

Locality and Accessibility in *Wh*-Questions

Philip Hofmeister, T. Florian Jaeger, Ivan A. Sag, Inbal Arnon, Neal Snider

1. Competing *Wh*-Orders¹

Even in relatively configurational languages, such as English, speakers frequently have a choice between different constituent orders. Many of these word order variations have been linked to complexity (Hawkins 2005; *inter alia*). For example, heavy-NP shift is more likely if the shifted NP is more complex than the NP it shifts over (Wasow 1997). Other cases of word order variations, however, have not been considered in these terms. The choice between different *wh*-phrase orders, as in (1), has been said to be determined by (categorical) grammatical constraints, such as Superiority (Kuno and Robinson 1972, Chomsky 1973; *inter alia*).

- | | | |
|--------|-------------------|-----------------------------|
| (1) a. | Who bought what? | Non-SUV |
| b. | What did who buy? | Superiority Violation (SUV) |

According to such accounts, (1b) is ungrammatical in English. These accounts, however, do not predict the findings of Arnon et al. (2005) and Clifton et al. (2006), both of whom present evidence from corpora, attesting the usage of Superiority violating examples. Nor can they accommodate the gradient nature of the contrast that has emerged in several studies (Featherston 2005; Fedorenko & Gibson 2006). In Arnon et al. (2005), we examined an alternative account, which we dubbed the *Wh*-Processing Hypothesis, which treats *wh*-phrase ordering as being subject to the same type of constraints as other word order variations. The *Wh*-Processing Hypothesis predicts that speakers disprefer more complex *wh*-dependencies. Here we examine to what extent factors known to affect the processing of filler-gap dependencies (FGDs) also affect the relative acceptability of different *wh*-phrase orders. We focus, in particular, on two factors in the processing of *wh*-questions: locality and accessibility. These factors play significant roles in the processing of FGDs in general, as we discuss below. One of our

goals in this paper is to explore the extent to which these factors can explain SUVs.

In the next section, we define and discuss the two factors of locality and accessibility, showing how these factors have been previously related to processing difficulty. In section 2.2, we present the *Wh*-Processing Hypothesis. In section 3, we present the results of three acceptability surveys and one reading time study which test the effects of the above-mentioned factors on the processing and acceptability of questions. Finally, in section 4, we discuss the implications of these results, other possible factors, and potential problems with this account.

2. Locality and Accessibility

The first factor we consider here is the locality of the dependency. Gibson (2000), Hawkins (2005), and many others observe that the distance between the filler and gap strongly affects the processing difficulty and relative acceptability of sentences with FGDs. For example, English object relatives, as compared to the shorter subject relatives, require more resources and increase processing difficulty, as indicated by reading times, question-answer accuracy, and lexical-decision tasks (King & Just 1991; *inter alia*). Since *wh*-interrogative dependencies are also non-local, it is reasonable to assume that they are subject to the same processing constraints as relative clauses. In fact, the lack of a specified, identifiable referent associated with a *wh*-interrogative filler potentially presents an additional cognitive challenge. Hence, we hypothesize that locality is also likely to play an important role in determining the acceptability of multiple *wh*-phrase (interrogative) constructions.

It has also been noticed that the type of the *wh*-filler (*which*-NP vs. bare *wh*-item) influences the acceptability of SUVs. Karttunen (1977) points out that examples like (2) sound better than (3):

- (2) Which class of drug will which patient get?
- (3) What will who get?

Pesetsky (1987) further notices that the type of the in-situ *wh*-phrase affects acceptability independently, so that (4) is judged better than (3):

- (4) What will which patient get?

Pesetsky ascribes this difference to "D(iscourse)-linking" of the *which*-NP, which exempts it from the normal conditions on *wh*-phrase ordering. The proposal, however, that the type of *wh*-filler and *wh*-intervener affect grammaticality is both ad hoc and without independent motivation. We propose that the factors explaining SUVs are both more general and independently motivated. We discuss next how *wh*-order preferences, widely discussed under the label of D-linking, relate to more general processing mechanisms. Specifically, we believe there is a strong relationship between the form and content of an expression and its degree of activation, which has been described in terms of accessibility (Ariel 1990) and that this degree of activation strongly impacts the processing of the FGD.

FGDs have been shown to be affected by the referential properties of material intervening between the filler and the gap. For example, in sentences like (5), verbs are read fastest when the relative clause subjects are pronouns, while first or famous names lead to faster reading times than definite descriptions:

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

Warren & Gibson (2002, 2005) interpret these results in terms of accessibility (Ariel 1990, 2001): the more accessible the intervening referents, the less burden there is on the processor, which is already taxed by maintaining the filler-gap dependency. Accessibility is a measure of activation level, which is partially indicated by the choice of referring expression. The form of an NP acts as a cue to the listener as to how much work is necessary to activate or retrieve the correct antecedent. As information and morphological complexity in the NP increase, the amount of work necessary to retrieve the antecedent also increases. Processing less accessible forms, therefore, requires more work and hence creates an additional processing difficulty while an FGD is being parsed.

Interrogative *wh*-dependencies, like other FGDs, also exhibit sensitivity to the properties of intervening material. Alexopolou and Keller (2003) show that words associated with a higher cognitive cost appearing between a *wh*-filler and gap impair the integration of *wh*-phrases with (the subcategorizer of) the gap. There is also evidence from German that certain intervening *wh*-phrases improve the acceptability of superiority-violating, multiple *wh*-questions: German speakers disprefer bare in-situ *wh*-phrases in SUVs (e.g. *wer*), as compared to complex *wh*-phrases (e.g. *welcher Mann*;

Featherston 2005). We interpret these results as reflecting the increased processing difficulty introduced by bare *wh*-words.

Locality and accessibility thus constitute the focus of this study. Before we turn to the predictions we make about these factors and how they influence the processing of *wh*-dependencies, we first address in detail how accessibility applies to *wh*-phrases.

2.1. Accessibility: *Wh*-phrases versus referential NPs

While accessibility has been almost exclusively applied to referential NPs, we propose that the same mechanisms that influence the processing of referential NPs are also at play during the processing of *wh*-phrases. We dwell on this subject here in order to address the issue of why the explicitness of intervening *wh*-phrases and referential NPs affect processing difficulty in seemingly different ways. As pointed out above, more explicitness correlates with more processing difficulty for referential NPs, but the opposite seems to be true for *wh*-phrases. To explain this difference, we consider here some hypotheses about the most important predictors of activation for *wh*-phrases and referential NPs.

For referential NPs, morphologically simple and less informative NPs (e.g. pronouns) are used to refer to entities of higher activation or salience, while morphological complexity and high informativity (e.g. definite descriptions) indicate that the referent is less activated at the time of utterance (Ariel 2001). Thus, the choice between a pronoun or a definite description is conditioned by the salience of that particular individual in the preceding discourse. Notice that it only makes sense to compare the accessibility of two phrases when they have the same intended interpretation (i.e. both phrases have the same referent).

In addition to marking a current degree of activation, the form of NPs also partially *determines* the degree of activation subsequent to their utterance—referred to as future accessibility by Ariel (2001). In short, the more explicit an NP is, the greater the subsequent increase in activation of the corresponding referent(s). Increases in activation not only make subsequent references with higher current accessibility markers more likely, they also facilitate other linguistic operations that involve that information, such as the integration of fillers and gaps. Thus, all other things being equal, the referent of an expression like *the gorilla approaching at breakneck speed*, as opposed to *it*, is more likely to become the discourse topic and have a higher activation level at subsequent points in the utterance.

In support of this view, Gernsbacher (1989), presents evidence that proper names reactivate an antecedent more strongly than a pronoun. From this perspective, current activation marking is in an inverse relation to future activation marking. A higher accessibility marker like a personal pronoun indicates high current accessibility, but does relatively little to increase activation. As Ariel (2001:68) notes, this "can explain why speakers shift to lower accessibility markers from time to time, even when they continue to discuss the same discourse entity." That is, to maintain topicality, speakers use longer and more explicit forms on occasion to compensate for normal activation decay and interference from other discourse entities. The same reasoning, we hypothesize, applies to *wh*-phrases: all other things being equal, the concept of politicians is more salient after an utterance of *which politician* (in context) than after *who*.

Wh-phrases, too, have a range of possible forms from morphologically simple and uninformative (e.g. *who*) to more complex forms that package more information (e.g. *which politician*) to ever more complex and informative forms (e.g. *which politician from Missouri*). Given the greater degree of morphological complexity and explicitness in *which*-NPs, we categorize them as higher future accessibility markers. Moreover, Frazier and Clifton (2002) provide evidence that *which*-NPs are better antecedents for pronouns than bare *wh*-words like *who* and *what*. Since high future accessibility phrases encourage the subsequent use of high current accessibility anaphors (i.e. pronouns), the relation between explicitness and future activation is thus the same for anaphoric and *wh*-expressions. Preliminary results from reading-time experiments conducted by the first author also favor this ranking. In unary, *wh*-island constructions with supporting contexts (*Which employee/Who did Albert learn whether they dismissed after the annual performance reviews?*), *which*-NPs lead to significantly faster reading times than a bare *wh*-item at the embedded verb and in subsequent regions. Accordingly, the evidence from Featherston and Frazier & Clifton can all be seen to reflect the fact that *which*-phrases are more accessible than simple *wh*-pronouns at the time that fillers and gaps are integrated.

If the difficulty of processing a head is a function (among other things) of the activation levels of its arguments, then the form preferences for both *wh*-questions and referential NPs emerge as a preference for high argument activation at the point when the head is processed. In examples like (7) from Warren & Gibson (2005), variants with highly salient personal pronouns will have the highest argument activation, because activation starts high and hence can withstand more decay and/or interference effects.

(6) It was $\left\{ \begin{array}{c} \text{you} \\ \text{Patricia} \\ \text{the lawyer} \end{array} \right\}$ who $\left\{ \begin{array}{c} \text{we} \\ \text{Dan} \\ \text{the businessman} \end{array} \right\}$ avoided at the party.

In contrast, argument activation starts low or at zero in multiple *wh*-questions, but is boosted higher when more information is expressed in the *wh*-phrases. Therefore, a *which*-phrase in either argument position of an SUV should satisfy the preference for higher activation at the verbal head.

This still leaves a noticeable distinction between the processing of referential NPs and *wh*-phrases. Recall that highly salient, but less informative NPs serve as the best kind of intervening referential NP (Warren and Gibson 2005). The above-cited data on *wh*-phrases, however, appears to indicate that more explicit and informative *wh*-phrases are preferred as interveners. Assuming that processing ease depends upon activation level, this means that *wh*-phrase interveners are most activated when the *wh*-phrase is explicit, while referential NP interveners are most activated when the form is not very informative but marks a highly salient referent.

One way to account for the apparently different effects of explicitness is to point to the simple fact that interrogative *wh*-phrases are not anaphoric. Anaphoric NPs are used to refer back to discourse referents previously mentioned. In other words, they evoke information already in the common ground (explicitly or implicitly). Hence, the primary task in processing a referential NP is retrieving the correct antecedent or, failing that, accommodating the existence of an antecedent. This whole process is expedited when the referent or mental entity is highly salient at the point the anaphor is reached.² A processing benefit for more explicit anaphoric forms is not apparent in the Warren & Gibson (2005) results.³ This does not preclude the possibility of some positive correlation between explicitness and activation boosting with respect to anaphoric NPs; instead, the results permit the view that the effect of activation boosting is obscured by the profound effect of salience in that study. One way to account for this is to argue that pronouns, proper names, and definites differ too much in their current activation levels (due to the need to express important differences in salience) for boosting to make much difference. On this view, it is the property of being an anaphor that causes activation boosting to be relatively unimportant.

In contrast, *wh*-phrases do not function as anaphors, although parts of their interpretation may derive from the preceding discourse. Rather, *wh*-phrases are used to construct complex objects—questions—which either

seek to gain information (as in main clause interrogatives) or else to make a clausal argument that can be predicated over (as in embedded interrogatives). Questions, therefore, will be more easily understood (and better answered, for that matter) when either a) the context strongly provides the focus of the question or b) the *wh*-phrase itself explicitly narrows down the scope of the inquiry. Under the assumption that *wh*-phrases have either low or zero activation prior to their utterance (which follows from their non-anaphoric function), using a more explicit *wh*-form should facilitate the retrieval and integration process. In other words, because the initial activation is so low, activation strength is largely dependent on activation boosts that are, in turn, dependent on explicitness. This hypothesis is consistent with all the *wh*-phrase data considered so far.⁴

In sum, we propose that the apparent differences in the effect (size) of explicitness can be attributed to *wh*-expressions being non-anaphoric. This proposal makes two interesting predictions for future research: a) an indefinite phrase should lead to faster processing at the verb if the indefinite phrase is more explicit (contains more information); b) in the right context, it may even be possible to observe effects of activation boost for anaphoric expressions (see footnote 3)—in such contexts, more explicit anaphoric NPs should lead to faster processing.

2.2. The *Wh*-Processing Hypothesis

Based on these observations of how locality and accessibility affect FGDs, we propose the following *Wh*-Processing Hypothesis to account for the relative rareness of examples like (1b) in English, as compared to non-superiority violating orders like (1a):

(7) The *Wh*-Processing Hypothesis

- a. Factors that have been shown to burden the processing of referential filler-gap dependencies (e.g. relative clauses) burden the processing of all FGDs, including *wh*-interrogative constructions.
- b. Many filler-gap sentences that have standardly been analyzed as ungrammatical (violating 'island' constraints) are in fact grammatical, but are judged to be less acceptable by speakers because they are harder to process.

The reasoning implicit in (7b) builds on recent proposals to better understand the relation between speaker judgments and processing factors. See, for example, Fanselow and Frisch 2004.

This hypothesis entails that speakers faced with a choice between several grammatical *wh*-orders, will disprefer those which (given the context) are associated with a greater processing cost. Combined with existing theories of processing complexity (e.g. Gibson 2000), the *Wh*-Processing Hypothesis makes the following predictions about *wh*-questions:

- (I) In filler-gap constructions, the greater the distance between the filler and its gap, the less acceptable the sentence.
- (II) Less accessible fillers make filler-gap sentences less acceptable.
- (III) Less accessible interveners make filler-gap sentences less acceptable.

Note that we make no assumptions about the relative importance of these predictions. That is, we do not conjecture whether the effect of distance is more important than accessibility or vice versa; nor does the *Wh*-Processing Hypothesis indicate if the accessibility of the filler is paramount to that of the interveners or vice versa.

3. Experimental Evidence

3.1. *Methods*

We present here the results of three surveys eliciting acceptability judgments and one experiment measuring comprehension complexity in *wh*-questions via self-paced reading.⁵ Acceptability judgments were elicited over the WWW using magnitude estimation (ME; Bard et al. 1996) with the WebExp software package (Keller et al. 1998). ME lets participants set their own continuous acceptability scale, allowing participants to express as many distinctions as desired. Acceptability judgments are made relative to a reference sentence. Participant's judgments are subsequently standardized by dividing by the reference sentence's score. All ME analyses are based on the *z*-score⁶ of these (log-transformed) standardized judgments. For the reading time study, residual reading times were used for the analysis. This method reduces variability due to individual differences in reading times.⁷

All experiments use Latin-square design: Each participant saw each item in exactly one condition, and all conditions occur equally often. All lists

include at least as many fillers as experimental items. All results were analyzed using repeated measures analyses of variance (ANOVAs).

Participants for the ME experiments were recruited via e-mail lists and online discussion forums. The reading-time study was conducted as part of another reading-time study at MIT's Tedlab.

3.2. Locality Effects on Acceptability (ME1)

3.2.1. Materials

ME1 investigates the effect of locality on the acceptability of *wh*-questions (Prediction I). Locality-based processing theories (e.g. Gibson 2000) predict that an increase in distance between filler and gap (measured in new discourse referents) makes *wh*-dependencies harder to process. We manipulated this distance by optionally attaching a six-word PP either to the *which*-phrase (8c,f) or to the other NP (8b,e). In addition, the *which*-phrase was either subject-extracted (8a-c) or object-extracted (8d-f):

- (8) a. Which man saw the girl?
- b. Which man saw the girl in the bar on California Ave?
- c. Which man in the bar on California Ave. saw the girl ?
- d. Which man did the girl see?
- e. Which man did the girl in the bar on California Ave. see?
- f. Which man in the bar on California Ave. did the girl see?

We hypothesized that longer filler-gap distances would engender higher processing costs, which would result in lower acceptability judgments. For example, the filler in (8d) is separated from the gap by only one new discourse referent, *the girl*; but in (8e), three new discourse referents intervene between the filler and the gap. Thus, we predict (8e) to be judged less acceptable than (8d). Notice that we further predict a difference between (8b) and (8e) for the same reasons, despite the roughly equivalent lengths of the questions. In general, Prediction I says that subject-extractions should be judged more acceptable than object extractions.

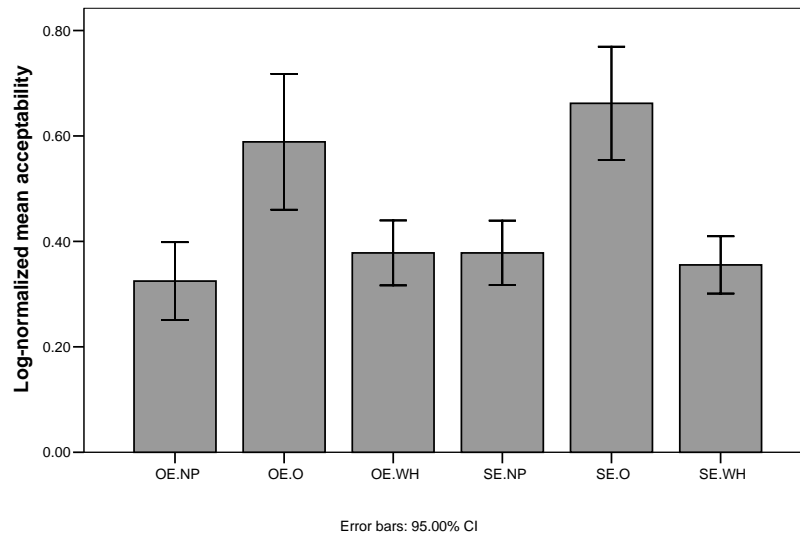
The study includes 36 items in six different conditions. In addition, 34 fillers were included in each list. 18 of these came from another multiple-*wh* experiment. 42 native English speakers completed the survey, but the results from one individual were removed because of incomplete data for that subject. Participation did not result in compensation.

3.2.2. Results

As shown in Table 1, object extractions (which have more intervening discourse referents) were judged as less acceptable than subject extractions in the subject but not the item analysis ($F(1,35) = 4.9, p < .05$; non-significant by items, $F(1,35) = 2.5, p = .12$). While the difference between examples like (8a) and (8d) turned out to be non-significant, this is not surprising since neither question involves more than one intervener and the number of interveners differed by only one. Notably though, the object *wh*-question with three intervening NPs (8e) was judged less acceptable than the subject *wh*-question (8b) of the same length with zero interveners. Overall, sentences were judged differently from each other if the difference in number of interveners was two or more. This may mean that, for simple unary *wh*-questions, it takes at least two interveners to invoke any measurable cognitive challenge.

Figure 1- Acceptability ratings from ME1 (OE = object extraction; SE = subject extraction; O = no attachment; WH = PP attached to *wh*-phrase; NP = PP attached to referential NP)

Extraction x Attachment



Pairs of Extraction.Attachment		Difference in # of interveners	Subj. analysis	Item Analysis
OE.NP (8e)	SE.NP (8b)	3	p<.01	p<.05
	SE.WH (8c)	3	p=.17	p=.29
	OE.WH (8f)	2	p<.1	p<.05
OE.WH (8f)	SE.NP (8b)	1	p=.52	p=.97
	SE.WH (8c)	1	p=.38	p=.14
SE.WH (8c)	SE.NP (8b)	0	p=.15	p=.18

Table 1. Pairwise comparisons of the six conditions in ME1, including the difference in interveners for each pair. (OE = object extracted; SE = subject extracted; NP = six-word PP is attached to referential NP; WH = six-word PP is attached to *wh*-phrase)

3.3. Accessibility Effects on Acceptability (ME2)

3.3.1. Materials

In ME2, we addressed the issue of how accessibility affects acceptability. To do this, we manipulated the accessibility of both the object-extracted *wh*-filler (*what* vs. *which book*) and the intervening subject *wh*-phrase (*who* vs. *which boy*). All questions were embedded SUVs, as in (9):

- (9) a. Mary wondered what who read.
 b. Mary wondered which book who read.
 c. Mary wondered what which boy read.
 d. Mary wondered which book which boy read.

According to our predictions, examples with higher accessibility fillers and interveners should be preferred to those with low accessibility fillers and interveners. In other words, examples like (9d) should be judged the most acceptable and examples like (9a) the least acceptable. We are agnostic about the possibility of an interaction between filler and intervener accessibility, and so do not make any claims about how cases like (9b) and (9c) will be ordered with respect to each other. However, since one preference is satisfied in both (9b) and (9c), we expect that these cases are more acceptable than SUVs with two low accessibility *wh*-phrases, but less acceptable than SUVs with two high accessibility *wh*-phrases.

Twenty items with 4 conditions each appeared in the experiment, as exemplified above. 42 people participated in this experiment over the web without any compensation.

3.3.2. Results

The results confirm the prediction that less accessible *wh*-interveners (the in-situ *wh*-phrase) decrease acceptability ($F_1(1,37) = 64.5$, $F_2(1,19) = 248.1$, $P_s < .001$): Interveners with a lower activation at the verbal head decreased acceptability: examples like (9a-b) were judged worse than those in (9c-d), as illustrated in the graph below.

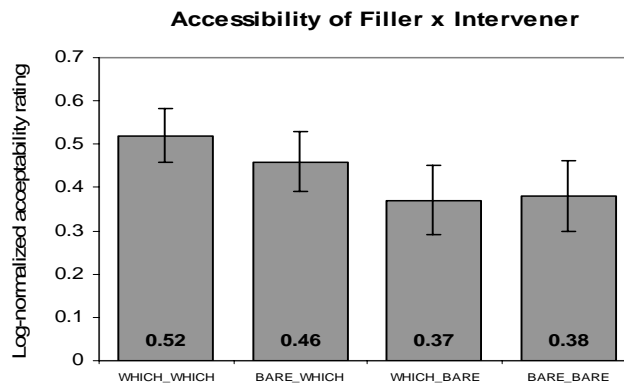


Figure 2 - Acceptability ratings of SUVs in ME2 with 95% confidence intervals shown.

We also observed a main effect of filler accessibility ($F(1,37) = 19.2$, $F(1,19) = 15.7$, $P_s < .001$). This effect is due to an interaction ($F(1,37) = 9.9$, $F(1,19) = 9.8$, $P_s < 0.01$): for *which*-interveners (9c,d), less accessible fillers reduce acceptability, but for bare *wh*-interveners, we found no effect of filler accessibility. That is, the accessibility of the filler had an effect when the in-situ *wh*-phrase was a *which*-NP, but not when the in-situ *wh*-phrase was *who* (as represented by the two rightmost bars of Figure 2).

According to the results, therefore, the effect of interveners actually outweighs the effect of fillers. Having a bare *wh*-intervener caused even the putatively "D-linked" examples like (9b) to be judged as badly as constructions with a bare filler and intervener. The prediction, therefore, that more accessible fillers always improve acceptability was not independently verified in this experiment.

3.4. Effects of Filler Accessibility on Acceptability (ME3)

3.4.1. Materials

The lack of an effect for filler accessibility in the presence of bare *wh*-interveners may seem surprising. ME3 addresses the possibility that the apparent lack of an effect for filler accessibility in the presence of a bare *wh*-intervener may be a spurious null result. The materials for this experiment consequently only varied the type of *wh*-filler (the intervener was always the bare *wh*-item *who*). ME3 also includes one more type of *wh*-expression, *what*-NPs, in order to test whether "complex" *wh*-phrases in general count as high future accessibility markers:

- (10) a. Tom revealed what who invented.
 b. Tom revealed what device who invented.
 c. Tom revealed which device who invented.

We did not entertain any predictions about how sentences like (10b) and (10c) should be judged with respect to each other, treating both simply as roughly equally more informative and syntactically more complex than the bare *wh*-word and hence as increasing future accessibility. ME3 also included non-SUV orders, resulting in 3 x 2 conditions. 18 experimental

items were mixed with 52 fillers, of which 36 were items from ME1. 42 native English speakers participated in ME3 without any compensation. Only the results for the SUV condition are relevant here.

3.4.2. Results

As per Prediction II, there was an effect of filler accessibility: compared to bare *what*-fillers, both *which*-NP and *what*-NP fillers were preferred ($F(1,43) = 12.546, p < .001, F(1, 17) = 5.235, p < .05$). As can be seen in Table 2, grouping *which*-NPs and *what*-NPs is justified. Post-hoc pairwise comparisons revealed that the acceptability of *which*-NP and *what*-NP fillers in SUVs did not differ from each other (subject and item t s $< 0.6, P$ s > 0.5). The pairwise comparisons of both *what*-NPs and *which*-NPs to bare *what*-fillers reached significance by subjects, but not quite by items.⁸

In contrast to ME2, we do see an effect of filler accessibility. Note that the stimuli in ME2 and ME3 are both binary embedded *wh*-questions. In light of this, we tentatively conclude that the lack of a filler accessibility effect for bare interveners in ME2 is a spurious null result.

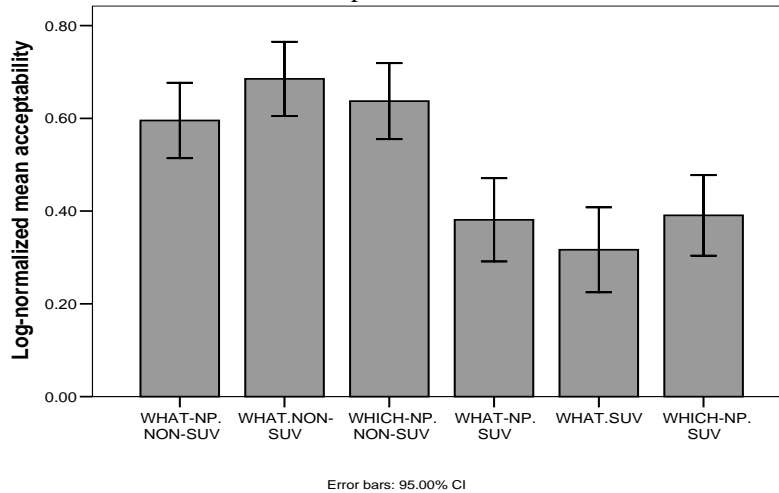


Figure 3 - Acceptability ratings of SUVs and non-SUVs in ME3 (WHAT = bare *wh*-item; WHAT-NP = *what*-phrase; WHICH-NP = *which*-phrase)

SUV condition (F1: subjects)	MEAN Z-SCORE	t	df	Sig. (2-tailed)
<i>what</i> -NP vs. <i>which</i> -NP	.00312	.051	40	.960
bare <i>what</i> vs. <i>which</i> NP	-.19764	-3.292	40	.002
bare <i>what</i> vs. <i>what</i> -NP	.20075	3.195	40	.003
SUV condition (F2: items)				
<i>what</i> -NP vs. <i>which</i> -NP	-.03791	-.563	17	.581
bare <i>what</i> vs. <i>which</i> -NP	-.20969	-1.752	17	.098
bare <i>what</i> vs. <i>what</i> -NP	.17178	1.604	17	.127

Table 2. Pairwise comparisons by subjects and items

Interestingly, we also find that *which*-phrases are not unique markers of high future accessibility. The equally explicit *what*-NP fillers did not induce significantly different judgments of acceptability. Compared to bare *wh*-fillers, *what*-NPs and *which*-NPs both have a greater degree of morphological complexity and explicitness (i.e. more information). This greater degree of explicitness leads to higher future activation, expediting linguistic operations which require retrieval and use of that information. The results thus support a view that demarcates multi-word, complex *wh*-items from less informative, bare *wh*-items in terms of processing difficulty.

3.5. Accessibility Effects on Comprehension Complexity

3.5.1. Materials

So far, we have worked under the assumption that current processing theories make correct predictions about comprehension complexity in *wh*-questions. The *Wh*-Processing Hypothesis in (7) allows for the possibility that differences in the acceptability of *wh*-orders are due to differences in the associated processing complexity. In order to test this assumption about processing complexity, we ran two self-paced, moving window reading time studies (SPR). In SPRs, participants read a sentence word-by-word at their own speed. To ensure proper comprehension, each experimental stimulus is followed by a true-false question about the participants or events described. Before the main experiment, a short list of practice items was presented to the participant in order to familiarize the participant with the task.

- (11) Ashley disclosed {what/which agreement}{who/which diplomat} signed after receiving permission from the president.

The stimuli were adaptations of those used in ME2—embedded SUVs in a 2 (filler accessibility) x 2 (intervener accessibility) design (with slight modifications, i.e. adding post-verbal PPs to control for reading time spill-over effects). Like ME2, 20 experimental items were included in the experiment. 41 subjects participated in this experiment that was conducted at MIT's TedLab, in conjunction with a separate, unrelated reading time experiment. Subjects were paid \$10 per hour for their participation.

The form of the *wh*-filler and *wh*-intervener were expected to affect reading times at the embedded verb (*signed* in (11) above). More specifically, we anticipated that the verb would be read fastest in the condition with the high accessibility filler and intervener (both *which*-NPs). Conversely, the slowest reading times were expected for the condition with the low accessibility filler (*what*) and intervener (*who*).

3.5.2. Results

As predicted, less accessible fillers result in slower processing at the verb ($F(1,40) = 17.7, p < .001, F(1,19) = 12.3, p < .003$), as do less accessible interveners ($F(1,40) = 10.5, F(1,19) = 11.5, P_s < .01$). This replicates the main effects found in ME2 and ME3. Unlike the case in ME2, there was no significant interaction between filler and intervener accessibility.

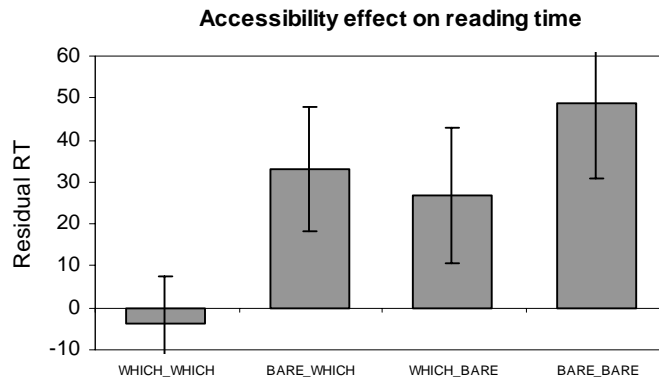


Figure 4 - Residual RTs with 95% confidence intervals, indicating type of superiority-violating object phrase (*which*-NP vs. *who*) and in-situ *wh*-phrase (*which*-NP vs. *who*).

Notice also that we find a difference between the two conditions that have a bare *wh*-intervener (the two rightmost columns in Figure 2), which we did not find in ME2. As in ME3, the more complex and informative *which*-fillers were appreciably better than the bare *wh*-items.

Interestingly, question-answer accuracy is also affected by accessibility (Figure 3). The results seem to mirror the results of ME2. First, question-answer accuracy was significantly lower for bare *wh*-interveners (83%) than for *which*-interveners (92.5%) ($F(1,40) = 18.6, p < 0.001$; $F(1,19) = 7.6, p < 0.02$). We found no main effect for filler-accessibility on answer accuracy, but we found an interaction between intervener and filler accessibility (marginal by subject, $F(1,40) = 3.6, p < 0.07$; significant by item, $F(1,19) = 5.6, p < 0.03$). For *wh*-questions with bare *wh*-interveners, filler accessibility does not affect accuracy. If the intervener is a *which*-phrase, however, high accessibility *which*-fillers result in better question-answer accuracy (95%, $SE = 2.5$) than low accessibility bare *wh*-fillers (89.9%, $SE = 3.1$). Again, this pattern replicates the acceptability results from ME2.

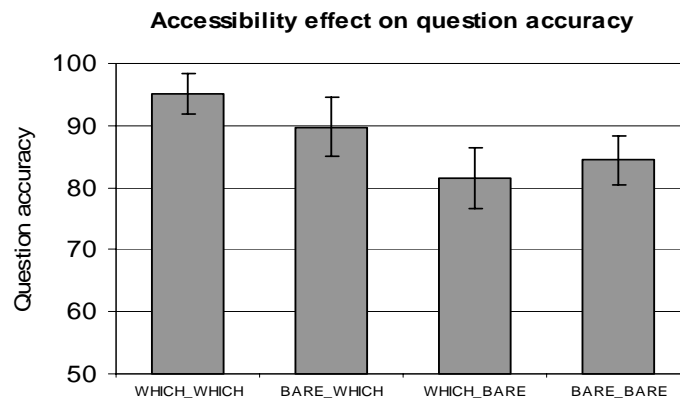


Figure 5 - Accessibility effects on question-answer accuracy with 95% confidence intervals.

4. Discussion

Cumulatively, the results described above demonstrate that configurations of multiple *wh*-phrases display gradient acceptability, affected by locality and the accessibility of the filler and intervener. In SUVs, *which*-NP fillers

improve acceptability judgments and reading times, as compared to bare *wh*-item fillers. Moreover, intervener accessibility impacts the processing of *wh*-dependencies as much as, or even more than filler accessibility: in-situ bare *wh*-items in SUVs decrease acceptability ratings and increase reading times at the verb. A similar dispreference for in-situ bare *wh*-subjects in multiple *wh*-questions has also been found for German (Featherston 2005). We conclude that the *Wh*-Processing Hypothesis can account for a considerable amount of *wh*-order variation using processing-based factors that have been independently introduced to explain other phenomena in sentence processing (e.g. locality- and accessibility-based effects).

One possible interpretation of these results is that mental grammars contain only minimal constraints licensing filler-gap dependencies, without complicated constraints specifying how fillers and gaps can be arranged. Instead, independently motivated processing constraints account for the space of judgments. Perhaps the most attractive aspect of this analysis is that it requires no ad hoc constraints to explain the observed variation. The earliest formulations of Superiority, as well as Pesetsky's D-linking proposal, lack any generality beyond the sphere of multiple *wh*-phrases. This research also bears important implications for other types of *wh*-dependencies that have been labeled as ungrammatical. Indeed, Kluender (1998), while discussing various syntactic islands, suggests that the processing cost of holding a filler in memory and additional referential processing "can interact to yield traditional grammaticality effects." The proposal made here adds support to this idea and identifies *wh*-accessibility as a factor that affects language users' ability to hold a filler in memory.

An interesting challenge for extreme versions of the *Wh*-Processing Hypothesis that attribute *all* variation in the acceptability of *wh*-orders to processing comes from cross-linguistic differences in *wh*-phrase ordering. We refer the reader to Arnon et al. (2006), where we address this challenge. We argue that, even under the assumption of universally processing strategies, the *Wh*-Processing Hypothesis is not only compatible with cross-linguistic differences, but also make predictions as to when they occur.

Our account *wh*-ordering is no doubt incomplete. Other relevant factors may include lexical frequency and collocation effects, as well as plausibility or the supportiveness of the context. We saw a considerable amount of item variability in our acceptability surveys, which is responsible for the lack of significance in some cases. This may be partly attributed to how strongly the embedding verb predicts an indirect question, but also to the degree of affinity between the embedded verb and its *wh*-phrase arguments.

As the multiple *wh*-questions we report on here were all presented without preceding context, the participants were likely faced with the task of imagining a proper context for the question (Fedorenko & Gibson (2006) provide corroboration of our results for English, though, with supporting contexts). In some cases, this may have been particularly challenging and affected the results. Multiple *wh*-questions seem in general suited to only a very particular kind of discourse setting and pragmatic purpose, and when the specific lexical choices cannot be easily reconciled with this purpose, additional difficulty may arise.

5. Conclusion

In this paper, we have identified two major factors that influence *wh*-ordering and the acceptability and processing of *wh*-dependencies: accessibility and locality. We have also examined a noticeable difference between the properties of *wh*-interveners and referential interveners. Our account explains this difference by proposing that explicitness more strongly predicts future activation levels for *wh*-phrases than it does for anaphoric NPs. Accessibility and locality not only explain the effects observed here, but also motivate them. This is in sharp contrast to the widely held views that a competence grammar must include a constraint like Superiority or Chomsky's Attract Closest principle, which seem to be both theoretically undesirable and empirically unnecessary.

Notes

1. This paper has benefited from the comments and input of numerous people including Tom Wasow, Joan Bresnan, Anubha Kothari, Perry Rosenstein and the participants at the 2006 Linguistic Evidence conference in Tübingen. We are also extremely grateful to Ted Gibson and Ev Fedorenko for sharing their knowledge and the resources of TedLab with us, as well as their invaluable expertise in running reading time studies. Any errors are our own.
2. This dichotomy between anaphoric and non-anaphoric NPs predicts that indefinites and definites used to introduce discourse referents should be easier to process as explicitness increases. We are, however, unaware of any results that reflect this preference. Data on the processing of definites from relevant experiments (e.g. Warren & Gibson 2005) only considers definites which require an anaphoric interpretation.

3. To be clear, Warren & Gibson were not looking for such an effect of activation enhancement. The results they present ultimately cannot say that much about the subject because the NP types they use are not forced to have the same interpretation, viz. *we*, *Dan*, and *the businessman* can each be interpreted differently. A true test of enhancement differences would require contextually situated examples with various NP types that can all be linked to the same referent.
4. Conceivably, some other difference between *wh*-phrases and referential NPs could explain the contrasting influences of explicitness. For instance, activation boosts may be stronger and therefore more predictive for *wh*-phrases than referential NPs. Explicitness thus would benefit *wh*-phrase processing more than referential NP processing. Our best hypothesis for this difference, however, is the functional disparity between the two kinds of NPs.
5. Processing difficulty cannot be ascertained merely by acceptability judgments. Fanselow and Frisch (2004) indeed point out that "processing difficulty (understood as including the need to revise an initial analysis) can thus have both positive and negative influences on acceptability." We proceed with the hypothesis that increased processing difficulty reduces acceptability in the case at hand, given findings that support this relationship for other *wh*-island phenomena (Kluender & Kutas 1993). In light of the possible criticisms of acceptability judgments, though, we corroborate the findings with more online data from reading time studies, which provide a more direct measure of processing difficulty.
6. A z-score is a standardization derived by subtracting the sample mean from the individual score and dividing the result by the sample standard deviation.
7. Residual reading times describe differences between the actual reading time and the expected reading time, given the word length (in characters). They are derived using linear regression and are standard in research on sentence processing.
8. The disparity between the results of the omnibus F-test and t-tests derives from a decrease of power in the t-tests.

References

- Alexopoulou, Theodora and Frank Keller
2003 Linguistic complexity, locality, and resumption. Proceedings of WCCFL 22. Somerville, MA: Cascadilla Press.
- Ariel, Mira
1990 *Assessing noun-phrase antecedents*. London: Routledge.

- Ariel, Mira
2001 Accessibility theory: an overview. In Sanders, T., Schilperoord, J., Spooren, W. (eds), *Text Representation: Linguistic and psycholinguistic aspects*. Amsterdam: John Benjamins.
- Arnon, Inbal, Bruno Estigarribia, Philip Hofmeister, T. Florian Jaeger, Jeanette Pettibone, Ivan A. Sag, & Neal Snider
2005 Long-distance dependencies without island constraints. Poster presented at HOWL 3: Hopkins Workshop on Language.
- Arnon, Inbal, Neal Snider, Philip Hofmeister, T. Florian Jaeger, & Ivan A. Sag
2006 Processing accounts for gradience in acceptability: the case of multiple *wh*-questions. Proceedings of *BLS 26*, Univ. of California-Berkeley.
- Bard, Ellen, Dan Robertson, & Antonella Sorace
1996 Magnitude estimation of linguistic acceptability. *Language*, **72.1**: 32-68.
- Chomsky, Noam
1973 Conditions on transformations. In S. Anderson & P. Kiparsky (Ed.) *A Festschrift for Morris Halle*. New York: Holt, Rinehart & Winston.
- Clifton, Charles, Gisbert Fanselow & Lyn Frazier
2006 Amnestying superiority violations: processing multiple questions. *Linguistic Inquiry*, **37.1**: 51-68.
- Fanselow, Gisbert & Stefan Frisch
2004 Effects of processing difficulty on judgments of acceptability. In G. Fanselow, C. Fery, M. Schlesewsky & R. Vogel (eds.) *Gradience in Grammar*. Oxford: Oxford University Press.
- Featherston, Sam
2005 Universals and grammaticality: *wh*-constraints in German and English. *Linguistics*, **43.4**: 667-711.
- Frazier, Lynn & Charles Clifton
2002 Processing 'D-linked' phrases. *Journal of Psycholinguistic Research*, **31.6**: 633-659.
- Fedorenko, Evelina & Edward Gibson

- 2006 Syntactic parallelism as an account of cross-linguistic superiority effects. Unpublished ms, MIT.
- Gernsbacher, Morton
1989 Mechanisms that improve referential access. *Cognition*, 32:99-156.
- Gibson, Edward
2000 The dependency locality theory: A distance-based theory of linguistic complexity. In Y. Miyashita, A. Marantz, & W. O'Neil (eds.), *Image, language, brain*. Cambridge, MA: MIT Press.
- Hawkins, John
2005 *Efficiency and complexity in grammars*. Oxford: Oxford University Press.
- Karttunen, Lauri
1977 Syntax and semantics of questions. *Linguistics & Philosophy*, 1:3-44.
- Keller, Frank, Martin Corley, Steffan Corley, Lars Konieczny & Amalia Todirascu
1998 Web-Exp: A Java toolbox for web-based psychological experiments (Technical report No. HCRC/TR 99). Univ. of Edinburgh. Human Communication Research Center.
- King, Jonathan & Marcel A. Just
1991 Individual differences in syntactic processing: the role of working memory. *Journal of Memory and Language*, 30:580-602.
- Kluender, Robert
1998 On the distinction between strong and weak islands: a processing perspective. In P. Culicover and L. McNally (eds.), *Syntax and Semantics 29: The Limits of Syntax*. San Diego, CA: Academic Press, 241-279.
- Kluender, Robert & Marta Kutas
1993 Bridging the gap: Evidence from ERPs on the processing of unbounded dependencies. *Journal of Cognitive Neuroscience*, 5: 196-214.
- Kuno, Susumu & Jane Robinson
1972 Multiple *wh*-questions. *Linguistic Inquiry*, 3:463-87.
- Pesetsky, David

- 1987 *Wh*-in-situ: Movement and unselective binding. In E. Reuland and A. ter Meulen (eds.), *The Representation of (In)Definiteness*. Cambridge, MA: MIT Press.
- Pesetsky, David
2000 *Phrasal Movement and Its Kin*. Cambridge, MA: MIT Press.
- Warren, Tessa & Edward Gibson
2002 The influence of referential processing on sentence complexity. *Cognition*, 85: 9-112.
- Warren, Tessa & Gibson, Edward
2005 Effects of NP type in reading cleft sentences in English. *Language and Cognitive Processes*, 20.6: 751-767.
- Wasow, Thomas
1997 Remarks on grammatical weight. *Language Variation and Change*, 9: 81-105.