

The effects of temporary representations on acceptability

Jon Sprouse
University of Maryland

Given the primary role played by acceptability judgments in the syntactic literature, much work has focused on the factors that affect judgments, from semantic plausibility to processing difficulties. Recent findings suggest that certain processing difficulties such as local ambiguity and object-before-subject fronting (see Fanselow and Frisch 2004 for a review) do indeed affect the global acceptability of sentences. Building on these findings, the experiments presented in this squib use temporary representations constructed and abandoned prior to the final, grammatical¹ representation to probe the nature of acceptability judgments. Experiment 1 uses two well-established self-paced reading paradigms, the *filled-gap effect* (Crain and Fodor 1985, Stowe 1986) and the *plausibility effect* (Garnsey et al 1989, Tanenhaus et al 1989), to test the effects of temporary representations that are syntactically and semantically ungrammatical respectively. The results from experiment 1 indicate that while the temporary syntactic violation at the root of the *filled-gap effect* does affect the judgment of the final (grammatical) representation, the temporary semantic violation at the root of the *plausibility effect* does not. Experiment 2 confirms that this asymmetry is not due to an unrelated difference between the filled-gap paradigm and the plausibility paradigm: namely that both conditions within the plausibility paradigm require abandoning one structure and building another (reanalysis), while only the filled-gap condition of the filled-gap paradigm requires reanalysis. In the end, the pattern of results indicate that judgment tasks are affected by temporary syntactically ungrammatical representations, but not by temporary semantically ungrammatical representations, and not by temporary grammatical-but-incorrect representations. This suggests that judgment tasks actually tap into syntactic knowledge in a qualitatively different way than semantic and parsing knowledge in that syntactic mistakes affect final judgments, whereas semantic and parsing mistakes do not.

1 Experiment 1: Syntactic versus semantic temporary ungrammaticality

Psycholinguistic Background

The Active Filling Strategy

The experiments in this squib build upon one of the major findings of sentence processing research: the *active filling strategy*. The active filling strategy is defined by Frazier and Flores d'Arcais (1989) as “when a filler has been identified, rank the possibility of assigning it to a gap above all other options.” Or in other words, the human parser prefers to complete long distance dependencies as quickly as possible. Because the

¹ For clarity, this paper will adhere to the following terminological convention: *grammatical/ungrammatical* are two possible properties of representations, and *acceptable/unacceptable* are two possible values of judgments.

quickest possible completion site is not always the correct one, the active filling strategy entails the construction of many temporarily incorrect representations. The experiments in this squib take advantage of these temporary representations.

The Filled-Gap Effect

The filled-gap effect in self-paced reading tasks is derived from the active filling strategy. Basically, if a verb is encountered while a wh-dependency is open, and if the verb’s thematic relations have already been filled (in other words, the gap is filled), then there will be a corresponding slow down in reading times at the object of the verb because the active filling strategy attempts to complete the wh-dependency at the verb, even though there is no empty thematic position. Stowe 1986 demonstrated this effect with the following quadruplet:

- (1) My brother wanted to know...
 - (a) if Ruth will bring us home to Mom at Christmas. (IF)
 - (b) who __ will bring us home to Mom at Christmas. (WH-S)
 - (c) who Ruth will bring __ home to Mom at Christmas. (WH-O)
 - (d) who Ruth will bring us home to __ at Christmas. (WH-P)

Stowe found a significant reading time slow-down at the position of the object *us* when there is a displaced wh-filler (WH-P) as compared to when there is no displaced wh-filler (IF):

- (2) Results from Stowe 1986

Dependency	Ruth	us	Mom
IF	661	755	755
WH-S	--	801	812
WH-O	680	--	833
WH-P	689	970	--

The Plausibility Effect

Much like the filled-gap effect, the plausibility effect stems from the active filling strategy of wh-dependencies, but in this case the slow-down in reading time stems from manipulating the plausibility of an association between the wh-filler and the verb. For instance, Pickering and Traxler (2003) found a significant slow-down at the verb when the displaced wh-filler is an implausible object of the verb *killed*:

- (3) Plausible: That’s the **general** that the soldier **killed** enthusiastically for __ during the war in Korea.
- Implausible: That’s the **country** that the soldier **killed** enthusiastically for __ during the war in Korea.

(4) Results from Pickering and Traxler 2003

Filler type	Reading time of <i>killed enthusiastically</i>
Plausible	1045
Implausible	1157

Rationale

Given that the question under investigation is whether temporary representations have an effect on the judgment of the final representation, it goes without saying that it must be established that the temporary representations being manipulated are actually constructed. The *filled-gap effect* (Crain and Fodor 1985, Stowe 1986) and the *plausibility effect* (Garnesey et al. 1989, Tanenhaus et al. 1989) were chosen because their effects are so well-established that they serve as tools for investigating filler-gap dependencies in the processing literature. Furthermore, despite both being reflexes of active filling, the source of the effect for each paradigm is different. The slow-down that occurs in the filled-gap paradigm is due to the failure of successfully integrating the filler with the verb's argument structure, resulting in a temporary structure in which there is an incomplete dependency. On the other hand, the slow-down in the plausibility paradigm is not due to an incomplete dependency, but rather a completed dependency that results in a semantically anomalous interpretation. Thus these two paradigms are ideal for investigating the effects of different types of temporarily illicit representations.

Because all of the target conditions must ultimately be acceptable sentences to avoid any interfering grammaticality effects, three design elements were incorporated into experiment 1 to ensure that any failure to detect either the filled-gap or the plausibility effect was not due to a lack of sensitivity. First, the task chosen was magnitude estimation. Recent findings have indicated that magnitude estimation is well suited for detecting differences among acceptable sentences (e.g. Featherston 2005). Second, a relatively large number of subjects were asked to participate in this study. Recent work has suggested that reliable results can be obtained from samples as small as 10 (Myers 2006), therefore the large sample size for experiment 1 (N=86) should be adequate to detect even very small differences.

Finally, a third condition set was included to determine whether the task and subject pool were sensitive to distinctions among acceptable sentences. The third condition set was taken from an ERP study of the distance between a wh-filler and its gap by Phillips et al (2005). Phillips et al. manipulated the distance by displacing the wh-filler either 1 or 2 clauses away from the gap position:

- (5) Short WH: The detective hoped that the lieutenant knew **which accomplice** the shrewd witness would recognize __ in the lineup.
 Long WH: The lieutenant knew **which accomplice** the detective hoped that the shrewd witness would recognize __ in the lineup.

Phillips et al. found a delay in the onset of the P600, a response that has been linked to the association of a wh-filler with its gap, for the Long WH condition, which they interpret as a reflex of the time it takes to retrieve the stored filler from working memory

(longer distance = longer retrieval time perhaps because of a decaying representation). But crucial to our purposes, Phillips et al. conducted a ratings survey in which they asked the subjects to rate the complexity of the two conditions on a scale from 1 to 5:

(6) Results from Phillips et al. 2005

Wh-Distance	Mean rating (Standard Deviation)
Short – 1 clause	2.71 (0.65)
Long – 2 clauses	3.51 (0.51)

The difference between the two conditions was highly significant ($t(23)=5.83$, $p<.001$), indicating that judgment tasks could indeed detect an effect that leads to unconscious processing effects. Thus this condition was included in experiment 1 as a baseline to test the sensitivity of the task and subject pool (although experiment 1 was an acceptability rating task rather than complexity rating task).

Participants

86 University of Maryland undergraduates participated in experiment 1 for extra credit. All of the participants were self-reported native speakers of English. The survey was 36 items long including practice items, and took about 15 minutes to complete.

Materials

The design of experiment 1 included 3 condition sets: the filled-gap paradigm to test temporary syntactically ungrammatical representation, the plausibility paradigm to test temporary semantically ungrammatical representations, and the wh-distance paradigm to check the sensitivity of the task and subject pool.

For the filled-gap condition set of this experiment, the WH-O and WH-P conditions from Stowe 1986 were reconstructed:

- (7) Gap: My brother wanted to know **who** Ruth will bring __ home to Mom at Christmas.
 Filled-Gap: My brother wanted to know **who** Ruth will bring **us** home to __ at Christmas.

In the filled-gap condition, the failure to integrate the displaced wh-filler with the verb creates a representation in which the dependency is incomplete, which persists until the gap in the prepositional phrase.

The materials for this plausibility condition set were taken directly from the published materials of Pickering and Traxler 2003, although an additional adverb was added to each token to increase the duration of the semantically ungrammatical representation, and the matrix clause was changed to match the style of the filled-gap conditions, i.e. declarative sentences:

- (8) Plausible: John wondered **which general** the soldier killed __ effectively and enthusiastically for __ during the war in Korea.
Implausible: John wondered **which country** the soldier killed __ effectively and enthusiastically for __ during the war in Korea.

The implausible version of (8) creates a dependency at the verb *killed* that is semantically ungrammatical, which persists through the two adverbs until the gap in the prepositional phrase.

The materials for the wh-distance condition set were taken directly from the published materials of Phillips et al. 2005:

- (9) Short WH: The detective hoped that the lieutenant knew **which accomplice** the shrewd witness would recognize __ in the lineup.
Long WH: The lieutenant knew **which accomplice** the detective hoped that the shrewd witness would recognize __ in the lineup.

Design

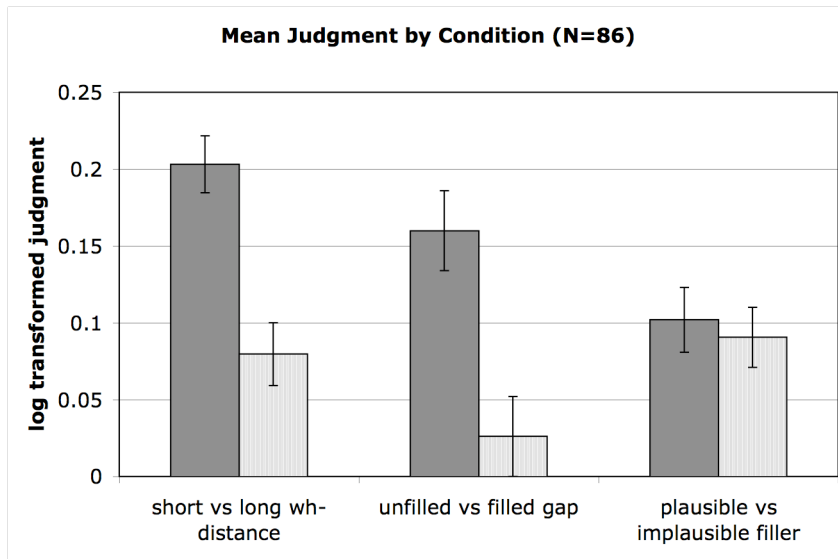
There were 6 total conditions (2 each of 3 sets) under investigation. 24 lexicalizations of the filled-gap conditions were reconstructed following the examples from Stowe 1986. 24 lexicalizations of the wh-distance conditions were taken from the materials of Phillips et al. 2005. Only 12 lexicalizations were available for the plausibility conditions from Pickering and Traxler 2003. A 24-cell Latin Square was constructed such that each list contained 2 tokens of each condition. 14 unacceptable fillers (various syntactic island violations) were added, and each list was pseudo-randomized such that no more than 2 target conditions were consecutive, and no related conditions were consecutive. 8 practice items were added, resulting in a 34 item survey.

The instructions were a modified version of the instructions distributed with the WebExp software suite (reference). The reference sentence for both the practice and experimental items was a three-clause long sentence containing a whether-island violation: *Mary figured out what her mother wondered whether she was hiding.*

Results

The chart below contains the log-transformed mean judgment values for the three condition sets (log transformation is standard practice for magnitude estimation data, see Lodge 1981, Bard et al. 1996). Shading is added as a convenience to indicate the direction of the effect in the psycholinguistic literature, where the effect would be either a reading time slow-down or delayed P600:

(10) Results for Experiment 1



As the chart indicates, there was a large and highly significant decrease in acceptability for longer wh-dependencies ($t(85) = 5.324$, $p < .001$, $r = .5$), and in exactly the same direction as obtained by Phillips et al 2005. There was also a large and highly significant decrease in acceptability for filled-gaps ($t(85) = 5.616$, $p < .001$, $r = .52$), mirroring the direction of the effect found by Stowe 1986. However, there was no effect of plausibility ($t(85) = -.514$, $p = .608$). Even though there are no direct statistical comparisons across the groups, it is clear that both of the significant p values are well under the conservative Bonferroni correction level of .0167.

Discussion

The pattern of results from experiment 1 is extremely puzzling. First, it is clear from the wh-distance effect that the design is capable of detecting differences that lead to processing effects. However, when it comes to the two active filling effects, a significant effect was only found for the filled-gap effect. Also, given the large sample size, it seems unlikely that increasing the sample size will lead to an effect of plausibility.² At first glance, this seems to corroborate the claim in the introduction that temporary syntactic ungrammaticality affects global judgments, whereas temporary semantic ungrammaticality does not. Unfortunately, there is a second possible explanation: reanalysis.

By definition, the filled-gap condition of the filled-gap paradigm involves abandoning one structure and constructing a second structure, a type of syntactic reanalysis: when the association between the wh-filler and the thematically saturated verb fails, the parser must reanalyze the structure such that the wh-filler is then associated with the preposition. In other words, the parser attempts to ‘drop’ the filler twice. However,

² As Colin Phillips (p.c.) points out, this is even more puzzling given that the plausibility effect is generally more robust than the filled-gap effect in reading-time studies.

the true gap condition of the paradigm involves no such reanalysis because the first association with the verb succeeds. It could be the case then that the difference in acceptability between the two conditions is an effect of reanalysis on the judgment. This would also account for the lack of effect in the plausibility conditions: in both conditions, the *wh*-filler is initially associated with the verb and later reanalyzed as the object of the preposition. Thus if reanalysis leads to a decrease in acceptability, one would expect an effect in the filled-gap paradigm but no in the plausibility paradigm. Experiment 2 was designed to tease apart these two hypotheses (asymmetry due to temporary unacceptability versus asymmetry due to reanalysis).

2 Experiment 2: Reanalysis

Rationale

Unfortunately, by definition there is no way to eliminate reanalysis from the filled-gap and plausibility paradigms. However, it is possible to add reanalysis to the true gap condition of the filled-gap paradigm, thus making it completely parallel to the plausibility paradigm in that both conditions will contain reanalysis. If the asymmetry in the presence of reanalysis across the two paradigms was the source of the asymmetry in the results for experiment 1, then eliminating the reanalysis asymmetry should eliminate the asymmetry in the results such that both paradigms return no effect. Experiment 2 was designed to test this hypothesis.

Furthermore, by adding the true-gap condition from experiment 1 that lacks reanalysis and comparing it to the new true-gap + reanalysis condition, it is possible to isolate the effect of reanalysis alone, if it should exist. This comparison investigates the effect of a temporary grammatical representation on the judgment of the final representation, or in other words, the effect of processing difficulty without ungrammaticality, setting up the three-way comparison of temporary representations discussed in section 1.

Materials and Design

The materials for experiment 2 were adapted from the materials for the plausibility conditions in experiment 1, which were themselves adapted from the published materials of Pickering and Traxler 2003. These materials were chosen for two reasons: (i) the plausibility materials already contained the necessary structure to include reanalysis in both the filled-gap and true-gap conditions; and (ii) if a filled-gap effect is indeed found using these materials, it would serve to exclude the possibility that the lack of effect for plausibility in experiment 1 was due to the meanings of the materials.

Three conditions were used to test whether the source of the asymmetry from experiment 1 was the reanalysis asymmetry. First, a filled-gap condition was constructed out of the materials from Pickering and Traxler 2003. Next a true-gap condition was constructed with an additional gap in the prepositional phrase to represent a gap+reanalysis condition. Finally, a standard true-gap condition was constructed with no gap in the prepositional phrase to serve as both a replication of the filled-gap effect in

experiment 1, and to serve as a control for gap+reanalysis condition so that an effect of reanalysis alone could be tested:

(11) Filled-gap + Reanalysis

John wondered which **general** the soldier killed **the enemy** effectively and enthusiastically for __ during the war in Korea.

Gap + Reanalysis

John wondered which **general** the soldier killed __ effectively and enthusiastically for __ during the war in Korea.

Gap

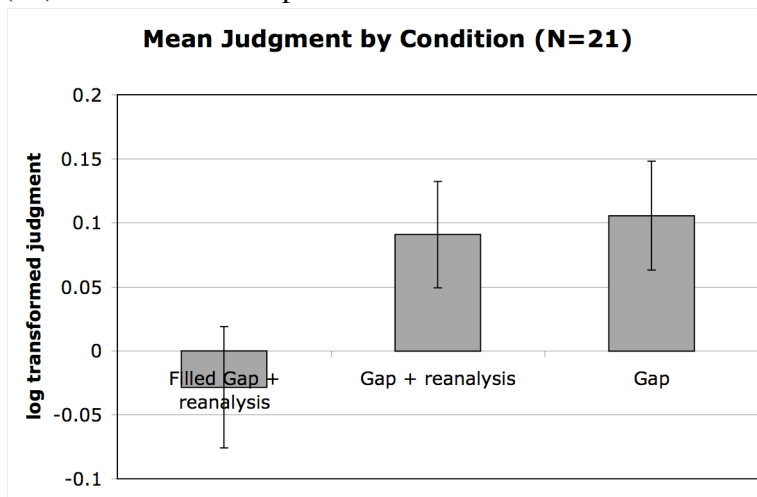
John wondered which **general** the soldier killed __ effectively and enthusiastically for our side during the war in Korea

Again, the competing hypotheses make different predictions: if reanalysis is the source of the asymmetry, then experiment 2 should yield no effect between FG+R and G+R because both conditions involve reanalysis, and a significant effect between G+R and G since there is an asymmetry in reanalysis; if the asymmetry is due to the nature of the representation constructed, then there should again be an effect between FG+R and G and also an effect between FG+R and G+R. This hypothesis makes no prediction about G+R and G, but that comparison would indicate whether reanalysis has any effect at all.

8 lexicalizations of each triplet were constructed and distributed using a Latin Square design. Each list contained 1 token of each condition. 10 additional conditions from an unrelated study were included as fillers. By hypothesis these 4 of these fillers were considered acceptable, while 6 were considered unacceptable, yielding a nearly balanced ratio of acceptable/unacceptable. 8 practice items were included for a total of 21 items. The task was magnitude estimation, and the instructions were identical to those of experiment 1. The reference sentence was also identical.

Results

(12) Results for Experiment 2



There was a large and significant effect of FG+R versus G+R ($t(20)=2.8$, $p=.005$, $r=.53$), and as expected of FG+R versus G ($t(20)=2.8$, $p=.005$, $r=.53$). There was no effect of G+R versus G ($t(20)=0.32$, $p=.37$). And although all of the p values were one-tailed, it should be noted that both of the significant p values were well below the Bonferroni corrected level of .017, even at their two-tailed value of $p=.01$.

Discussion

By introducing a second gap within the prepositional phrase of the true-gap condition, experiment 2 was able to eliminate the asymmetry of reanalysis from the design of the filled-gap paradigm, and thus tease apart the two possible explanations of the asymmetry in the results of experiment 1. The persistence of the effect despite the introduction of reanalysis into both conditions confirms that there is something peculiar to the filled-gap effect that affects the judgment of the final representation. Furthermore, the lack of effect between the two gap conditions suggests that reanalysis has no lasting effect on the judgment of the final representation: apparently there is no lasting cost associated with abandoning one well-formed representation for another.³

3 Conclusion

The results from experiments 1 and 2 reveal a surprising asymmetry in the effects of temporary representations on global acceptability, suggesting that syntactic difficulties are treated by the judgment process in a qualitatively different way than semantic or processing difficulties. This seems to indicate that judgment tasks are tapping directly into syntactic knowledge in a very real sense. At a methodological level, these results demonstrate the potential power of formal judgment experiments: the ability to detect significant differences between two acceptable sentences opens the possibility of using judgment experiments to explore phenomena that are typically the domain of sentence processing studies. Although it is just a promissory note at this point, these results offer the possibility of uncovering further convergences and divergences between syntactic and processing data. If that were to prove possible, it could provide a new source of data for determining the division of labor between grammar and parser.

³ Because there was no comprehension task included in experiment 2, it is possible that the lack of effect of reanalysis actually represents a lack of reanalysis, in that the subjects might not notice the gap position in the string *for during*. Of course, if it was the case that *for during* was not an appropriate cue for a gap, then it would be unclear why there was no effect of plausibility in experiment 1, as without reanalysis the implausible condition is actually unacceptable, and should have received a correspondingly low judgment.

References

- Bard, Ellen Gurman, Dan Robertson, and Antonella Sorace. 1996. Magnitude estimation of linguistic acceptability. Language, 72: 32--68.
- Cowart, Wayne. 1997. Experimental syntax: Applying objective methods to sentence judgments. Thousand Oaks, CA: Sage Publications.
- Crain, Stephen. and Janet Fodor. 1985. How Can Grammars Help Parsers?, in Natural language parsing: psycholinguistic, computational, and theoretical approaches. eds.D. Dowty, L. Karttunen and A. Zwicky, 94-128. Cambridge, UK: Cambridge University Press
- Fanselow, Gisbert and Stefan Frisch. 2004. Effects of processing difficulty on judgments of acceptability. In Gradience in Grammar. eds. G. Fanselow, C. Fery, M. Schlesewsky & R. Vogel. Oxford, UK: Oxford University Press.
- Garnsey, S. M., M. K. Tanenhaus, R. C. Chapman. 1989. Evoked potentials and the study of sentence comprehension. Journal of Psycholinguistic Research 18. 51-60.
- Featherston, Sam. 2005. Universals and grammaticality. Linguistics 43.
- Myers, James. 2006. MiniJudge: Software for minimalist experimental syntax. In Proceedings of ROCLING 18 Conference on Computational Linguistics and Speech Processing. 271-285.
- Phillips, Colin, Nina Kazanina, and Shani Abada. (2005) ERP Effects of the Processing of Syntactic Long-Distance Dependencies. Cognitive Brain Research 22: 407-428.
- Phillips, Colin. 2006. The real-time status of island phenomena. Language 82:4.
- Pickering, M. J., and Traxler M. J. 2003. Evidence against the use of subcategorization frequency in the processing of unbounded dependencies. Language and Cognitive Processes 18 (4): 469–503.
- Schütze, Carson. 1996. The empirical base of linguistics: Grammaticality judgments and linguistic methodology. Chicago: University of Chicago Press.
- Stowe, Laurie. 1986. Parsing WH-Constructions: Evidence for On-Line Gap Location", Language and Cognitive Processes 1.3: 227-245.
- Tanenhaus, M., G. Carlson, and J. Trueswell. 1989. The role of thematic structures in interpretation and parsing. Language and Cognitive Processes 4: 211-234.
- Traxler, M. J., and Pickering, M. J. 1996. Plausibility and the processing of unbounded dependencies: An eye-tracking study. Journal of Memory & Language 35: 454-475.