FANTASY ISLANDS? TESTING SEMANTIC CONSTRAINTS ON *WH*-EXTRACTION [*]

Dennis Ryan Storoshenko
University of Calgary

Overt *wh*-extraction in English is subject to a series of constraints which have been well-defined in the literature for several decades (Ross 1967). While the terminology and details of the analysis have evolved over the years (from 'islands' to 'subjacency' and most recently to 'phases'), one core aspect of the analysis has remained constant: these constraints are describable in terms of the syntactic structure of the given example. In this paper, I report on a study testing the validity of a set of *wh*-extraction contrasts which defy an easy structural account, and have been used to argue for a set of semantic constraints on extraction which operate alongside structural syntactic ones. The results show that there is indeed evidence for the existence of semantic constraints on *wh*-extraction, but this evidence only fully emerges in a task where specific contexts are controlled, making it possible to define a taxonomy of such constraints.

1. Constraining movement

The familiar constraints on *wh*-movement, such as the inability to extract out of a relative clause or a complement clause to a DP, exemplified in (1), have familiar structural accounts within the literature.

(1) a. * What$_i$ did you see the man who was wearing t$_i$?
b. * What$_i$ did you hear rumours that Benny had cooked t$_i$?

While the details are slightly different in each of these cases, the unifying idea within the GB/Minimalism mainstream is that the proposed movement is ill-formed in that it violates principles governing cyclic movement. In the case of the relative clause, the most immediate landing site for the moving *what* is occupied, while in the complement clause case, the account rests on the DP failing to provide the necessary landing site for cyclic movement. While technical descriptions evolve, the grounding of these constraints in the syntactic structure remains consistent.

Less easy to account for are cases reported in Truswell (2007):

(2) a. What did Will arrive whistling t$_i$?

[*] Thanks to the CLA reviewers for their comments on this work, as well as all those who attended the poster session to give feedback. Thanks also to Robert Frank for many discussions of the examples in this paper, including pointing me in the direction of the Cattell literature, and on the potential for semantic constraints on syntax. This project was funded by a University of Calgary Faculty Start-Up Grant.

Actes du congrès annuel de l'Association canadienne de linguistique 2016.
Proceedings of the 2016 annual conference of the Canadian Linguistic Association.

b. * What did Will run whistling t$_i$?

Truswell's account is based on the telicity of the predicates, noting that the extraction is licit with a telic predicate such as *arrive*, but not the atelic *run*. He formalizes this through a constraint based in a neo-Davidsonian semantics, stating that the extraction is permitted "only if the event denoted by the secondary [adjunct] predicate is identified with an event position in the matrix predicate" (Truswell 2007: 1359). While space does not permit a full re-casting of Truswell's event semantics and its relationship to telicity, the generalization which emerges is that extraction is only licit when the matrix predicate is telic, as in (2), and when the adjunct predicate is atelic. This accounts for the further contrast in (3):

(3) a. What did Joyce drive Jonathan crazy trying to fix t$_i$?
b. * What did Joyce drive Jonathan crazy fixing t$_i$?

This contrast is key in Truswell's argumentation, as he points out that it would be most unusual for a structural account to be ameliorated by adding additional structure into the adjunct clause. While it may be possible to assimilate the simpler contrast in (2) to a structural account of verbal aspect which somehow gives rise to the relevant extraction domain contrasts, the semantic account can easily capture the facts in (2) and (3) in one generalization.

While Truswell is one of the most recent to argue for a position that semantics has a role to play in *wh*-extraction, he is not the first to do so. Cattell (1976, 1979) brings forward examples along the lines in (4):

(4) a. Which car does the mechanic like the gears in t$_i$?
b. * Which car does the mechanic like the kids in t$_i$?

To fully grasp the contrast here, one must keep in mind that the intended answer is *He likes the ones in the Camaro*. This obviates a spurious parse of (4b) in which the mechanic is choosing among cars in which to place a single set of kids; rather, the mechanic's preference among sets of kids in different cars is being questioned. Crucially, there is no clear case to be made for assigning different syntactic structures to the PP adjuncts in (4a) and (4b). Rather, Cattell provides this along with a number of other examples to show that the essential contrast lies in the fact that an essential part-whole relationship obtains between *cars* and *gears*, but not *cars* and *kids*. He contrasts this with cases where altering the verb can induce an extraction contrast by introducing possible structural ambiguity in the PP placement:

(5) a. Who$_i$ did Jim read an article about t$_i$?
b. * Who$_i$ did Jim burn an article about t$_i$?

Extraction is possible where the adjunct can readily be seen to modify the verb, but not when it modifies the nominal complement. This ambiguity underlies the spurious parse of (4b), which is grammatical under a reading where the PP modifies *likes*. However, Cattell notes that, holding the nominal adjunct reading constant, almost any matrix verb will induce the contrast in (4). He further elucidates this notion of a part-whole relationship by describing two different contexts for (6), one of which yields a grammatical judgement, while the other is, in his term, “deviant”:

(6) Which book$_i$ do you like the pictures in t$_i$?
I like the ones in that book.

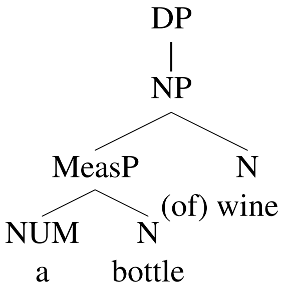
Cattell’s judgement, which I share, is that (6) is felicitous when choosing among a set of art books which all have printed pictures as a core component of the volume. However, the question cannot be used to ask about sets photographs which have each been stored inside of heavy novels (think *Ulysses*, *Crime and Punishment*, *Infinite Jest*) to keep them flat. Where the pictures in question are only incidentally contained within the book, (6) is degraded. While no formal analysis is presented, this case most clearly shows that the relevant is contextual and semantic, not syntactic.

A third contender for an extraction contrast based on semantics derives from a potential ambiguity in container nouns (Robert Frank p.c.):

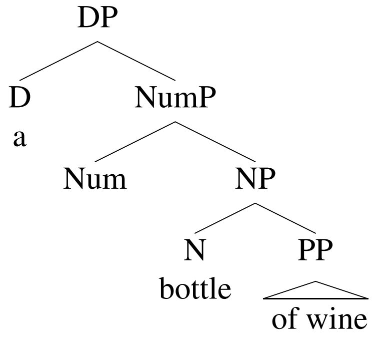
(7) a. What$_i$ did you drink a bottle of t$_i$ after dinner?
b. ? What$_i$ did you break a bottle of t$_i$ after dinner?

The essential semantic contrast here lies in the relationship of the predicate to its internal argument. With *drink*, the theme is the contents of the bottle (e.g. wine), while the entity acted upon with *break* is the physical bottle itself. However, this can also be described in terms of the relationship between the two lexical items in the DP. In (7a), *bottle* is acting as a measure phrase, and could even be felicitously used if 750mL (typical bottle volume) of wine from a cardboard bag-in-a-box had been consumed. However, in (7b), the bottle is a very literal container. Unlike the two cases above, this contrast does have a syntactic contrast, as proposed in Rothstein (2010):

(8) a. Measure



b. Container



However, taking Rothstein's structural analysis to be correct leads to a similar quandary as in (3), as it is not clear how the different structures are relevant to extraction domains. Furthermore, it is easily demonstrated that this type of complement PP as in (8b) allows subextraction when the right relationship obtains between the verb and the DP:

(9) Who$_i$ did you take a picture of t$_i$ (with that camera)?

Thus, the container structure itself is not easily definable as opaque to extraction on structural grounds. This means that if the contrast in (7) holds, it too should be described as semantic, rather than syntactic.

Aside from this common theme of having unlikely structural syntactic accounts, the three cases above, which I will henceforth distinguish as 'Telicity', 'Part-Whole', and 'Container', share another common feature: subtle judgements. The Part-Whole cases are most susceptible to this as there is a spurious parse of the expected unacceptable cases which is acceptable. But, for native speakers all three judgements can be difficult, and disagreements are common. Before theoretic weight can be given to these cases, evidence of the stability of the data reported should be obtained. In the next section, I describe the study which was conducted to seek proof of this stability.

2. Study design

As described above, the primary goal of the study is to test whether the judgement contrasts described above are stable. Recalling though that the overarching idea is that the judgements derive from semantic contrasts, this study also tests that idea by treating context as a variable. In the first case, the hypothesis is that stimuli which meet the conditions for the various extraction constraints should be significantly different from minimal pair variants which do not. For this second question of semantics, the hypothesis is that the observed effects should be stronger when more context is given, enriching the semantics. Further treating each of the three contrasts as variables leads to a 2x2x3 design, as shown in Table 1. Conditions under 'No Context' are simple acceptability judgements examining the *wh*-questions in isolation, while the judgements in context are more accurately described as felicity judgements, as they examine the coherence of the context as a whole.

The details of the stimuli and participants are given in the following sections.

Table 1: Study Design Condition

	No Context		Context	
Telicity	Acceptable	Unacceptable	Felicitous	Infelicitous
Part-Whole	Acceptable	Unacceptable	Felicitous	Infelicitous
Container	Acceptable	Unacceptable	Felicitous	Infelicitous

2.1 Stimuli

For each of the three constraints under examination, six minimal pair question and answer sequences were constructed; each pair differed only by the crucial lexical item which would determine whether the question on its own would be predicted to be acceptable or unacceptable. Sample minimal pairs for each of the three are given below in the context-enriched question-answer form:

(10) a. What did the dog **leave** chewing?
The dog left chewing a bone. Telicity, Felicitous

b. What did the dog **run** chewing?
The dog ran chewing a bone. Telicity, Infelicitous

(11) a. Which class does Nancy like the **students** in?
Nancy likes the ones in the math class. Part-Whole, Felicitous

b. Which class does Nancy like the **chairs** in?
Nancy likes the ones in the math class. Part-Whole, Infelicitous

(12) a. What did Dustin **eat** a bowl of this morning?
Dustin ate a bowl of oatmeal this morning. Container, Felicitous

b. What did Dustin **crack** a bowl of this morning?
Dustin cracked a bowl of oatmeal this morning. Container, Infelicitous

The Telicity cases all conform to the constraint of having atelic predicates in the adjunct clause, while varying the telicity of the matrix. For the Part-Whole cases, the manipulation is in the head noun of the internal argument, varying on whether the extracted *which* phrase forms a natural part-whole relationship with the noun or not. In this case, students are more essential to a class than chairs. Also, to avoid the spurious parse, the answer to the Part-Whole cases reinforces the reading where the PP is a nominal modifier via *one*-anaphora replacing the head noun. This has the effect of making the answers constant across both members of the pair. For the Container cases, the verb is again manipulated on the basis of whether it most readily brings out a measure or a container reading for the following noun.

To avoid any judgements based on a prescriptive bias against ending the question with a preposition, all stimuli in this condition ended with either a temporal or a locative adverbial PP. Overall, 18 minimal pairs were constructed. These were divided into two stimulus lists such that each list contained only one member from each minimal pair. Each of these lists formed the basis for two further lists, two containing questions only (the 'Q' lists), and two containing the question-answer sequences (the 'QA' lists).

In addition to the 18 experimental stimuli, each list also contained 36 filler items. 18 of the filler items were grammatical *wh*-questions of varying complexity:

(13) a. What did Mike believe he saw after dinner?
Mike believed he saw a cougar after dinner.

b. What did the waitress who just started bring to the table first?
The waitress who just started brought coffee to the table first.

c. Which house did the girls all go to?
All the girls went to the blue house on the hill.

The relatively complex structures were designed to mirror the complexity of the experimental stimuli, as well as the more marginal fillers discussed below. Six out of the 18 grammatical fillers were *which* questions, again mirroring the form of the experimental stimuli, a third of which are *which* questions.

The other 18 filler items were more marginal. 12 of these were straightforward island violations similar to those in (1):

(14) a. What did Barb surprise the person who bought?
Barb surprised the person who bought the vase. (Relative Clause)

b. What did the children all understand who won?
The children all understood who won the game. (Embedded *Wh*-Question)

These were included to ensure that some stimuli in the list would be rejected by all participants, controlling for participants making random judgements.

The final six filler items were superiority violations mitigated by D-linking:

(15) Which game did which team need to win?
The home team needed to win Tuesday's game.

These items were also included to ensure that there would be some more marginal items making use of *which* within the whole stimulus set. Furthermore, these were included to further test the effectiveness of the context versus no context variable, as it would be expected that D-linking is context-sensitive. In total, each list contained 54 items.

2.2 Participants and task

Participants were all native speakers of English recruited from the Linguistics Experiment Pool, as well as the broader university community. Pool participants received bonus course credit for their participation[1], while other community members received a cash payment. All data was collected in the Language Research Centre at the University of Calgary. Upon arrival at the lab, each participant was randomly assigned to one of the four experiment lists; this means that the context variable (Q list vs. QA list) was between participants, while the other variables were within participants. 48 participants were recruited, with 12 assigned to each of the four stimulus lists.

Data collection was on a PC running WebExp2 experiment software (Mayo et al. 2005). Stimulus presentation was randomized to avoid any potential priming effects impacting the study as a whole; each participant saw the stimuli in a unique order. Participants in the Q list condition received instructions that their task was simply to rate the questions on a 7-point Likert scale, with 7 being most acceptable and 1 being completely unacceptable. This group was told that some questions could be ambiguous, and to only rate a stimulus as completely unacceptable if there was no possible acceptable reading. Those in the QA group were asked to rate the felicity of the overall exchange, keeping in mind that their rating needed to take into account whether the question as presented could bring out the given answer, not other possible ones they could imagine. Again, 7 was for most felicitous, while 1 was for completely infelicitous. Participants in both groups were encouraged to use the whole scale, with the contrast in instructions designed to maximize the distinctions in the between-group context variable: the Q group was free to construct as much or as little context as desired, while the QA group was given a very specific context to focus on.

Participants in both groups received three initial training trials using examples parallel to the fillers, to get used to the interface, as well as to highlight the types of contrasts which may fall at different points on the scale. While there was some variation in the absolute values, almost all the participants made the expected relative rankings of a relative clause island being quite unacceptable/infelicitous, with the D-linking superiority violation scoring somewhat higher, and the categorically grammatical/felicitous examples always highest. The study was self-paced with no time limit, clicking the appropriate rating button for the currently displayed stimulus would advance to the next.

3. Results and discussion

The mean ratings for each condition are given in Table 2. Based on an initial impression of the numbers, two facts are striking. Firstly, even the trials which are predicted to be grammatical/felicitous are degraded. As a baseline, the mean for all of the grammatical filler trials was 6.51. This result alone may explain some of the reason the judgement of these cases is subtle: the sentence structures are dispreferred. The second initial impression

[1] Ineligible non-native speakers have alternate means of obtaining the same bonus course credit.

is that at least based on the means, the (un)acceptable/(in)felicitous pairings, hereafter (a) vs. (b), come out as predicted, with degradation of judgements moving in the expected direction.

Table 2: Mean Results by Condition

	No Context (Q)		Context (QA)	
	Acceptable (a)	Unacceptable (b)	Felicitous (a)	Infelicitous (b)
Telicity (10)	3.55	2.54	3.99	3.21
Part-Whole (11)	5.54	5.36	5.13	4.69
Container (12)	6.04	5.72	6.16	5.58

To verify these impressions, the data were subjected to analysis using a Generalized Estimating Equation (GEE) in SPSS v.20. First, to test the significance of the different semantic constraints, all results were pooled, treating island type and acceptability/felicity as factors. Both of these are individually significant ($\chi^2(2)$ = 243.32, $p < 0.001$ and $\chi^2(1)$ = 24.61, $p < 0.001$ respectively), but no significant interaction is observed ($\chi^2(2)$ = 4.935, $p = 0.085$). In pairwise comparisons, again pooled across Q and QA groups, the (a) vs. (b) contrast only has a significant effect on the Telicity and Container cases, significant at the $p \leq 0.05$ threshold. These results suggest that the telicity contrast identified by Truswell and the Container contrast proposed here do have an effect on *wh*-extraction which is detectable in the population. However, the lack of an interaction between the factors does suggest that overall, the effect of the lexical manipulations to trigger the various extraction constraints was consistent across the three types.

The reason for the lack of a significant effect within the Part-Whole cases clearly lies in the no context Q group, where there is less than a 0.20 difference in the means between the acceptable and unacceptable conditions. This suggests that participants in the Q group were accessing the spurious parse for these sentences, thereby avoiding the contrast in acceptability. This speculation is supported by a comparison with the numbers for the Part-Whole examples in the QA group, where the difference in between the means is more than doubled.

A second analysis was conducted to address the question of whether the observed effects could be attributed to semantics. Here, the Q vs. QA factor was tested against the different constraint types. Not surprisingly after examining the table of means, a significant interaction was found: all three constraints show different effects when moving from the Q to the QA group. The Container contrast behaves as one might most reasonably expect, with the contrast sharpening as context is added: the (a) cases get slightly better, and the (b) cases get slightly worse. For Part-Whole, the effect of eliminating the spurious parse is very clear: where context forces the PP adjunct to be taken as modifying the nominal complement, the (a) cases get slightly worse, and the contrast between the (a) and the (b) cases sharpens considerably. The most unexpected finding was in the Telicity data, where adding context ameliorated the both the (a) and the (b) cases, narrowing the distinction between the two. It thus seems that the Part-Whole and the Container cases are behaving as a semantic constraint, with added context sharpening the judgement, while the Telicity cases are doing the opposite, even though it seems to be the sharpest of the three contrasts, with the widest gap between the (a) and (b) cases.

Table 3: Mean Results for Marginal Fillers

	No Context (Q)	Context (QA)
Relative Clause (14a)	1.64	1.82
Embedded *wh* (14b)	2.16	2.43
D-Linking (15)	4.97	4.67

A potential explanation for this can be found in an examination of the results for the marginal fillers, shown in Table 3. The two most clearly structural cases show this same effect where providing a cogent answer to the question boosts the rating. This may suggest that in contexts where a structural constraint is violated, a coherent discourse context mitigates the violation as an accommodation strategy. In turn, this may suggest that unlike the Part-Whole and Container cases, the Telicity cases are better treated as structural, motivating a closer analysis of the relationship between aspectual projections in the verbal spine and potential phase heads. Most unexpected in the results for the marginal fillers are the means for the D-Linking cases between conditions. Given that D-Linking is already a phenomenon in which context mitigates a structural constraint on extraction, one might expect that adding context should be an even more mitigating factor. However, there is a crucial difference in that D-Linking relies on prior context; it may be the case that only providing a following context, along with the instructions directing participants to attend only to the given question answer sequence, discouraged participants in the QA group from assuming an appropriate prior discourse to license the extraction. A dedicated study would be required to determine whether this is indeed the case.

4. Conclusion and future work

The research questions of this paper are twofold. First, the study outlined above sought to corroborate the claims in Section 1, namely that the three constraints on *wh*-movement do exist within the population at large, and are not merely reflective of a narrow ideolect spoken only by certain syntacticians. The significant result in the first statistical analysis bears this out, demonstrating that all three effects are indeed real, if subtle. The second research question concerned the identification of these constraints as semantic in nature rather than syntactic. To establish this, the study tested whether or not adding context in the form of an answer would sharpen the contrast. In two of the cases, this did happen. However, the interaction of this variable with the fillers suggests that adding richer context may be a way to tease apart two types of *wh*-extraction constraints, as constraints which have been traditionally (and uncontroversially) described as structural are mitigated in the presence of a cogent answer. This opens up the possibility that the methodology employed here, contrasting the Q with the QA groups, could lead to a taxonomy of constraints on *wh*-extraction. At least in the case of the Part-Whole and Container constraints, the results reported here do suggest that the effect lies in the semantic relationship between the matrix predicate and its internal argument, not something that can be easily distinguished in the syntax. Or, at least, not something that could be easily expressed in the syntax as creating a domain that is opaque to A$'$ movement without incorrectly predicting the ungrammaticality of other cases.

The theoretical implications are clear. If the constraint lies in the compositional semantics, then the derivational timing between semantic composition and overt syntactic movement needs to be re-thought. As an overt move, *wh*-extraction should precede LF, and should therefore be "immune" to semantic considerations where these relationships are calculated. Instead, a parallel architecture closer to that proposed in Culicover and Jackendoff (2005) may be more well-adapted to capture the semantic constraints. In this way, syntax and semantics can be derived in parallel (along with the phonological form), allowing semantic effects to have impact on earlier portions of the syntactic derivation.

Synchronous Tree Adjoining Grammar (STAG) is another formalism which treats syntax and semantics as being derived in parallel. Syntactic trees not unlike the kernel sentences of early Chomskyan work are assembled, and all have associated lambda expressions. The trees are then composed into a larger structure, and each compositional step must simultaneously meet certain syntactic well-formedness constraints and yield a licit combination in the typed lambda calculus (Schabes and Shieber 1994, Nesson and Shieber 2006). Interestingly, within this framework, two different mechanisms for *wh*-extraction have been proposed. Frank (2002) uses a model involving strictly local overt syntactic movement interacting with the tree composition process to derive the familiar structural constraints on *wh*-extraction. However, other researchers have treated *wh*-expressions as having the syntactic and semantic forms of generalized quantifiers, and derive the meanings of *wh*-questions using the exact same STAG mechanism that derives quantifier scope, mimicking quantifier raising (QR). This leads to the incorrect prediction that domains which are opaque to QR are also opaque to *wh*-extraction. However, it is well-known that quantifiers cannot scope out of a finite embedded clause, while cyclic *wh*-movement is possible:

(16) a. Some kid said [(that) every electrician was evil].
$\exists > \forall, *\forall > \exists$

b. What did Lucas say [that the evil electricians were driving t$_i$]?

Addressing this issue in a recent paper, Storoshenko and Frank (2016) argue that there are (at least) two types of *wh*-dependency[2] formation in STAG, and that syntactic and semantic

[2] The term 'dependency' is used in place of 'extraction' in this literature as not all cases are formed via syntactic movement.

constraints on the dependencies can co-exist within a single grammar. The findings of the study reported in this paper support that claim, identifying additional semantic constraints. The potential corollary is that the *wh*-dependencies which are subject to semantic constraints may derive from this more QR-like treatment rather than overt movement. To test this, a follow-up study can be designed to determine whether the environments which have been demonstrated to trigger this semantic constraint on *wh*-extraction are also opaque to QR:

(17) a. A student worked whistling every song.
b. A mechanic liked the kids in every car.
c. A waiter broke a bottle of every wine after dinner.

If it emerges that examples along the lines of (17) are indeed scope islands barring inverse scope, while their extraction-permitting counterparts are not (i.e. swapping *arrived* for *worked* in (17a) and so on), then there will be further evidence for the larger claim that the heterogeneity of *wh*-extraction constraints may derive from a heterogeneity of *wh*-dependency formation. This is because it is already well-established as in (16) that the canonical long-distance movement cases, which receive a different account in STAG, are able to cross scope islands. Reporting on the results of this proposed follow-up is left for future work.

References

Cattell, Ray. 1976. Constraints on movement rules. *Language* 52(1): 18–50.

Cattell, Ray. 1979. On extractability from quasi-NPs. *Linguistic Inquiry* 10(1): 168–172.

Culicover, Peter W. and Ray Jackendoff. 2005. *Simpler syntax*. Oxford: Oxford University Press.

Frank, Robert. 2002. *Phrase structure composition and syntactic dependencies*. Cambridge, MA: MIT Press.

Mayo, Neil, Martin Corley, and Frank Keller. 2005. *WebExp2 experimenter's manual*. School of Informatics, University of Edinburgh.

Nesson, Rebecca and Stuart M. Shieber. 2006. Simpler TAG semantics through synchronization. In *Proceedings of the 11th conference on formal grammar*, ed. Shuly Winter. 129–142.

Ross, John R. 1967. *Constraints on variables in syntax*. Ph.D. thesis, Massachusetts Institute of Technology.

Rothstein, Susan. 2010. Counting, measuring, and the semantics of classifiers. *The Baltic International Yearbook of Cognition, Logic and Communication* 6: 1–42.

Schabes, Yves and Stuart M. Shieber. 1994. An alternative conception of tree-adjoining derivation. *Computational Linguistics* 20(1): 91–124.

Storoshenko, Dennis Ryan and Robert Frank. 2016. Parasitic gaps and the heterogeneity of dependency formation in STAG. In *Proceedings of the 12th international workshop on tree adjoining grammars and related formalisms (TAG+12)*, ed. David Chaing and Alexander Koller. 112–120.

Truswell, Robert. 2007. Extraction from adjuncts and the structure of events. *Lingua* 117: 1355–1377.