Locality and Accessibility in *Wh*-questions

P. Hofmeister, T.F. Jaeger, I. Arnon, I.A. Sag, and N. Snider
Stanford University
{philiph, tiflo, inbalar, sag, snider}@stanford.edu

1. Competing *Wh*-Orders

There is growing evidence (Pesetsky, 2000; Featherston, 2005) that the ordering of multiple *wh*-phrases within a sentence is conditioned by non-syntactic factors, unlike early proposals along the lines of Kuno and Robinson (1972), which appeal to the grammatical notion of Superiority to account for the contrast between examples like (1a) and (1b):

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

Such grammatical accounts, though, cannot explain the gradience of the data or the existence of naturally occurring examples found in corpora. We propose the ***WH*-Processing Hypothesis** to account for the relative rareness of examples like (1b), as compared to "non-superiority violating" orders like (1a): given the choice between several grammatical *wh*-orders ((e.g. (1a) vs. (1b)), speakers disprefer those which (given the context) are associated with a greater processing cost. Combined with existing theories of processing complexity (for the current purpose, Gibson, 2000), the *WH*-Processing Hypothesis makes the following predictions, which we discuss below:

I. Gaps that are further from the filler are harder to process
II. Less accessible fillers are harder to process
III. Less accessible interveners are harder to process

Evidence from both corpora and relative acceptability judgments indicate that constraints on *wh*-order are non-categorical in nature. Arnon et al. (2005) demonstrate that examples which violate the Superiority condition occur naturally in corpus data. Furthermore, experimentally elicited acceptability judgments do not provide support for a categorical contrast: sentences like (1b) were judged less acceptable than ones like (1a) but more acceptable than clearly ungrammatical sentences (e.g. *who what said*).

2. Experimental Evidence

We present three surveys eliciting acceptability judgments and one experiment measuring comprehension complexity in wh-questions. The purpose of these experiments was to investigate the extent to which variance in the acceptability of different *wh*-orders (and especially Superiority effects) can be accounted for in terms of comprehension complexity. Acceptability judgments were elicited over the WWW using magnitude estimation (ME; Bard et al., 1996) with the WebExp software (Keller et al., 1998). This method lets participants set their own continuous scale of acceptability. Naturalness judgments are made relative to a reference sentence. Comprehension complexity was assessed using a self-paced moving window reading time paradigm (RE). In all experiments, each participant saw each item in exactly one condition (Latin-square design). Table 1 summarizes participant and item numbers:

Table 1 Participant and Item Summary for all Experiments

	ME1	**ME2**	**ME3**	**RE1**
N1 Participants (excluded)	41 (1)	42	42	41
N2 Items	36	20	36	20

1.1 Locality Effects on Acceptability (ME1)

ME1 investigates the effect of locality on the acceptability of *wh*-questions (Prediction I). Locality-based processing theories (Gibson, 2000) predict that *wh*-dependencies are harder to process the greater the distance between a filler and its head (measured in new discourse referents). We manipulated this distance by optionally attaching a six word PP either to the *which*-phrase (2c,f) or to the other NP (2b,e). Furthermore, the *which*-phrase was either subject-extracted (2a-c) or object-extracted (2d-f).

(2) 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?*

Object extractions (which have more intervening discourse referents) are judged as less acceptable than subject extractions ($F1(1,35) = 4.9$, $p < .05$; non-significant by items, $F2(1,35) = 2.5$, $p = .12$). Further comparisons revealed that locality was a good but not perfect predictor of the observed variation in acceptability. This provides partial support for Prediction I (further experiments are in preparation). In short, locality affects the acceptability of *wh*-questions even in the absence of an SUV.

1.2 Accessibility Effects on Acceptability (ME2)

ME2 examines the influence of accessibility on the acceptability of binary wh-questions (prediction II and III). We assume that *which*-NPs are higher accesssibility markers than bare *wh*-words, in the sense that their content is more salient (e.g. *which*-phrases make better antecedents than bare wh-phrases; Clifton & Frazier, 2002). Accessibility-sensitive theories predict that dependencies are easier to process if interveners are high in accessibility (Gibson, 2000). 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 (3):

(3) Mary wondered what/which book who/which boy read.

As predicted, less accessible interveners (the in-situ *wh*-phrase) decrease acceptability significantly ($F1(1,37) = 64.5$, $F2(1,19) = 248.1$, $Ps < .001$). We also observed a main effect of filler accessibility ($F1(1,37) = 19.2$, $F2(1,19) = 15.7$, $Ps < .001$), but this effect is due to an interaction ($F1(1,37) = 9.9$, $F2(1,19) = 9.8$, $Ps < 0.01$): for *which*-interveners (3b,d), less accessible fillers reduce acceptability, but for bare *wh*-interveners, ME2 did not reveal any effect of filler accessibility. The intervener effect was confirmed in another ME experiment ($N1 = 23$, $N2 = 36$) which presented wh-questions in short contexts (unlike in the experiments presented). In sum, sentences with bare *wh*-interveners were considered worse than those with *which*-interveners.

1.3 Effects of Filler Accessibility on Acceptability (ME3)

Given the focus on the filler's (discourse) status in previous syntactic literature (e.g. Pesetsky, 2000), the lack of an effect for filler accessibility in the presence of bare *wh*-interveners may be surprising. ME3 addresses the possibility of a spurious null-result. ME3 also includes one more contextually-linked (high accessibility) type of *wh*-expression—*what*-NPs.. We manipulated *wh*-phrase ordering (SUV (4a) vs. non-SUV (4b)) and the accessibility of the object *wh*-phrase (*which/what*-NP/bare *what*):

(4) a. *Ted indicated what/what law/which law who broke.*
b. *Ted indicated who broke what/what law/which law.*

Here we focus on the results for SUV cases (4a). As predicted (Prediction II), there was an effect of filler accessibility: both *which*-NP and *what*-NP fillers were preferred to bare *what* fillers (pairwise comparisons yielded the following statistics: highly significant by subject $ts > 3.0$; marginal by items $ts > 1.6$). The acceptability of *which*- NP and *what*-NP fillers in SUVs, though, did not differ from each other (subject and item $ts < 0.6$, $Ps > 0.5$). Consistent with Prediction I, SUVs (4a) are judged worse than non-SUVs (4b), but they are still judged a lot better than ungrammatical sentences.

1.4 Accessibility Effects on Comprehension Complexity (RE1)

So far, we have worked on the assumption that current processing theories make correct predictions about comprehension complexity in *wh*-questions. The *Wh*-Processing Hypothesis states that differences in the acceptability of *wh*-orders are due to differences in the associated processing complexity. We ran two self-paced, moving window reading time studies (RE) to test our assumption about processing complexity. In REs, participants read a sentence word-by-word at their own speed. To ensure proper comprehension, each experimental stimulus is followed by a question about the participants or events described. Here we limit ourselves to the description of one RE. RE1 was run to investigate accessibility effects on processing complexity. Stimuli were taken from ME2 (with minimal revisions). The form of the *wh*-filler and *wh*-intervener (as indicators of accessibility) were expected to significantly influence reading times of the embedded verb (*read* in (3) above). Higher accessibility fillers and interveners should decrease reading times.

As predicted, less accessible fillers result in slower processing at the verb ($F1(1,40) = 17.7$, $p < .001$, $F2(1,19) = 12.3$, $p < .003$), as do less accessible interveners ($F1(1,40) = 10.5$, $F2(1,19) = 11.5$, $Ps < .01$). These results provide further support for Prediction II and III and for the *Wh*-Processing Hypothesis in general. Interestingly, question-answer accuracy is also affected by accessibility. 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%) ($F1(1,40) = 18.6$, $p < 0.001$; $F2(1,19) = 7.6$, $p < 0.02$). We found no main effect for filler-accessibility on question accuracy, but we found an interaction between intervener and filler accessibility (marginal by subject, $F1(1,40) = 3.6$, $p < 0.07$; significant by item, $F2(1,19) = 5.6$, $p < 0.03$). For *wh*-questions with bare *wh*-interveners, filler accessibility does not affect question 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).

3. Discussion and Conclusion

The results of all the experiments provide support for the influence of the three proposed processing factors. They demonstrate that configurations of multiple *wh*-phrases display gradient acceptability, affected by factors such as locality and the accessibility of the filler and intervener. In SUV contexts, *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 that are interveners in SUVs decrease both acceptability ratings and 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).

The results presented here are compatible with the English data reported in Fedorenko & Gibson (submitted), which details a similar gradience in the acceptability of SUVs. That study, though, also reports that Russian speakers do not judge SUV and non-SUV orders differently. Note that, even under the assumption of universally applicable processing strategies, the Wh-Processing Hypothesis is compatible with cross-linguistic differences as long as they can be attributed to differences in when and what information becomes available during incremental processing. Nevertheless, cross-linguistic differences pose an interesting challenge for accounts that attribute all variation in the acceptability of wh-orders to processing (i.e. an extreme form of the Wh-Processing Hypothesis). Possible sources for such differences will be addressed in the talk (e.g. the availability of case-marking; word order differences; the amount of case syncretism). For now, we close with the observation that all our experiments find that English SUVs are judged significantly better than other island violations or clearly ungrammatical structures (see also Featherston, 2005). Fedorenko & Gibson (submitted) even find that double center-embeddings, structures generally considered complex but grammatical are judged to be much less acceptable than English SUVs. This fact seems incompatible with accounts that treat SUVs as simply ungrammatical.

References

Arnon, I., B. Estigarribia, P. Hofmeister, T. Jaeger, J. Pettibone, I. Sag, and N. Snider (2005). Long-distance dependencies without island constraints. Poster presented at HOWL 2005.

Bard, E., D. Robertson, and A. Sorace (1996). Magnitude estimation of linguistic acceptability. Language, **72.1**: 32-68.

Frazier, L. and C. Clifton (2002). Processing ‘d-linked’ phrases. Journal of Psycholinguistic Research, **31.6**: 633-659.

Fedorenko, E. and E. Gibson (submitted). The asymmetry in superiority violations in English and Russian: evidence against a processing account.

Gibson, E. (2000). The dependency locality theory: A distance-based theory of linguistic complexity. In Miyashita, Y., Marantz, A., & O’Neil, W., eds., Image, language, brain. MIT Press, Cambridge, MA, pp. 95-126.

Keller, F., Corley, M., Corley, S., Konieczny, L., & Todirascu, A. (1998). Web-Exp: A Java toolbox for web-based psychological experiments (Technical report No. HCRC/TR 99). Univ. of Edinburgh. Human Communication Research Center.

Kuno, S. and J. Robinson (1972). Multiple wh-questions. Linguistic Inquiry, **3**:463-87.

Pesetsky, D. (2000). Phrasal Movement and Its Kin. MIT Press, Cambridge, MA.