Hebrew were compared with different types of wh-phrases.

An acceptability judgment experiment on questions in Hebrew with CNPCC violations (5.2) and another similar experiment on questions with CNPRC violations (5.3) were performed. Since only the first experiment showed an effect for the different wh-phrases, it was complemented with a self-paced reading task on the same stimuli (5.4).

5.1 About the analysis

Data analysis was conducted in the R programming environment (R Development Core Team, 2010), using the linear mixed-effects model (LME) available as the package lme4 (Bates and Maechler, 2010). LMEs are regression models that take into account group-level variation and they include both fixed effects (such as predictors) and random effects.

The traditional by-participants and by-items calculation of ANOVA is not necessary in LMEs because participant and item level variation can be taken simultaneously into account in the model which increases statistical power reducing false negatives (Type II errors) while not increasing the risk of false positives (Type I errors) (Baayen et al., 2008; Baayen and Milin, 2010). Moreover, the ability to model unbalanced and incomplete repeated-measures data makes LMEs ideal for the analysis of the self-paced reading experiment carried out in the Experiment 3 (5.4).

However, estimating p-values for LMEs regression is a complex matter and different approximations can lead to different p-values. For large samples, the t distribution approximates the normal distribution and an absolute value of t larger than 2 indicates a 5% significance level (Baayen, 2008). However, for both large and small samples Markov chain Monte Carlo (MCMC) simulations can be used (Baayen et al., 2008). The p-values reported (p_{MCMC}) were estimated using MCMC sampling with 10000 samples using the pvals.fnc function from languageR package (Baayen, 2010) for R.

The complete results of the statistical analysis of the experiments are condensed in Appendix A.

5.2 Experiment 1 - Acceptability judgment task on questions with CNPCC violations.

Materials and Design The stimuli in this experiment consist of questions with CNPCC which varied with respect to properties of the filler-phrase. Each token set includes questions with CNPCC violations without a D-linked wh-phrase (D0), with a D-linked phrase with two degrees of informativity (D1 - D2), and a grammatical and acceptable baseline (B).

(13) (a) *ma doron cien et ha-uvda še-hu ra'a _ be-šavu'a še-avar? (D0)*
what Doron stressed ACC the fact that he saw _ in week that past?
'What did Doron stress the fact that he saw _ last week?'

(b) *eize seret doron cien et ha-uvda še-hu ra'a _ be-šavu'a še-avar?*
which movie Doron stressed ACC the fact that he saw _ in week that past?
(D1)

'Which movie did Doron stress the fact that he saw _ last week?'

(c) *eize me-ha-sratim haxi meša'amemim doron cien et ha-uvda še-hu ra'a*
which of the movies most boring Doron stressed ACC the fact that he saw
_ be-šavu'a še-avar? (D2)
_ in week that past?
'Which of the most boring movies did Doron stress the fact that he saw _ last week?'

(d) *ma doron cien še-hu ra'a _ be-šavu'a še-avar? (B)*
what Doron stressed that he saw _ in week that past?
'What did Doron stress that he saw _ last week?'

The experimental items for the experiment consisted of 16 sets of the same 4 sentences types presented above but with different lexical items, half of them included an inanimate wh-element (as the example before) and the other half an animate wh-element:

(14) (a) *et mi omri hikxiš et ha-te'ana še-hu harag _ lifney kama xodašim?*
ACC who Omri denied ACC the claim that he killed _ before some months?
(D0)

'Who did Omri deny the claim that he killed _ some months ago?'

(b) *et mi me-ha-mafionerim omri hikxiš et ha-ta'ana še-hu harag* _
ACC who of the mafia.members Omri denied ACC the claim that he killed _
lifney kama xodašim? (D1)
before some months?
'Which of the mafia members did Omri deny the claim that he killed _ some months ago?'

(c) *et mi me-ha-mafionerim ha-mevukašim omri hikxiš et ha-ta'ana še-hu*
ACC who of the mafia.members the wanted Omri denied ACC the claim that he
harag _ lifney kama xodašim? (D2)
killed _ before some months?
'Which of the most wanted mafia members did Omri deny the claim that he killed _ some months ago?'

(d) *et mi omri hikxiš še-hu harag _ lifney kama xodašim? (B)*
ACC who Omri denied that he killed _ before some months?
'Who did Omri deny that he killed _ some months ago?'

The 16 sets of items were distributed among four lists in a Latin square design, such that each list contained one version of each item, equal numbers of each condition and half of the items containing an inanimate wh-element while the other half an animate wh-element. Each list was combined with 32 filler items and was pseudo-randomized so that at least one filler item separated any two experimental items. Each questionnaire had instructions that asked the participants to rate the acceptability of each question using a 7-point Likert-type scale (from completely unacceptable (1) to perfectly acceptable (7))[2], and began with the same 7 practice items. Appendix B provides a complete list of the stimuli.

So each subject judged the sentences of one of the 4 surveys consisting of 55 sentences: 7 practice items at the beginning followed by one sentence of each of the 16 token sets together with the 32 fillers. Each participant saw only one of the four versions of each sentence, and each version was read by the same number of participants.

Predictions Regarding CNPCC violation in Hebrew, since the construction is parallel to the English one, similar results to Hofmeister and Sag's (2010) experiment were expected. In particular the CNPCCs with D-linking (13 b and 13 c) were expected to receive better judgments than the CNPCCs without D-linking (13 a) (but equal or worse than the baseline 13 d). Moreover if the informativity or semantic richness and not just D-linking of the filler is the relevant factor that eases processing (as was claimed in Hofmeister and Sag, 2010), it was expected to find also a significant difference between conditions with a less informative D-linking (13 b) and with a more informative one (13 c).[3]

Since D-linking is not associated with a specific structure of the wh-phrases, the structures used for inanimate wh-phrases *eize* N' (parallel to which N') and for animate wh-phrases *et mi me-N'* (ACC who from-N') should render the same results. However, since each of the matrix verbs of all the stimuli

[2] For a comparison of judgment tasks using Likert-type scale with the use of the magnitude estimation method see Dąbrowska (2010); Murphy and Vogel (2008); Cowart (1997); Weskott and Fanselow (2008).

[3] The more informative wh-phrases are also longer and more syntactically complex, this fact is addressed in the following discussion section and in section (5.5).

is only semantically compatible with an inanimate wh-phrase (since it is always a saying verb such as deny, claim, say, etc), there may be implications for the processing of animate and inanimate wh-phrases independently of their D-linking. These implications are developed in the following discussion part.

Participants 28 subjects aged between 22-32 received the questionnaire as an excel file and rated the sentences at their own rate in their personal computers. All participants reported to be native speakers of Hebrew and were naïve to the purpose of the study.

Analysis and Results One of the items was removed since not even the grammatical baseline was perceived as acceptable leaving 15 items. A linear mixed-effects model was built as specified in (15).

(15) *rating ~ condition * animacy + (1/tokenset) + (1/subject)*

The dependent variable, *rating*, appears to the left of the tilde operator, which indicates that the acceptability rating is a function or depends on the following parameters. In (15), the *rating* depends on the fixed effects *condition* and *animacy* which may interact between them as indicated by the '*'. The animacy factor is coded using -1 for inanimate wh-phrases and 1 for animate ones. The four levels of the condition factor are coded using Helmert contrast, comparing the conditions with islands (D0, D1 and D2) against the baseline (B), the D-linked CNPCCs (D1, D2) against the non-D-linked CNPCCs (D0) and the more informative D-linked CNPCC (D2) vs the less informative D-linked CNPCC (D1). In order to evaluate the difference between B and D0-2 conditions, another model is also fit using sum coding (comparing B to each condition). The specific coding does not affect the significance of any of the manipulations, however, α-value is now 0.025 (0.05/2) for the conditions and their interactions.

The random effects for Subject are specified as (1|*subject*); this factor introduces by-subject adjustments to the intercept (denoted by 1). The random intercept for sets of sentences is specified with *(1/tokenset)*, which indicates a random effect introducing adjustments to the intercept (again denoted by 1) conditional on which set the item belongs to. This random effect captures potential differences according to the choice of words of each token set; native speakers reported that they preferred certain saying verbs with NPs rather than with clauses or the other way around. Once "token set" is included as a random effect, a likelihood ratio test shows that the random effect of items is unnecessary.

The model is fitted using restricted maximum likelihood estimation (REML), a modification of maximum likelihood estimation that is more precise for mixed-effects modeling. REML seeks to find those parameter values that make the model's predicted values most similar to the observed values (Baayen et al., 2008).

The results[4] indicate that the difference between sentences with CNPCC islands and non islands (D0, D1, D2 vs B) is significant ($|t| = 24.6$, coef $= 0.85$, $p_{\mathrm{MCMC}} = 0.0001$) and the positive coefficient

[4] The full summary of model is reported in Table 1 on page 28 in the Appendix section.

indicates that non islands received higher acceptability judgments as expected.

The results also reveal that D-linked items as a whole are significantly more acceptable than the non-D-linked items with CNPCC violations (|t| = 4.11, coef =-0.2, $p_{\mathrm{MCMC}} = 0.0001$). This means that D-linking improves significantly the acceptability of sentences with CNPCC violations as predicted. However, there is no statistically significant difference between the two D-linked conditions (D2 vs D1).

After coding the condition factor using sum codes, the result reveals that D-linked CNPCC question ratings are well below the grammatical baseline and that this difference is also significant (for D1 vs B: |t| = 6.4, coef = -0.63 , $p_{\mathrm{MCMC}} = 0.0001$; and for D2 vs B: |t| = 6.1, coef = -0.66, $p_{\mathrm{MCMC}} = 0.0001$).

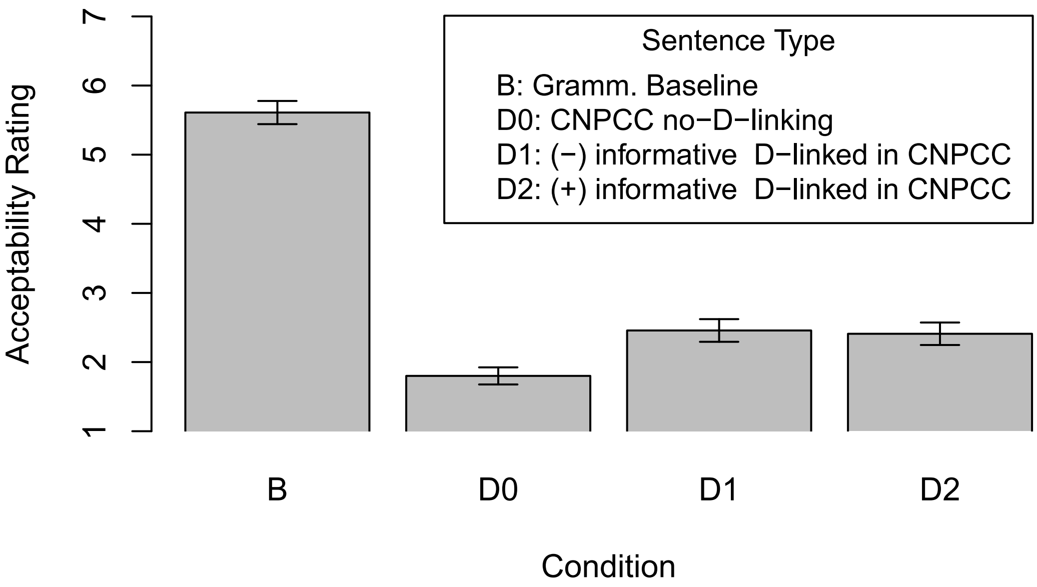

Figure 1: Mean judgment ratings for the different sentence types

Discussion The results clearly indicate that D-linking does improve the acceptability of the questions with CNPCC violations. At first glance there are two possible explanations: a processing account and an initial misparsing account. The first account is based on the fact that D-linked wh-phrases are assumed to facilitate processing at retrieval points (Kluender, 1998; Hofmeister, 2007). Then, the improvement in acceptability judgments implies that filler-gap dependencies in D-linked questions with CNPCC violations were processed more easily than filler-gap dependencies in their non-D-linked counterparts.

The initial misparsing account is based on the fact that D-linked wh-phrases carry more semantic information than non-D-linked wh-phrases. Since acceptability judgments are also reduced by an initial incorrect parsing, the extra information could help, in principle, to locate the right gap raising the acceptability. When the questions with CNPCC violations are processed, there are two potential places for the gap: an incorrect place, after the matrix verb (indicated as *_ in (16)) and the correct place after the embedded verb (indicated as _ in (16)).

The D-linked wh-phrases are, in all questions from the stimuli list, semantically implausible as complements of the matrix verbs (e.g. “which movie” as the complement of“stressed” in (16)), while the non D-linked ones are not (e.g. “what” as the complement of “stressed”). Then, it could be said that the questions with D-linking are more acceptable since an initial misparsing is avoided.

(16) *{ma; eize seret} doron cien *_ et ha-uvda še-hu ra’a _ be-šavu’a*
{what; which movie} Doron stressed *_ ACC the fact that he saw _ in week
še-avar?
that past?
‘{What; which movie} did Doron stress _ the fact that he saw _ last week?’

However, there is evidence that temporary implausible semantic representations do not affect the judgment of (acceptable) sentences with two possible gaps (Sprouse, 2008). More importantly, the results from the experiment show that animacy did not have an effect on judgments even though the animacy of the wh-phrase can be used as a semantic clue by the parser and could help to locate the right gap:

(17) (a) *{et mi; et mi me-ha-mafionerim} omri hikxiš *_ et ha-ta’ana*
{ACC who; ACC who of the mafia.members} Omri denied *_ ACC the claim
še-hu harag _ lifney kama xodašim?
that he killed _ before some months?
‘{Who; Which of the mafia members} did Omri deny _ the claim that he killed _ some months ago?’

None of the matrix verbs used would allow an animate complement; and items (both questions with CNPCC violations and baselines) with animate wh-phrases did not receive significantly better judgments than their counterparts with inanimate wh-phrases. There are two possible explanations and both reject the misparsing account presented here. Either semantic clues do not help avoiding an initial misparsing, or the help of the semantic clues do not have an effect on the acceptability judgments.

Both of these explanations suggest that the contribution of D-linking on judgments is not related to the possibility of avoiding (or helping to avoid) an initial misparsing. This strongly suggests the first interpretation: questions with CNPCC violations with D-linked wh-phrases received better judgments because filler-gap dependencies were processed more easily than in their non-D-linked counterparts. This supports the idea that once the parser finds the gap, the D-linked wh-phrases are more easily integrated into the sentence representation (Hofmeister, 2007; Kluender, 1998). However, these findings cannot base the view that the low acceptability of *CNPCC violations as a whole* is due to processing limitations. Experiment 3 complements the presented results and suggests that it is not the case.

D-linking showed to improve the judgments regardless of their semantic richness. Two hypotheses can explain this. The first hypothesis is that the relevant distinction is between D-linked and non-D-linked wh-phrases, independently of their informativity. The second hypothesis is that there is a trade-off between the processing difficulty of holding a longer wh-element (which even includes an embedded sentence) and the processing facilitation at the retrieval point. These competing hypotheses are tested with the online experiment since the trade-off will be translated in longer reading times after the wh-phrase is read and in shorter reading times at the embedded sentence inside the island.

Despite the improvement of D-linking on acceptability, figure 1 shows that the acceptability of the grammatical baseline is still much higher than the acceptability of the D-linked questions with CNPCC violations. As said before, this is unsurprising since D-linked sentences with CNPCC still possess a hard-to-process structure: a filler which depends on a gap that is in a clause dominated by a noun.

5.3 Experiment 2 - Acceptability judgment task on questions with CNPRC violations.

Materials and Design The stimuli in this experiment consists of questions with CNPRC violations which varied with respect to the properties of the wh-phrase. Each token set includes questions with CNPRC violations without a D-linked wh-phrase and with a D-linked phrase. In this case, however, an acceptable baseline minimally different from the experimental conditions cannot be devised, and the comparison is performed between the sentences with CNPRC violations.

(18)	*{ma;*	*eizo*	*mexonit}*	*Itamar*	*pagaš*	*et*	*ha-iš*	*še-maxar*	_	*lifney*	*šavu'a?*
	{what;	which	car}	Itamar	met	ACC	the man	that sold	_	before	week?

For this experiment, 16 sets of items were created and distributed among two lists in a Latin square design, such that each list contained one version of each item, equal numbers of each condition and half of the items containing an inanimate wh-element while the other half an animate wh-element. Each list was combined with the same 32 filler items used in Experiment 1 and was pseudo-randomized so that at least one filler item separated any two experimental items. Each questionnaire had instructions that asked the participants to rate the acceptability of each question using a 7-point Likert-type scale, and began with the same 7 practice items. Appendix C provides a complete list of the stimuli.

Predictions A previous experiment showed sentences with CNPRC violations to be extremely unacceptable. In fact, there was no significant difference between the acceptability of non-sense ungrammatical sentences and sentences with CNPRC violations (p = 0.83 (2-tails), t(100) = -0.21). While finding an improvement for the acceptability of D-linked CNPRC violations in Hebrew would be in line with Kluender's prediction that also the unacceptability of CNPRC can be explained with processing factors; a lack of improvement would support a pure grammatical account.

Participants 28 subjects aged between 22-32 (different from the subjects of Experiment 1) received the questionnaire as an excel file with instructions and rated the sentences at their own rate in their personal computers. All participants reported to be native speakers of Hebrew, and were naïve to the purpose of the study.

Analysis and results A linear mixed-effects model fitted using REML was built as in Experiment 1:

(19) *rating ~ condition * animacy + (1|tokenset) + (1|subject)*

The dependent variable, *rating* is again a function of the *condition*, *animacy* and their interaction. However, the condition fixed effect has only two levels: the non-D-linked CNPRCs (D0 coded as -1) and the D-linked CNPRCs (D1 coded as 1).

Contrarily to Experiment 1, D-linking does not have a significant effect on judgments and only animacy shows a significant effect ($|t| = 2.09$, coef = -0.13, $p_{\mathrm{MCMC}} = 0.04$)[5]. The positive value of the coefficient indicates that inanimate wh-phrases received better judgments than their animate counterparts.

Discussion Even though the results of the model indicate that an effect cannot not be found rather than there is no effect, the comparison with Experiment 1 strongly suggests that D-linking cannot improve the acceptability of CNPRC violations.

Regarding the effect of animacy, since the most unmarked kind of transitive construction is the one where the subject is animate and the object is inanimate (Comrie, 1989, p. 128), the most unmarked questions that include a transitive construction should have an inanimate wh-phrase. Even if CNPRC violations are found highly unacceptable by the subjects, it could be that an inanimate wh-phrase makes them more typical questions and slightly less unacceptable.

5.4 Experiment 3 - Self-paced reading task on questions with CNPCC violations.

This experiment examines reading times at critical regions for questions with CNPCC violations and complements the first acceptability judgment experiment.

Design In this experiment, subjects read sentences at their own pace on a computer screen word by word (self-paced reading technique), with each press of a key, a new word appears and the previous word disappears. The same four lists of sentences used in Experiment 1 were used.

The experiment was run online using Ibex[6] (previously WebSPR) 0.3-beta7 and 0.3-beta10 software by Alex Drummond adapted for Hebrew in his server. In self-paced reading, the amount of time between each keystroke is saved, and when comparing two minimally different conditions, longer RTs at a particular region are interpreted as an indication of processing difficulty (Just et al., 1982). In order to ensure attentive reading, on 10% of the stimuli, the sentences (including both fillers and experimental items) were followed by the task of choosing the most appropriate answer. The options presented included a

[5] The full summary of model is reported in Table 2 on page 28 in the Appendix section.

[6] Ibex is available at http://code.google.com/p/webspr/

possible and an impossible answer. The impossible answer had either an incorrect case (as in (20) with (21 a) and (21 b)) or it was semantically incorrect (as in (22) with (23 a) and (23 b)).

(20)	*eizo*	*mexonit*	*Yossi*	*šama*	*et*	*ha-sipur*	*še-hem*	*kanu*	_	*lifney*	*kama*	*yamim?*
	which	car	Yossi	heard	ACC	the story	that they-masc	bought	_	before	some	days?

‘Which car did Yossi hear the story that they bought _ some days ago?’

(21)	(a)	*et*	*ha-mercedez*	*ha-xadaša*	*(Right answer)*
		ACC	the Mercedez	the new	

(b)	*ha-mercedez*	*ha-xadaša*	*(Wrong answer)*
	the Mercedez	the new	

(22)	*ma*	*Dana*	*xasfa*	*et*	*ha-uvda*	*še-hem*	*ziyfu*	_	*lifney*	*šavu'a?*
	what	Dana	revealed	ACC	the fact	that they-masc	falsified	_	before	week?

‘What did Dana reveal the fact that they falsified _ a week ago?’

(23)	(a)	*še-hem*	*ziyfu*	*et*	*ha-mismaxim*	*(Right answer)*
		that they-masc	falsified	ACC	the papers	

(b)	*še-hi*	*ziyfa*	*et*	*ha-mismaxim*	*(Wrong answer)*
	that she	falsified	ACC	the papers	

Predictions The regions of interest for the RTs, are the words that follow the wh-phrase and the words in the embedded clauses of the different experimental conditions.

The CNPCCs with D-linking (13 b and 13 c) were expected to have faster RTs in the embedded clause than the CNPCCs without D-linking (13 a) (but equal or slower than the baseline).

Moreover, the hypothesis that the informativity of the filler is the relevant factor that eases processing (as claimed by Hofmeister 2007) would predict faster RTs in the embedded clause for CNPCCs with a more informative D-linking (13 c) than for the ones with a less informative one (13 b). On the other hand, longer RTs in the area after the wh-phrase for more informative D-linked wh-phrases would explain the lack of variation in acceptability ratings.

Participants 54 subjects aged between 22-37 (different from the subjects of Experiment 1 and 2) performed the self-paced reading task online in their personal computers and 200 shekels were raffled between the participants. All participants reported to be native speakers of Hebrew, and were naïve to the purpose of the study.

Data cleaning Before fitting the model the data was cleaned. Data sets for 14 subjects whose mean question-answer accuracy was below 80% were dropped from the analysis. The relatively large number of subjects with low overall percentage question-answer accuracy is likely to be due to the complexity of the stimuli and because the experiment was online. Since the experiment was done online over the Internet, some participants may have only played with it and not completed it seriously. Moreover, results of 3 subjects with global reading time average that differed from the entire data-set's global average by 2 standard deviations or more were excluded.

Individual outliers were removed from the data setting an upper threshold at log RT = 8 (∼ 3000ms) and a lower threshold at log RT = 5.2 (∼180 ms) which comprises 0.66% of the observations. RTs of less than 180 milliseconds are probably erroneous button presses (according to Baayen, 2008, ch. 7 visual uptake and response execution require at least 200 milliseconds) and RTs longer than ~3000 milliseconds do not represent real reading time and their source may be temporary distraction.

The same token set which was excluded from Experiment 1 since the baseline was not acceptable, was also removed from the analysis.

Analysis For the purposes of the analysis, there are two distinct areas of interest. The first area comprises two regions: the subject (*NP1*) and matrix verb (*V1*) which appear after the wh-phrase. The second area is the embedded sentence either after the noun head of the clause, in the case of the islands, or after the main verb, in the case of the baselines. This area comprises the complementizer prefixed to the subject which is always a pronoun (*C-NP2*), the embedded verb (*V2*) and the word appearing immediately after the embedded verb, which is the first word of an adjunct. The last word of the sentences was omitted from the analysis.

(24)	*eize*	*seret*	*doron*	*cien*	*et*	*ha-uvda*	*še-hu*	*ra'a*	_	*be-šavu'a*	*še-avar?*
	which	movie	Doron	stressed	ACC	the fact	that he	saw	_	in week	that past?
			[NP1	V1	]		[C-NP2	V2		Ad	]

For the area which comprises the embedded sentence (*C-NP2*, *V2* and *Ad*), a linear mixed-effects model was built as specified in 25.

(25) logRT ~ condition * animacy + c_wordlength + c_spillover + (1|subject) + (1|tokenset)

In the LME, the logarithm of the RTs is used instead of raw RT to reduce the effect of extreme reaction times (see Baayen, 2008). The dependent variable, *logRT,* is as in the previous two experiments a function of the *condition*, *animacy* and their interaction. The four levels of the condition factor were coded using Helmert contrast in the main analysis and sum coding only when it was relevant. The factor *c_wordlength* models the fact that shorter words are usually read faster than longer words; this factor is centered to eliminate possible spurious correlations (Baayen, 2008). Including *c_spillover* (centered spillover effect) from the preceding word eliminates the possibility that it is the source of the changes in RTs. Since the end of one response measure is immediately followed by the beginning of another, together with a new portion of text, any uncompleted processing will *spill over* from one response measure to the next one (Mitchell, 1984, p. 76 as cited in Vasishth et al., 2010). Since the regions preceding the critical ones were not identical between the conditions with and without island violations, the amount of spillover from the preceding word could significantly differ in the contrasting conditions. Mitchell (1984, p. 76) explains that since certain aspects of processing will be postponed and will join a buffer so that they can be dealt with later, then the RTs may be influenced also by any processing that may have built up in the buffer (as cited in Vasishth et al., 2010).

The random factors *(1/subject)* and (1|*tokenset*) are used as in Experiment 1 and 2 to model the individual differences among participants' reading rates and the differences between the sets of sentences.

For the first area which comprises the matrix subject and the matrix verb, the baseline and the non-D-linked CNPCC conditions do not differ since the participants cannot distinguish between them yet. For the LME, instead of condition, the factor wh-phrase type is used.

(26) *logRT ~ wh_type * animacy + c_wordlength + c_spillover + (1/subject) + (1/tokenset)*

Results The complete results of the statistical analysis are condensed in Tables 3, 4, 5, 6 and 7 in Appendix A.

As expected the spillover factor had a significant effect in all the regions (*NP1*: |t| = 9.28, coef = 0.38, p_{MCMC} = 0.0001; *V1*: |t| = 8.56, coef = 0.35, p_{MCMC} = 0.0001; *C-NP2*: |t| = 8.06, coef = 0.3, p_{MCMC} = 0.0001; *V2*: |t| = 11.11, coef = 0.38, p_{MCMC} = 0.0001; *Ad*: |t| = 11.64, coef = 0.36, p_{MCMC} = 0.0001). The target region was read more slowly, the higher the log RT on the preceding word, as can be seen from the positive coefficients.

The effect of word length is only significant for region *Ad* (|t| = 2.31, coef = 0.037, p_{MCMC} = 0.03), the region was read more slowly, the longer the word. Even though shorter words should be associated with shorter reading times in every region, two reasons may explain the difference between region *Ad* and the rest of the regions.

Eyes move in a series of jumps, remaining relatively stationary between these jumps (Staub and Rayner, 2007), these jumps, known as “saccades” have been found to have an average length of 5.5 characters for Hebrew (Pollatsek et al., 1981). Words in all the examined regions were between 3 and 6 characters, either the absence of more than one saccade or the relatively small variation on length can explain the lack of word length effect.

However, only *Ad* presents a high variation on the types of lexical items, namely, different types of prepositions, adverbs, bare nouns and nouns with an attached preposition. The effect of word length may be the result of the correlation between highly frequent words or functional words with shorter length. Since frequent and functional words have been shown to be read faster (Staub and Rayner, 2007), the correlation may explain the effect of word length.

First Area - NP1 and V1 At the subject of the matrix sentence (*NP1*), the first region after the wh-phrase is read, there is no effect of D-linking. Only the informativity effect of the D-linked wh-phrases is significant, this region is processed the fastest when the D-linked wh-phrase is more informative (|t| = 2.19, coef = 0.033, $p_{MCMC} = 0.032$).

At the matrix verb (*V1*), the RTs are not significantly different for different wh-phrases types. However, there is a marginal interaction between animacy and the effect "informativity" (|t| = 2.03, coef = 0.037, $p_{MCMC} = 0.05$).

Second Area - V2, C-NP2 and Ad The results reveal that D-linking provokes significantly faster RTs in sentences with CNPCC violations (|t| = 2.79 , coef = 0.031, $p_{MCMC} = 0.005$), as predicted, for the beginning of the embedded clause (*C-NP2*). Both D-linking conditions (D1 and D2) as a whole elicit faster RTs than the non-D-linked items with CNPCC violations; while there is no significant difference between D1 and D2. In this region, RTs for the baseline are not significantly different from RTs for condition D0. Moreover, the region is read faster than the baseline for conditions with CNPCC violations and D-linking (significantly for D1 condition (|t| = 2.31, coef = -0.054, $p_{MCMC} = 0.024$) and mildly significant for D2 (|t| = 2.16 , coef = -0.05 , $p_{MCMC} = 0.032$[7]).

The embedded verb region (*V2*) shows a different pattern, it is read faster for all the island conditions (D0, D1, D2) than for the grammatical baseline (B) (|t| = 2.84 , coef = 0.019, $p_{MCMC} = 0.004$).

Finally, the region after the verb (*Ad*) shows no significant difference between the conditions. However, this is a region with high variation in lexical items (and on RTs). The lexical items include variation in word classes and according to the *Word-frequency Database in Hebrew* (Frost and Plaut, 2010), also very different frequencies. Further investigation into the data reveals that if frequency and its interaction is included into the model, this region behaves similarly to the beginning of the embedded clause: RTs for D1-D2 are shorter than for D0 (|t| = 0.44 , coef = 2.74 , $p_{MCMC} = 0.011$) and there is no significant difference between D1 and D2; RTs for condition B are larger than for condition D1 (|t| = 2.84 , coef = -0.10, $p_{MCMC} = 0.004$) but shorter than for D0 (|t| = 3.24, coef = 0.12, $p_{MCMC} = 0.002$). There are many complex interactions between frequency and the conditions (see Table 7 on page 32 for the whole output of the model) that are significant; in fact, for the highly frequent words (such as *lifney* "before" and *rak* "only") D-linking does not speed processing but slows it down.

Discussion The results identify a slowdown immediately after processing a more informative and complex D-linked wh-phrase, but only at the first word that follows the wh-phrase. In fact, the less "informative" D-linked form is also read slower than the bare wh-phrase as illustrated in Figure 2, however, if the spillover from the previous word is controlled, this difference is not significant. The different results between this work and previous investigations (Hofmeister, 2007; De Vincenzi, 1996)

[7] Since two models are fit with different coding of the condition factor, α is 0.025.

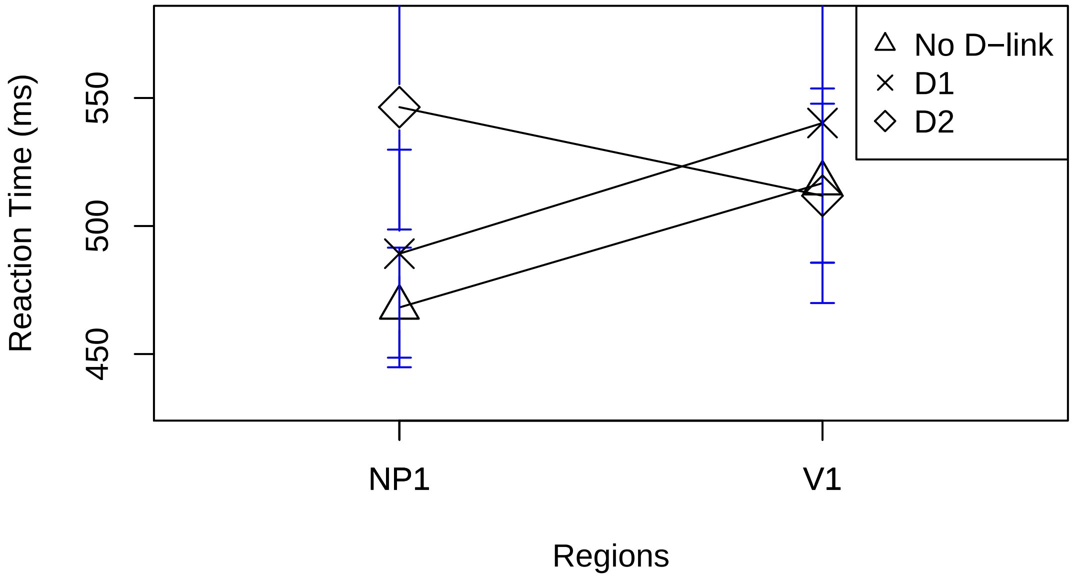

Figure 2: RTs for different wh-phrases at the first area.

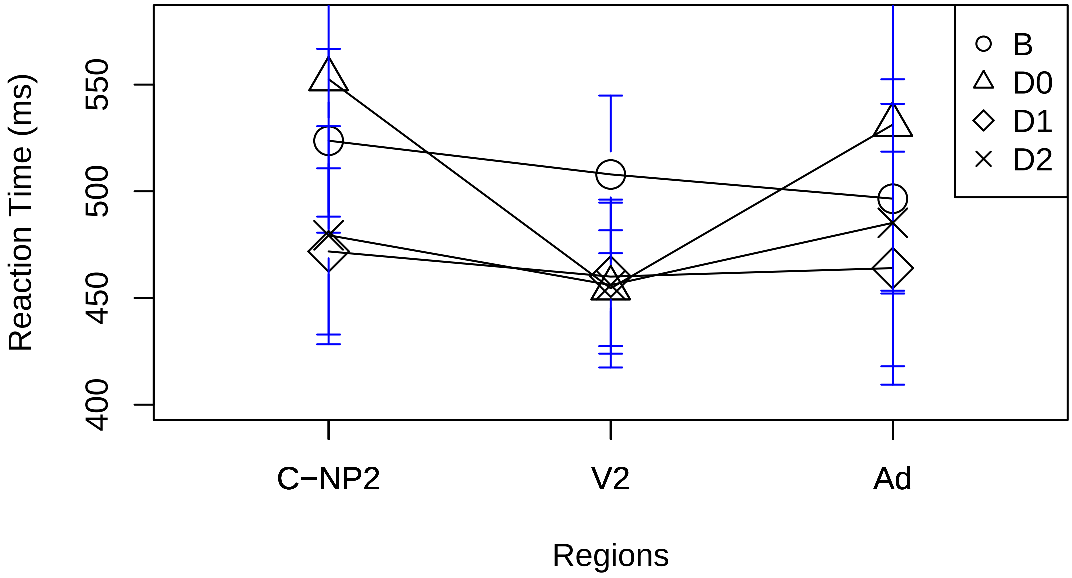

Figure 3: RTs for the different conditions at the second area (with high frequent words removed from Region *Ad*).

that report a slow down after a D-linked wh-phrase is read can be attributed to the lack of the spillover control. In fact, if it is removed from the model, the difference between D-linked conditions *is* significant.

Since the more informative wh-phrases are also the longest and the more complex syntactically, the slowdown may be attributed to either the more complex semantic representation as claimed by Hofmeister (2007); Kluender (1998), or the more complex syntactic form independently of its semantic value.

The faster RTs at the beginning of the embedded sentence when the wh-phrase is D-linked suggest that they can reduce the computational effort of building the new sentence. The lack of difference in RTs, when the wh-phrase are D-linked but with different levels of semantic richness or informativity, may indicate that either this factor is not relevant, a ceiling effect (meaning that the fastest RTs are already achieved with the less informative phrase) or some interaction with their syntactic complexity. For further investigation, D-linked wh-phrases with the same syntactic complexity but different levels of semantic richness should be compared, e.g. “which man” vs. “which astronaut”.

Even though there is a complex interaction with frequency, D-linking as whole facilitates the processing in the region immediately after the verb (at least for medium frequent words).

The fact that the last region of the embedded sentence is processed faster for the baseline than for the D0 conditions together with the results of Experiment 1 (where the baseline got better acceptability judgments) may in principle support the view that the unacceptability of CNPCC violations is due to the processing effort of retrieving the wh-element inside the embedded sentence. On the other hand, even though results of Experiment 1 show that D-linking does not improve the acceptability of D-linked sentences with CNPCC violations to the level of the grammatical baseline, all the regions of the embedded sentence are processed even faster for the D-linked sentences with CNPCC violations than for the baseline.

The discrepancies between these results and the results of Hofmeister and Sag (2010) fall into place given the fact that the grammatical baseline they use has a D-linked wh-element; while the one used here has a bare non-D-linked wh-element. The comparison of their results with the current ones suggests that the processing of filler-gap dependencies is independent of CNPCC effects.

The results show that processing differences within CNPCC violations and the baseline are not matched by contrasts in acceptability judgments, which is not compatible with the view that limitations on the cognitive resources related to the filler-gap dependencies process are responsible for CNPCC violation effects. If the processing of the filler-gap dependencies is facilitated, the overall acceptability arises. However, the filler-gap dependency processing in an embedded sentence governed by a noun (a CNPCC violation) can yield faster RTs than in a non-island embedded sentence and the overall acceptability level of the non-island would still be much higher.

Still, it may be that the unacceptability is a product of processing difficulties, figure 4 reveals that questions with CNPCC violations have a peak of long RT at the head noun (H in the graphic) of the

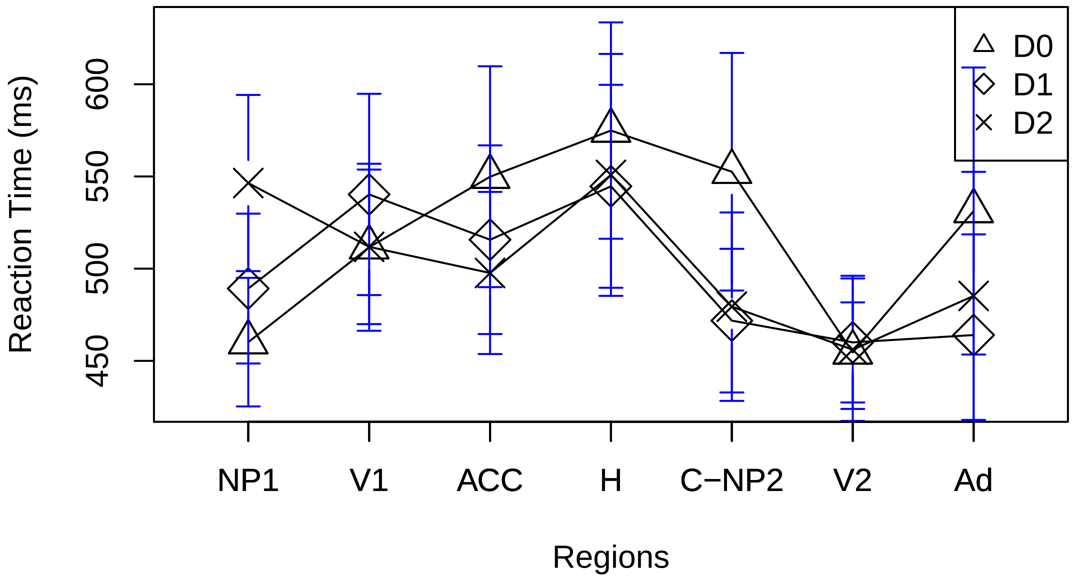

Figure 4: RTs for the island conditions (with high frequent words removed from Region *Ad*).

island which is absent in non-islands and it is not significantly different for the island conditions (D0-2). However, this is the region where the parser discovers a violation. At the region of the head noun, the parser has to assume that either there is no gap at all or that the gap is inside the CNPCC. Then the low acceptability and long RTs can also be explained as the reaction to a grammatical violation.

5.5 General Discussion

The results strongly suggest that CNPRC and CNPCC are two very distinct phenomena. Firstly, it was previously attested that the acceptability judgments of CNPCC was significantly higher than the acceptability of CNPRC. Secondly, while CNPCC violations are susceptible to manipulations attributed to sentence processing, such as D-linking of the wh-phrase, CNPRCs seem to be not affected by those.

While the absence of an effect is virtually impossible to prove, the comparison between Experiment 1 where an effect is found on CNPCCs and Experiment 2 where an effect is not found on CNPRCs, even though each subject is exposed to more items of each condition, strongly suggests a lack of effect.

Hence, the lack of effect of D-linking on sentences with CNPRC violations shows that the parser cannot be "helped" by the change in referentiality, suggesting that the parser do not try to assign the filler to the right gap, and "gives up" the sentence's processing. Either the processing difficulties associated with CNPRC violations are of a kind which cannot be improved with the change of referentiality provoked by the D-linking of the wh-phrase, or the source of the unacceptability of the CNPRC violations is not processing. The first option seems implausible due to the evidence that D-linking does improve the acceptability of many other different structures associated with processing difficulties. It seems more plausible that the filler-gap dependency cannot be constructed and the sentence is unacceptable due to the fact that it is ungrammatical.

Regarding the CNPCC, the implications are less straightforward. From Experiment 1, it seems that at least part of the unacceptability is due to processing factors. Hofmeister and Sag (2010) claim that "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". The results from Experiment 3 show that processing differences within islands and non-islands are not matched by contrasts in acceptability judgments. In fact, the processing of the filler-gap dependency is not substantially different in an embedded sentence of a question with a CNPCC violation than in an embedded sentence of a grammatical baseline.

However, the retrieval site of the gap in the embedded sentence is processed faster in sentences without a CNPCC violation than in a similar sentence with the CNPCC violation (and the same kind of wh-phrase). This difference can be attributed to the fact that questions with CNPCC violations are longer since they have an extra head that governs the embedded sentence. But if the processing of the filler-gap dependency in the CNPCC question is alleviated by D-linking the wh-element, this trend is reverted. In that case, the retrieval site of the gap is processed faster in a sentence with a CNPCC violation than in a non-island as shown in Experiment 3.

In sum, part of the unacceptability of the sentences with CNPCC violations is due to the processing difficulties involved in parsing a sentence with a filler-gap dependency and these processing difficulties can be ameliorated. However, these processing difficulties seem not to be the reason for the unacceptability of the violation, and they seem to explain the same kind of acceptability variation that is found between yes-no questions and wh-questions (Kluender, 1998).

On the other hand, as figure 4 shows, the unacceptability of CNPCC violations *is* related with processing efforts. There is a major slowdown in processing at the head of the island which is absent in non-islands. However, since this is the region where the parser "discovers" the island, the source of the processing difficulties is uncertain. Either the unacceptability and the processing efforts at the head of the island are a consequence of the ungrammaticality of sentences with CNPCC violations or these sentences are grammatical and the unacceptability is a consequence of the processing efforts of building the derivation of the island. In order to verify the latter hypothesis, further research should compare the RTs at the head nouns of complement clauses in questions with and without CNPCC violations such as:

(27) (a) {What; Which house} did Itay hear the **rumor** that Roy bought _ in Kansas?

(b) When did Itay hear the **rumor** that Roy bought a house in Kansas ?

The treatment of CNPCC violations if the source of the processing difficulties is grammatical depends on certain assumptions about the grammar.

If grammar is assumed to allow gradient grammaticality, violations such as CNPCC would affect grammaticality (and then acceptability) relatively mildly allowing these kind of sentences to be parsed. The processing slowdown in violations as seen in the head of the complex NP would be a “penalty” for violating a grammatical constraint.

If grammar is only allowed to be gradient-free, all violations should affect grammaticality in the same way. Some have identified a class of semi-sentences, meaning ungrammatical utterances that are comprehensible (Katz, 1964 as cited in Schütze, 1996) and sentences with CNPCC violations would enter to this class. The interpretation of ungrammatical sentences has been shown not to be random but common to most listeners and the process of interpretation seems to be governed by syntactic, morphophonemic, semantic-pragmatic, and heuristic considerations (Shanon, 1973). Under this view, speakers would interpret sentences with CNPCC violations according to some heuristic considerations, and their parsing would be as close as can be of a sentence without such violation. Then it would be susceptible to the same processing factors as a grammatical sentence. The unacceptability elicited by this violation would be then attenuated by the fact that the sentence is interpretable. Under this view, even though ungrammatical sentences may not derived by the computational system, speakers may understand the intended meaning of an ungrammatical sentences by “repairing” their form with syntactic, morphophonemic, semantic-pragmatic, and heuristic considerations when it is possible.

6 Conclusion

This study aims to investigate to what extent processing factors can account for Complex NP island violations in Hebrew.

By eliciting acceptability judgments and reading times of sentences with CNPCC and CNPRC violations in Hebrew, while a factor which eases processing is manipulated, it is possible to examine which violations are susceptible to processing factors.

While CNPCC violations are susceptible to a manipulation attributed to sentence processing such as D-linking of the wh-phrase, CNPRCs seem to be not affected by it. It seems that the filler-gap dependency cannot be constructed when there is a CNPRC violation and sentences that violate this constraint are unacceptable due to the fact that they are ungrammatical.

Regarding the CNPCC, the implications are less straightforward. The results from Experiment 3 show that processing differences within islands and non-islands are not matched by contrasts in acceptability judgments. The results of Experiment 1 together with Experiment 3 show that the processing of the filler-gap dependency is not substantially different in an embedded sentence of a question with a CNPCC violation than in an embedded sentence of a grammatical baseline. The processing difficulties associated with the filler-gap dependency can be ameliorated but they do not seem to be the main source of unacceptability of sentences with CNPCC violations.

However, there is a major slowdown in processing at the head of the island which is absent in non-islands. Since this is the region where the parser "discovers" the island, the source of the processing difficulties is uncertain.

Appendices

A Statistical Results

Predictor	Coef	SD	t	p_{MCMC}
(Intercept)	3.078125	0.231381	13.303	0.0001*
D2 vs D1	0.015482	0.084290	0.184	0.8582
D1, D2 vs D0	-0.199958	0.048665	-4.109	0.0001*
D0, D1, D2 vs B	0.845064	0.034411	24.558	0.0001*
D2 vs B	-0.66059	0.10323	-6.399	0.0001*
D1 vs B	-0.62962	0.10323	-6.099	0.0001*
D0 vs B	-1.24498	0.10323	-12.06	0.0001*
animacy = *animate*	0.136161	0.159160	0.855	0.3494
D2 vs D1 : animacy = *animate*	0.069054	0.084290	0.819	0.4176
D1, D2 vs D0 : animacy = *animate*	-0.021387	0.048665	-0.439	0.6492
D0, D1, D2 vs B : animacy = *animate*	-0.007615	0.034411	-0.221	0.8486
D2 vs B : animacy = *animate*	-0.04005	0.10323	-0.388	0.7098
D1 vs B : animacy = *animate*	0.09806	0.10323	0.95	0.3292
D0 vs B : animacy = *animate*	-0.03516	0.10323	-0.341	0.7308

(log-likelihood = -732.6; $\alpha = 0.025$ for the conditions and their interactions with animacy; $\alpha = 0.05$ for animacy)

Table 1: Summary of the fixed effects in the LME for Experiment 1

Predictor	Coef	SE	t	p_{MCMC}
(Intercept)	1.94196	0.16020	12.122	0.0001*
D0 vs D1	0.01786	0.04959	0.360	0.7196
animacy = *animate*	-0.12500	0.05988	-2.087	0.0422*
D0 vs D1 : animacy = *animate*	0.01339	0.04959	0.270	0.7794

(log-likelihood = -697.4; $\alpha = 0.05$)

Table 2: Summary of the fixed effects in the LME for Experiment 2

Predictor	Coef	SD	t	p_{MCMC}
(Intercept)	6.1335533	0.0322572	190.15	0.0001*
D2 vs D1	-0.0393553	0.0179967	-2.19	0.0318
D1, D2 vs no-D-linking	-0.0116303	0.0085874	-1.35	0.2294
animacy = *animate*	-0.0005109	0.0138224	-0.04	0.965
c_wordlength	0.0250935	0.0150885	1.66	0.13
c_spillover	0.3796899	0.0409042	9.28	0.0001*
D2 vs D1 : animacy	-0.0307007	0.0179669	-1.71	0.0938
D1, D2 vs no-D : animacy	-0.0017364	0.0084643	-0.21	0.8176

(log-likelihood = -167.4 ; α = 0.05)

Table 3: Summary of the fixed effects in the LME for Experiment 3, Region *NP1*

Predictor	Coef	SD	t	p_{MCMC}
(Intercept)	6.141605	0.038950 1	57.68	0.0001*
D2 vs D1	0.026954	0.018269	1.48	0.1238
D1, D2 vs no-D-linking	0.003317	0.008652	0.38	0.6066
animacy = *animate*	-0.033615	0.01873	-1.79	0.0834
c_wordlength	0.016101	0.019416	0.83	0.3376
c_spillover	0.350690	0.040969	8.56	0.0001*
D2 vs D1 : animacy	-0.036876	0.018198	-2.03	0.0516(*)
D1, D2 vs no-D : animacy	0.009187	0.008579	1.07	0.2918

(log-likelihood = -180.7 ; α = 0.05)

Table 4: Summary of the fixed effects in the LME for Experiment 3, Region *V1*

Predictor	Coef	SD	t	p_{MCMC}
(Intercept)	6.0888563	0.0323997	187.93	0.0001*
D2 vs D1	-0.0019183	0.0190803	-0.10	0.9262
D1, D2 vs D0	0.0309609	0.0110905	2.79	0.0052*
D0, D1, D2 vs B	0.0212755	0.0078068	2.73	0.0052*
D2 vs B	-0.0503182	0.0232966	-2.16	0.0322(*)
D1 vs B	-0.0541547	0.0234892	-2.31	0.0240*
D0 vs B	0.0406463	0.0235474	1.73	0.1002
animacy = *animate*	-0.0093735	0.0171472	-0.55	0.5768
c_wordlength	-0.0678783	0.036267	-1.87	0.0798
c_spillover	0.3001283	0.037259	8.06	0.0001*
D2 vs D1 : animacy	-0.0267916	0.0191586	-1.40	0.1868
D1, D2 vs D0 : animacy	0.0006402	0.0110612	0.06	0.9776
D0-2 vs B : animacy	0.0018439	0.0078151	0.24	0.8418
D2 vs B : animacy	0.0243076	0.0233185	1.04	0.3164
D1 vs B : animacy	-0.0292757	0.0235925	-1.24	0.2572
D0 vs B : animacy	-0.0005636	0.0234751	-0.02	0.9546

(log-likelihood = -206.4; $\alpha = 0.025$ for the conditions and their interactions with animacy; $\alpha = 0.05$ for the rest of the factors)

Table 5: Summary of the fixed effects in the LME for Experiment 3, Region *C-NP2*

Predictor	Coef	SD	t	p_{MCMC}
(Intercept)	6.073726	0.025441	238.73	0.0001*
D2 vs D1	0.004731	0.016473	0.29	0.7694
D1, D2 vs D0	-0.007506	0.009664	-0.78	0.4142
D0, D1, D2 vs B	0.019181	0.00675	2.84	0.0036*
animacy = *animate*	-0.005144	0.013748	-0.37	0.7244
c_wordlength	-0.00353	0.017766	-0.20	0.839
c_spillover	0.384470	0.034607	11.11	0.0001*
D2 vs D1 : animacy	0.001945	0.016534	0.12	0.8518
D1, D2 vs D0 : animacy	0.001420	0.009572	0.15	0.8938
D0-2 vs B : animacy	0.008944	0.006721	1.33	0.1794

(log-likelihood = -124.3; $\alpha = 0.05$)

Table 6: Summary of the fixed effects in the LME for Experiment 3, Region *V2*

Predictor	Coef	SD	t	p_{MCMC}
(Intercept)	6.137e+00	3.157e-02	194.38	0.0001*
D2 vs D1	-3.740e-02	2.760e-02	-1.36	0.1854
D1, D2 vs D0	4.382e-02	1.599e-02	2.74	0.0106*
D0, D1, D2 vs B	-5.397e-04	1.114e-02	-0.05	0.9844
D2 vs B	-2.991e-02	3.526e-02	-0.85	0.4266
D1 vs B	-1.027e-01	3.616e-02	-2.84	0.004*
D0 vs B	1.169e-01	3.602e-02	3.24	0.002*
animacy = *animate*	9.294e-03	1.388e-02	0.67	0.4988
c_frequency	-4.153e-05	2.703e-05	-1.54	0.1394
c_spillover	4.099e-01	4.080e-02	10.04	0.0001*
D2 vs D1 : animacy	-3.466e-02	1.730e-02	-2.00	0.0496
D1, D2 vs D0 : animacy	6.985e-03	9.962e-03	0.70	0.5118
D0-2 vs B : animacy	-3.175e-03	7.005e-03	-0.45	0.6374
D2 vs B : animacy	3.055e-02	2.218e-02	1.38	0.1738
D1 vs B : animacy	-4.999e-02	2.237e-02	-2.23	0.0294(*)
D0 vs B : animacy	1.818e-02	2.225e-02	0.82	0.4378
D2 vs D1 : c_frequency	9.770e-05	3.657e-05	2.67	0.0066*
D1, D2 vs D0 : c_frequency	-5.591e-05	2.048e-05	-2.73	0.0134*
D0-2 vs B : c_frequency	2.381e-06	1.448e-05	0.16	0.9482
D2 vs B : c_frequency	-2.722e-05	4.711e-05	-0.58	0.537
D1 vs B : c_frequency	1.654e-04	4.704e-05	3.52	0.0004*
D0 vs B : c_frequency	-1.704e-04	4.611e-05	-3.70	0.0004*

(log-likelihood = 31.73; $\alpha = 0.025$ for the conditions and their interactions with animacy; $\alpha = 0.05$ for the rest of the factors)

Table 7: Summary of the fixed effects in the LME for Experiment 3, Region *Ad*

B Sentences used in Experiment 1 and 3[8]

Set	Condition	Sentence
1	D0	מה יוסי שמע את הסיפור שהם קנו לפני כמה ימים?
1	D1	איזו מכונית יוסי שמע את הסיפור שהם קנו לפני כמה ימים?
1	D2	איזו מהמכונית שנבחנו יוסי שמע את הסיפור שהם קנו לפני כמה ימים?
1	B	מה יוסי שמע שהם קנו לפני כמה ימים?
2	D0	מה דורון ציין את העובדה שהוא ראה בשבוע שעבר?
2	D1	איזה סרט דורון ציין את העובדה שהוא ראה בשבוע שעבר?
2	D2	איזה מהסרטים הכי משעממים דורון ציין את העובדה שהוא ראה בשבוע שעבר?
2	B	מה דורון ציין שהוא ראה בשבוע שעבר?
3	D0	מה שמעון אישש את החשד שהוא יבטל בעוד שבוע?
3	D1	איזה פרויקט שמעון אישש את החשד שהוא יבטל בעוד שבוע?
3	D2	איזה מהפרויקטים שנבחנו שמעון אישש את החשד שהוא יבטל בעוד שבוע?
3	B	מה שמעון אישש שהוא יבטל בעוד שבוע?
4	D0	מה חן הוכיחה את ההשערה שהוא איבד לפני שבוע וחצי?
4	D1	איזה מסמך חן הוכיחה את ההשערה שהוא איבד לפני שבוע וחצי?
4	D2	איזה מהמסמכים שנחתמו חן הוכיחה את ההשערה שהוא איבד לפני שבוע וחצי?
4	B	מה חן הוכיחה שהוא איבד לפני שבוע וחצי?
5	D0	מה דנה חשפה את העובדה שהם זייפו לפני בניית המלון?
5	D1	איזה אישור דנה חשפה את העובדה שהם זייפו לפני בניית המלון?
5	D2	איזה מהאישורים שהעירייה דורשת דנה חשפה את העובדה שהם זייפו לפני בניית המלון?
5	B	מה דנה חשפה שהם זייפו לפני בניית המלון?
6	D0	מה שרה הבינה את הדרישה שהיא תעזוב בסוף החודש?
6	D1	איזו מפלגה שרה הבינה את הדרישה שהיא תעזוב בסוף החודש?
6	D2	איזו מהמפלגות ששייכות לאופוזיציה שרה הבינה את הדרישה שהיא תעזוב בסוף החודש?
6	B	מה שרה הבינה שהיא תעזוב בסוף החודש?
7	D0	מה טלי גילתה את העובדה שהם גנבו במהלך המסיבה?
7	D1	איזה דיסק טלי גילתה את העובדה שהם גנבו במהלך המסיבה?
7	D2	איזה מהדיסקים שקשה להשיג טלי גילתה את העובדה שהם גנבו במהלך המסיבה?
7	B	מה טלי גילתה שהם גנבו במהלך המסיבה?
8	D0	מה גילי הפנימה את העובדה שהם יאבדו במהלך המשבר?
8	D1	איזה נכס גילי הפנימה את העובדה שהם יאבדו במהלך המשבר?
8	D2	איזה מהנכסים המשתלמים גילי הפנימה את העובדה שהם יאבדו במהלך המשבר?
8	B	מה גילי הפנימה שהם יאבדו במהלך המשבר?

Table 8: Stimuli of experiment 1 and 3 - Sets 1-8

[8] Set 13 was presented to the subjects in both experiments but was excluded from the analysis.

Set	Condition	Sentence
9	D0	את מי רון הדגיש את העובדה שהוא הכיר רק אתמול בצהרים?
9	D1	את מי מהעובדים רון הדגיש את העובדה שהוא הכיר רק אתמול בצהרים?
9	D2	את מי מהעובדים שפוטרו רון הדגיש את העובדה שהוא הכיר רק אתמול בצהרים?
9	B	את מי רון הדגיש שהוא הכיר רק אתמול בצהרים?
10	D0	את מי עמיר אישר את הידיעה שהוא מינה אתמול בערב?
10	D1	את מי מהפוליטיקאים עמיר אישר את הידיעה שהוא מינה אתמול בערב?
10	D2	את מי מהפוליטיקאים שנבחרו עמיר אישר את הידיעה שהוא מינה אתמול בערב?
10	B	את מי עמיר אישר שהוא מינה אתמול בערב?
11	D0	את מי עומרי הכחיש את הטענה שהוא הרג לפני כמה חודשים?
11	D1	את מי מהמאפיונרים עומרי הכחיש את הטענה שהוא הרג לפני כמה חודשים?
11	D2	את מי מהמאפיונרים המבוקשים עומרי הכחיש את הטענה שהוא הרג לפני כמה חודשים?
11	B	את מי עומרי הכחיש שהוא הרג לפני כמה חודשים?
12	D0	את מי מיכל הציעה את הרעיון שהיא תעסיק מחר בצהרים?
12	D1	את מי מהסטודנטים מיכל הציעה את הרעיון שהיא תעסיק מחר בצהרים?
12	D2	את מי מהסטודנטים הכי חכמים מיכל הציעה את הרעיון שהיא תעסיק מחר בצהרים?
12	B	את מי מיכל הציעה שהיא תעסיק מחר בצהרים?
13	D0	את מי גילה הדליפה את הידיעה שהוא שיחד לפני הפגישה?
13	D1	את מי מהשוטרים גילה הדליפה את הידיעה שהוא שיחד לפני הפגישה?
13	D2	את מי מהשוטרים המעורבים בפרשה גילה הדליפה את הידיעה שהוא שיחד לפני הפגישה?
13	B	את מי גילה הדליפה שהוא שיחד לפני הפגישה?
14	D0	את מי טלי הבהירה את הטענה שהם הטעו במהלך המשפט?
14	D1	את מי מעורכי הדין טלי הבהירה את הטענה שהם הטעו במהלך המשפט?
14	D2	את מי מעורכי הדין הנחשבים טלי הבהירה את הטענה שהם הטעו במהלך המשפט?
14	B	את מי טלי הבהירה שהם הטעו במהלך המשפט?
15	D0	את מי דביר כתב את ההודעה שהוא תבע ביום ראשון?
15	D1	את מי מהעיתונאים דביר כתב את ההודעה שהוא תבע ביום ראשון?
15	D2	את מי מהעיתונאים שנחקרו דביר כתב את ההודעה שהוא תבע ביום ראשון?
15	B	את מי דביר כתב שהוא תבע ביום ראשון?
16	D0	את מי מזל פרסמה את ההודעה שהיא תתגמל במהלך החודש?
16	D1	את מי מהעובדים מזל פרסמה את ההודעה שהיא תתגמל במהלך החודש?
16	D2	את מי מהעובדים שהצטיינו מזל פרסמה את ההודעה שהיא תתגמל במהלך החודש?
16	B	את מי מזל פרסמה שהיא תתגמל במהלך החודש?

Table 9: Stimuli of experiment 1 and 3 - Sets 9-16

C Sentences used in Experiment 2

Set	Condition	Sentence
1	D0	מה יוסי מכיר מישהו שקנה לפני כמה ימים?
1	D1	איזו מכונית יוסי מכיר מישהו שקנה לפני כמה ימים?
2	D0	מה דורון פגש מישהי שראתה בשבוע שעבר?
2	D1	איזה סרט דורון פגש מישהי שראתה בשבוע שעבר?
3	D0	מה דביר מצא מישהי שמוכרת לפני ימים בודדים?
3	D1	איזה ספר דביר מצא מישהי שמוכרת לפני ימים בודדים?
4	D0	מה דן ראיין מישהו ששדד בשנה שעברה?
4	D1	איזה בנק דן ראיין מישהו ששדד בשנה שעברה?
5	D0	מה מזל הכירה מישהי שגנבה בסוף השבוע?
5	D1	איזה מסמדמזל הכירה מישהי שגנבה בסוף השבוע?
6	D0	מה איה צריכה מישהו שמדבר לשבוע הבא?
6	D1	איזו שפה איה צריכה מישהו שמדבר לשבוע הבא?
7	D0	מה דנה תבעה מישהו שכתב לפני חודש?
7	D1	איזו כתבה דנה תבעה מישהו שכתב לפני חודש?
8	D0	מה שרה העסיקה מישהי שעיצבה לפני כמה שנים?
8	D1	איזה מוצר שרה העסיקה מישהי שעיצבה לפני כמה שנים?
9	D0	את מי עמיר חקר מישהו שמינה לפני חצי שנה?
9	D1	את מי מהפוליטיקאים עמיר חקר מישהו שמינה לפני חצי שנה?
10	D0	את מי עומרי הסטיר מישהו שהרג לפני כמה חודשים?
10	D1	את מי מהמאפיונרים עומרי הסטיר מישהו שהרג לפני כמה חודשים?
11	D0	את מי מיכל פיטרה מישהי שהעסיקה לפני שבועיים?
11	D1	את מי מהסטודנטים מיכל פיטרה מישהי שהעסיקה לפני שבועיים?
12	D0	את מי עוז עצר מישהו ששיחד לפני יומיים?
12	D1	את מי מהשוטרים עוז עצר מישהו ששיחד לפני יומיים?
13	D0	את מי טלי גילתה מישהו שהטעה במהלך המשפט?
13	D1	את מי מעורכי הדין טלי גילתה מישהו שהטעה במהלך המשפט?
14	D0	את מי דביר הטריד מישהו שתבע ביום ראשון?
14	D1	את מי מהעיתונאים דביר הטריד מישהו שתבע ביום ראשון?
15	D0	את מירון קידם מישהי שפיטרה לפני שבוע וחצי?
15	D1	את מי מהעובדים רון קידם מישהי שפיטרה לפני שבוע וחצי?
16	D0	את מיגילי הרגיזה מישהו שראיין במהלך החודש האחרון?
16	D1	את מי מהשחקנים גילי הרגיזה מישהו שראיין במהלך החודש האחרון?

Table 10: Stimuli of experiment 2

References

Allwood, J. 1982. The complex NP constraint in Swedish. In *Readings on unbounded dependencies in scandinavian languages*, ed. E. Engdahl and E. Ejerhed, 15–32. Stockholm: Almqvist & Wiksell.

Baayen, R. H. 2008. *Analyzing linguistic data: A practical introduction to statistics using R.* Cambridge: Cambridge University Press.

Baayen, R. H. 2010. *languageR: Data sets and functions with "Analyzing Linguistic Data: A practical introduction to statistics"*. URL http://CRAN.R-project.org/package=languageR, r package version 1.0.

Baayen, R. H., D. J. Davidson, and D. M. Bates. 2008. Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language* 59:390–412.

Baayen, R. H., and P. Milin. 2010. Analyzing reaction times. *To appear in International Journal of Psychological Research.* Submitted.

Bates, D., and M. Maechler. 2010. *lme4: Linear mixed-effects models using s4 classes.* URL http://CRAN.R-project.org/package=lme4, r package version 0.999375-34.

Bever, T.G., and J.M. Carroll. 1981. On some continuous properties of language. In *The cognitive representation of speech*, ed. T. Myers, J. Laver, and J. M. Anderson. Amsterdam: North-Holland.

Boeckx, C. 2008. Islands. *Language and Linguistics Compass* 2:151.

Boston, M. F. 2010. The role of memory in wh-island gradience. In *Workshop on Cognitive Modeling and Computational Linguistics*. Uppsala, Sweden: Association for Computational Linguistics.

Chen, E., E. Gibson, and F. Wolf. 2005. Online syntactic storage costs in sentence comprehension. *Journal of Memory and Language* 52:144–169.

Chomsky, N. 1965. *Aspects of the Theory of Syntax*. Cambridge, MASS: MIT press.

Chomsky, N. 1978. *Topics in the theory of generative grammar*. The Hague: Mouton De Gruyter.

Chomsky, N., and G. A. Miller. 1963. Introduction to the Formal Analysis of Natural Languages. *Handbook of Mathematical Psychology* 269.

Cinque, G. 1990. *Types of A'-dependencies*. MIT press.

Comrie, B. 1989. Language universals and linguistic typology: Syntax and morphology .

Cowart, W. 1997. *Experimental syntax: Applying objective methods to sentence judgments*. Thousand Oaks: Sage Pubns.

Dąbrowska, E. 2010. Naïve v. expert intuitions: An empirical study of acceptability judgments. *The Linguistic Review* 27:1–23.

De Vincenzi, M. 1996. Syntactic analysis in sentence comprehension: Effects of dependency types and grammatical constraints. *Journal of Psycholinguistic Research* 25:117–133.

Engdahl, E. 1997. Relative clause extractions in context. *Working Papers in Scandinavian Syntax* 60:51–79.

Fanselow, G., and S. Frisch. 2006. Effects of processing difficulty on judgments of acceptability. *Gradience in grammar* 291–316.

Featherston, S. 2007. Data in generative grammar: The stick and the carrot. *Theoretical linguistics* 33:269–318.

Fodor, J.D. 1978. Parsing strategies and constraints on transformations. *Linguistic Inquiry* 9:427–473.

Frazier, L., and C. Clifton. 2002. Processing d-Linked Phrases. *Journal of Psycholinguistic Research* 31:633–659.

Frazier, L., and A. Flores. 1989. Filler driven parsing: A study of gap filling in Dutch. *Journal of Memory and Language* 28:331.

Frost, R., and D. Plaut. 2010. The word-frequency database for printed Hebrew. Retrieved at September 2010.

Gibson, E. 2000. The dependency locality theory: A distance-based theory of linguistic complexity. In *Image, language, brain: Papers from the first mind articulation project symposium*, 95–126.

Goodluck, Helen. 1997. Islands, parsing and learnability: A commentary on some experimental assessments of children's knowledge of island constraints. In *10 th Annual CUNY Conference on Human Sentence Processing, Santa Monica, CA, USA*.

Grewendorf, G. 2007. Empirical evidence and theoretical reasoning in generative grammar. *Theoretical Linguistics* 33:369–380.

Grosu, A. 1982. The extragrammatical content of certain "Island Constraints". *Theoretical Linguistics* 9:17–68.

Hawkins, J. A. 1999. Processing complexity and filler-gap dependencies across grammars. *Language* 244–285.

Hofmeister, P. 2007. Retrievability and gradience in filler-gap dependencies. In *Proceedings from the Annual Meeting of the Chicago Linguistic Society*, volume 43, 109–123. Chic Ling Society.

Hofmeister, P., T. F. Jaeger, I. A. Sag, I. Arnon, and N. Snider. 2007. Locality and accessibility in wh-questions. *Roots: Linguistics in search of its evidential base* 185.

Hofmeister, P., and I. A. Sag. 2010. Cognitive constraints and island effects. *Language* 86.

Just, M. A., P. A. Carpenter, and J. D. Woolley. 1982. Paradigms and processes in reading comprehension. *Journal of Experimental Psychology: General* 111:228–238.

Katz, J. J. 1964. Semi-sentences. In *The structure of language: Readings in the philosophy of language*, ed. J. A. Fodor and J. J. Katz, volume 1964, 400–416. Englewood Cliffs, New Jersey: Prentice-Hall.

Kluender, R. 1992. Deriving island constraints from principles of predication. *Island constraints: Theory, acquisition and processing* 15:223–58.

Kluender, R. 1998. On the distinction between strong and weak islands: A processing perspective. *Syntax and semantics* 241–280.

Kluender, R. 2004. Are subject islands subject to a processing account. In *Proceedings of WCCFL*, volume 23, 475–499.

Kluender, R., and M. Kutas. 1993a. Bridging the gap: Evidence from ERPs on the processing of unbounded dependencies. *Journal of Cognitive Neuroscience* 5:196–214.

Kluender, R., and M. Kutas. 1993b. Subjacency as a processing phenomenon. *Language and Cognitive Processes* 8:573–633.

Kuno, S. 1976. Subject, theme, and the speaker's empathy -a reexamination of relativization phenomena. *Subject and topic* 417:444.

Lakoff, R. 1973. Questionable answers and answerable questions. *Issues in linguistics: Papers in honor of Henry and Renee Kahane* 453–467.

Mitchell, D.C. 1984. An evaluation of subject-paced reading tasks and other methods for investigating immediate processes in reading. *New methods in reading comprehension research* 69–90.

Murphy, B., and C. Vogel. 2008. An Empirical Comparison of Measurement Scales for Judgements of Linguistic Acceptability. In *Poster presented at the Linguistic Evidence Conference*.

Pesetsky, D. 1987. Wh-in-situ: Movement and unselective binding. *The representation of (in) definiteness* 98–129.

Pollatsek, A., S. Bolozky, A. D. Well, and K. Rayner. 1981. Asymmetries in the perceptual span for Israeli readers. *Brain and Language* 14:174–180.

R Development Core Team. 2010. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL `http://www.R-project.org`, ISBN 3-900051-07-0.

Ross, J.R. 1967. Constraints on variables in syntax. Doctoral Dissertation, MIT.

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

Shanon, B. 1973. Interpretation of ungrammatical sentences. *Journal of Verbal Learning and Verbal Behavior* 12:389–400.

Sorace, A., and F. Keller. 2005. Gradience in linguistic data. *Lingua* 115:1497–1524.

Sprouse, J. 2007. A program for experimental syntax. *College Park, MD: University of Maryland dissertation* .

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

Staub, A., and K. Rayner. 2007. Eye movements and on-line comprehension processes. *The Oxford handbook of psycholinguistics* 327–342.

Szabolcsi, A. 2006. Strong vs. weak islands. *The Blackwell Companion to Syntax* 4:479–531.

Vasishth, S., and R. L. Lewis. 2006. Argument-head distance and processing complexity: Explaining both locality and antilocality effects. *Language* 82:767–794.

Vasishth, S., K. Suckow, R. L. Lewis, and S. Kern. 2010. Short-term forgetting in sentence comprehension: Crosslinguistic evidence from verb-final structures. *Language and Cognitive Processes* 25:533–567.

Warren, T., and E. Gibson. 2002. The influence of referential processing on sentence complexity. *Cognition* 85:79–112.

Weskott, T., and G. Fanselow. 2008. Variance and Informativity in Different Measures of Linguistic Acceptability. In *Proceedings of the 27th West Coast Conference on Formal Linguistics, ed. Natasha Abner and Jason Bishop*, 431–439.