Going the distance: memory and decision-making in active dependency constructions

Matthew W. Wagers¹ & Colin Phillips²,³
¹ Department of Linguistics, University of California, Santa Cruz
² Department of Linguistics, ³Neuroscience and Cognitive Science Program, University of Maryland, College Park

1 Introduction

The functional description of the language comprehension architecture must specify both how the parser encodes and retrieves information about intermediate products of the parsing process as well as the decision principles for using or disposing of those encodings – roughly, what Lewis (2000) decomposes into memory and control processes. The comprehension of filler-gap dependencies, as exemplified by the relative clause in (1), has provided a fruitful paradigm for understanding the interaction between the memory and control processes of the parser. There are two reasons for their interest¹. Firstly, filler-gap dependencies are unbounded: the distance between filler and gap can be arbitrarily large, which example (2) illustrates. Consequently the encoding of the filler must be therefore durably encoded and reliably retrievable such that it can survive spans of material during which it is not being processed.

(1) One clause filler-gap dependency
[DP The stones [CP that the pilgrim [VP placed ___ on the cairn ]]] were smooth.

(2) (a) Two clause dependency
The stones [ that the monk advised the pilgrim [ to place ___ on the cairn ]] ...

(b) Three clause dependency
The stones [ that the scriptures advised the monk [ to tell the pilgrim [ to place ___ on the cairn ]]] ...

Secondly, filler-gap dependencies are ‘island-sensitive’; there are certain domains that filler-gap dependencies cannot cross, which example (3) illustrates for phrases in subject position.

(3) *The stones [ that [ cairns of ___ ] impressed the pilgrim ] were smooth.

Ross (1967) identified a number of these domains, which he dubbed ‘islands’, including adjuncts, coordinate structures, relative clauses, wh-clauses, and factive clauses. Therefore a diverse array of restrictions exists on where a gap may be found. While syntactic theories have attempted to explain the diversity of island domains as the surface manifestation of fewer, more abstract rules (Chomsky, 1977, 1986, inter alia), it remains the case that, from the perspective of the parser, not every argument position can be a gap site. The control processes for dependency construction are therefore not likely to be governed by a single, simple rule, but must incorporate a number of contingencies.

Real-time filler-gap dependency construction has been studied intensely in the past 20 years and a notably consistent, convergent account of the properties of dependency construction has emerged. In section 1.1, we briefly review the evidence that dependency construction occurs as-soon-as-possible and that it respects the constraints of the grammar. In section 1.2, we discuss the question of how fillers are stored and retrieved from memory during the comprehension process. A strong psycholinguistic tradition, beginning with Wanner & Maratsos (1978), holds that fillers are maintained in a distinguished memory state until they can be discharged to complete the dependency. This viewpoint can be challenged empirically, as there is little direct evidence that such a
distinguished state exists. It can also be challenged conceptually, as some otherwise well-supported memory models simply do not easily tolerate maintenance of the kind envisioned (McElree, 2006). The alternative is to assume that the filler is retrieved later at its integration site. However we have previously observed that the timing of dependency construction is apparently shifted from the verb to the gap site when the length of the filler-gap dependency is increased (Wagers & Phillips, 2009). This shift in timing is difficult to explain on either a pure maintenance or pure retrieval account. In a series of three experiments, we explore under what conditions the timing of dependency construction appears to shift from early, pre-gap sites to late, post-gap sites. We find that for early dependency formation, coarse-grained, categorical details are reliably present and more robust than finer-grained lexically-anchored detail. We propose a two-phase maintenance-retrieval model to account for these new data.

1.1 Constructing unbounded dependencies on-line

A fairly detailed account of how filler-gap processing proceeds has been compiled, mostly related to control processes. Intuitively, processing a filler-gap dependency requires recognizing filler, identifying a gap position, linking the filler to the gap position, and verifying that the resulting dependency is licit. An important finding regarding filler-gap processing is that comprehenders attempt to construct dependencies in advance of information about where constituents have actually gone missing. That is, comprehenders anticipate the location of potential gap sites and predictively construct a relation with the filler. This property is referred to as active dependency formation, and the supporting evidence comes from a broad variety of measures and languages (Phillips & Wagers, 2007).
One type of evidence comes from a reading-time slowdown in sentences like (4a). In this sentence, the comprehender encounters an overt direct object, ‘us’, where a gap can be expected. Reading times at that word are elevated relative to reading times for the same word in a closely matched sentence that lacks a filler-gap dependency, where no gap is expected (4b). This effect is termed the filled gap effect (Crain & Fodor, 1985; Stowe 1986; Lee 2004).

(4) a. My brother wanted to know who Ruth will bring us home to __ at Christmas.
    b. My brother wanted to know if Ruth will bring us home to Mom at Christmas.

Another type of evidence comes from processing disruptions at the verb when the filler would be an implausible argument of the verb, as in (5). Traxler & Pickering (1996) found elevated reading times for implausible filler-verb combinations at the verb, and Garnsey, Tanenhaus & Chapman (1989), an enhanced N400, the ERP component that’s often associated with lexico-semantic integration or expectancy.

(5) That’s the pistol/garage with which the heartless killer shot the hapless man __ yesterday afternoon.

It is because processing disruptions precede direct evidence in the input for the actual location of the gap that filler-gap processing is called ‘active’. In other words, comprehenders eagerly posit gaps, preferring to construct a dependency early on that might have to be retracted, than to wait for unambiguous evidence. An important question is whether this eagerness is restricted by the grammar. A number of studies have therefore examined whether island constraints guide the comprehender. Several ERP studies have indicated a processing disruption that occurs as soon as the parser encounters the edge of an island domain, if there is also an unsatisfied wh-phrase (Neville, Nicol, Barss, Forster, & Garrett, 1991; Kluender & Kutas, 1993; McKinnon & Osterhout, 1996). Behavioral studies
have further shown that active dependency formation effects are absent inside island domains. For example, Stowe (1986) found no filled gap effect inside a subject noun phrase. Likewise, Traxler & Pickering (1996) found no plausibility effect inside a relative clause island. Evidence from additional behavioral studies (Bourdages, 1992, Pickering, Barton & Shillcock, 1994, McElree & Griffith, 1998) is convergent. In other words, comprehenders recognize island domains and subsequently refrain from positing gaps inside them. More recently, studies have shown that comprehenders are sensitive in real-time to important subtleties of the filler-gap grammar, including where parasitic gaps may be licensed (Phillips, 2006) and where across-the-board extraction must occur (Wagers & Phillips, 2009).

A relatively straightforward picture emerges for the control processes that support filler-gap dependency construction. Comprehenders attempt to terminate an open filler-gap dependency wherever the grammar allows. The comprehension mechanism does not ‘wait and see’ where the actual gaps in a sentence are, but posits the gaps wherever they might occur, based on the partial input and the upcoming structure it entails. Active dependency construction thus reflects a general trend in incremental comprehension to make parsing decisions on the basis of anticipated representations.

1.2 Memory resources for filler-gap dependencies

The temporal delay between the encoding of the filler and sites in the sentence where it can be integrated may be considerable. More importantly, this delay is filled with potentially irrelevant processing. For example, consider the sentence in (4), which seems readily understood. After processing the filler and before encountering the gap site, the
comprehender must (at least) assemble the subject phrase, attach it the verb, build that verb’s complement clause, and link the matrix clause subject to its thematic role assigner in the embedded clause.

(6) **Which shrines** is the famous monk most likely to have visited __? 

The encoding of the filler has to therefore sufficiently durable so it can survive arbitrary spans of material during which it is not being processed. Note that this problem is in principle true both of cases where there is considerable hierarchical distance between filler and gap site, but also where hierarchical distance is minimal and the temporal delay is nonetheless substantial, as in (5), where the subject is modified.

(7) **Which shrines** did the monk from the ancient monastery visit ____?

Perhaps the most influential proposal is that the filler phrase is stored in a distinguished memory register until it can be assigned a role in the sentence. Wanner & Maratsos (1978) outlined an Augmented Transition Network (ATN) model of parsing which included a separate buffer, called the hold cell, for displaced NPs. When an NP is identified as the head of a relative clause, that the contents of the NP are placed in the hold cell. The hold cell allows the assignment of grammatical function to be put off until an appropriate subsequent context. For example, in an object-extracted relative clause, after a transitive verb has been processed, the ATN attempts to analyze subsequent input as an NP. If it is unsuccessful, as it would be were there a gap, then the system checks to see whether or not hold is empty; if it isn’t, it retrieves its contents, treating them as input.
This parsing sequence does not qualify as an active strategy, because a gap is not postulated unless the parser fails to recognize the verb’s lexical argument. But that is a claim about the model’s control structure, which can be adjusted independently of the memory architecture. Indeed Frazier (1987) suggested that the non-emptiness of the \texttt{HOLD} cell itself could serve as a signal to postulate gap before lexical arguments are encountered. In later work (Frazier & Flores d’Arcais, 1989), the Active Filler Strategy is formulated such that the parser is “immediately predispose[d]” to rank gaps more highly than lexical arguments, if a filler has been encountered. The authors suggest that this implicates a filler that is not ‘inert’: at predicates following the filler, gap analyses are considered before lexical argument ones because the unintegrated filler is effectively always an input. What it means to be not ‘inert’ is open to interpretation, but a strong view would be that a filler, while a dependency is incomplete, has the same representational status as bottom-up input. Considered this way, dependency completion is active because the filler is in the processing workspace before subsequent inputs, and it will effectively out-compete incoming categories for attachment.

There have been many demonstrations that filler-gap dependencies exact a cost on the comprehender. Wanner & Maratsos (1978) provided initial evidence that memory costs were higher when a dependency was incomplete. They compared subject and object relatives in a dual-task paradigm, in which participants had to both comprehend a sentence and recall a list of names that at some point intruded upon word-by-word reading. That result has since been criticized (Ford, 1983). More generally it seems that the real-time cost of open fillers is not a diffuse cost spread across the dependency, but can be localized to specific processing events. For example, the differential difficulty in subject- and object-
relatives can be localized to the verb, where reading times are elevated for object relatives (King & Just, 1991). Electrophysiological evidence has also been brought to bear on the question of memory load during an open dependency. Both King & Kutas (1995) and Fiebach, Schlesewsky & Friederici (2002) showed that in object-extracted filler-gap dependencies, the averaged EEG record reveals a sustained anterior negativity. This effect is modulated by performance and participants’ memory span and has been implicated in explicit memory load tasks (e.g. Ruchkin et al. 1990). For these reasons, the presence of a sustained anterior negativity in open filler-gap dependencies has been interpreted as a direct reflection of the memory load consumed by actively maintaining the filler. But, more pointedly, work by Phillips and colleagues (2005) has provided evidence that the sustained anterior negativity does not reflect a cumulative effect that accrues at each word, but derives mainly from the first few words of the dependency. Such a finding raises doubt that the effect reflects active maintenance of the filler representation.

Fewer studies have examined the actual contents of memory when a dependency is open. Fiebach et al. (2002), discussing the sustained anterior negativity, are cautious to point out that the electrophysiological effect does not choose between alternative accounts of what precisely is being maintained. It could be a full semantic or syntactic representation of the filler, or perhaps just a few features. Or, it may not be the content of the filler that’s represented at all, but rather that prediction for a category that allows completion of the dependency, as in Dependency Locality Theory (Gibson 2000). Swinney and colleagues, using the cross-modal lexical decision task (Nicol & Swinney 1989; Nicol, Fodor & Swinney 1994), showed that a filler’s semantic associates are primed in lexical decision immediately following the introduction of the filler in the sentence. However this
priming does not persist, declining across the sentence. Only immediately following the verb, when the filler must be contacted and thematically integrated, does the priming effect re-emerge. These results were interpreted in favor of a re-activation model, in which the contents of the filler are not actively maintained but must be re-activated into a state suitable for integration. McElree (2001) reports a related study in which sentences were interrupted at different points with a probe word, and the comprehender had to judge whether that word was synonymous with a previously encountered word or not. As in the cross-modal lexical decision task, accuracy at this task declined smoothly for positions after the filler, but grew for positions after the verb\(^2\). However, growth in accuracy was not significant until after the gap location was unambiguously signaled by the input.

On balance, the above data seem to choose against a maintenance account of filler memory. The two relevant observations are the punctate nature of the costs in long-distance dependency formation and the fact that reactivation is observed only post-verbally. These data are far from resolving the issue, however. If the filler is not maintained, it must be retrieved when it is to be integrated with the verb (McElree et al., 2003). The question arises, how is the correct constituent retrieved from memory? One possibility is that lexical information, such as a verb’s subcategorization or selectional restrictions, provides the necessary cues to retrieve the filler. The semantic anomaly studies discussed above (Garnsey, et al. 1989, Traxler & Pickering, 1999) provide a useful

---

\(^2\) The cross-modal lexical decision task has been extensively criticized by McKoon & Ratcliff (1994), who have argued the effect could be mediated by the integrability of the probe itself and not its relation to the filler. Note that in McElree (2001), a target probe word was synonymous with a modifier in the filler, and not the head noun. For example, for the filler phrase “the courageous mountaineer”, a target probe would be ‘brave’. Therefore the ability of the probe to be integrated, at least strictly compositionally, is unlikely to drive the effect.
constraint on this idea. In those studies, an anomalous combination of verb and filler was detected very early, either at the verb or immediately thereafter. This suggests that, even if a filler does not match a verb’s selectional restrictions, that it is nonetheless is retrieved. Subcategorization information may be better suited as a cue, and Van Dyke & McElree (2009) have suggested more generally that syntactic cues gate semantic cues in the retrieval of constituents form memory. Evidence remains somewhat mixed on whether fillers give rise to active effects at verbs that do not subcategorize for them (Staub, 2007, Omaki, Lau & Phillips, 2009).

It is important to re-emphasize the memory and control distinction. The process of retrieval alone – whatever the cues are – is insufficient to explain the active construction of filler-gap dependencies. If fillers were retrieved simply by bottom-up cues, with no special parsing state associated with the filler-gap dependency, we would be at a loss to explain the lack of active filling observed in island domains or in post-verbal parasitic gaps. Whatever the retrieval cues are for the filler, those results strongly suggest that they must be subsidiary to the process that initiates the retrieval.

In connection with that logic, Wagers & Phillips (2009) make a surprising observation. They compared the semantic anomaly effect for short versus long dependencies. Dependencies were serially lengthened five words by adding a PP to the subject phrase. While there was an immediate anomaly effect for short dependencies at the verb, that effect disappeared from long dependencies and was not reliable until after the gap site (an oblique gap that occurred much later in the sentence than the verb). The reason this observation is surprising is not because there is no effect at the verb per se, but because a strong, active effect at the verb became a strong, non-active effect at the gap. The
shift in the locus of the effect raises the very real possibility that active filling actually only occurs at very short distances and, possibly, that we've systematically undersampled longer-distance dependencies. Under the account we've sketched forward thus far, this would mean that the control processes that triggers retrieval at licit sites fail at a certain distance. Constructing the filler-gap dependency would then shift from top-down, global processing to local, bottom-up processing.

In the present study, we test this hypothesis by looking at three different indices of gap formation at longer dependency lengths: the filled-gap effect, the semantic anomaly effect, and a novel subcategorization index. We replicate Wagers & Phillips (2009) original result and extend it to biclausal dependencies. However, we find that the filled-gap effect survives all dependency lengths. We account for these results with a hybrid model that incorporates a minimal amount of maintenance. Such a model covers the extended data set and its modest cost is repaid with the benefit of interference robustness.

2  Experiment 1: The Filled-Gap Effect

2.1  Rationale

In the first experiment, we applied filled-gap logic to test whether active dependency formation persists over long dependency lengths. We used a modified version of the standard filled-gap logic, devised by Lee (2004). Lee (2004) contrasted displaced DPs with displaced PPs. Only displaced DPs can occupy subject or object position (in unmarked word order). In comparison to the displaced PP, displaced DPs should engender difficulty if those positions are filled. Example (8) illustrates the experimental contrast.

(8)  (a)  The stones which the pilgrim toppled the cairn for ___ ...
(b) The **stones for which** the pilgrim toppled **the cairn** ___ ...

In either (a) or (b) it is possible to form a dependency at the critical verb (‘topple’). But only when the displaced category is DP can the well-formedness of dependency be diagnosed in the direct object position. If at the verb in (a), the comprehender hypothesized that the gap occupied direct object position, then encountering the overt DP, ‘the cairn’, would disconfirm that hypothesis. If at the verb in (b), the comprehender were sensitive to the category identity of the filler, she would not hypothesize that there was a gap in direct object position. Consequently for processing the direct object, the (a) condition is expected to be more difficult than (b). This design is an improvement over the standard filled-gap logic, because it contrasts sentences that have both filler-gap dependencies and identical interpretations.

We created contrasts like (8) for three different dependency lengths: a short single-clause dependency, a longer single-clause dependency, and a longer bi-clausal dependency. If comprehenders remained sensitive to the category identity of the filler across the different dependency lengths, then we expected to observe a filled-gap effect in the object position of the critical verb in each.

2.2 Procedure

Participants were 36 native speakers of English from the University community, who received $10 for taking part in the experiment.

This experiment crossed the factors Filler Category and Length in a 2 × 3 factorial design. Filler Category was either ‘DP’ or ‘PP’. Length was either ‘Short, ‘+PP’, ‘+CP’. In Baseline Length conditions, a two-word subject and an adverb separated the filler from the
verb. In +PP Length conditions, the subject was modified by a five-word prepositional phrase (PP). Finally, in +CP Length conditions, the sentence inside the Baseline condition relative clause was embedded under a verb like ‘say’ or ‘think’, verbs do not (readily) host a DP object. Like the +PP condition, the embedding clause was comprised of five words.

Table 1 illustrates a full materials set. The category manipulation is indicated with bold face, the critical region with double underlining, and the gap position with an underscore.

<table>
<thead>
<tr>
<th>Length</th>
<th>Filler category</th>
<th>Item Frame:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Short</td>
<td>DP</td>
<td>“... [T]he chemicals ... might still become contaminated ...”</td>
</tr>
<tr>
<td></td>
<td>PP</td>
<td>... for which the technician carefully prepared the clean tubes while wearing a mask ...</td>
</tr>
<tr>
<td>+PP</td>
<td>DP</td>
<td>... which the technician carefully prepared the clean tubes for ___ while wearing a mask ...</td>
</tr>
<tr>
<td></td>
<td>PP</td>
<td>... for which the technician at the medical research facility carefully prepared the clean tubes ___ while wearing a mask ...</td>
</tr>
<tr>
<td>+CP</td>
<td>DP</td>
<td>... which the young biologist said that technician at the medical research facility carefully prepared the clean tubes for ___ while wearing a mask ...</td>
</tr>
<tr>
<td></td>
<td>PP</td>
<td>... for which the young biologist said that technician at the medical research facility carefully prepared the clean tubes ___ while wearing a mask ...</td>
</tr>
</tbody>
</table>

Table 1 Sample material set for Experiment 1

The object DP comprised the critical region in this design because it was where a filled-gap effect can first occur. Object DPs were always three-word sequences of the form ‘determiner-adjective-noun’ or ‘determiner-noun-noun’ (i.e., a noun-noun compound). The relative clause always contained an adverbial phrase in final position (in the example, “while wearing a mask”). This clause served a dual purpose. First, it provided a strong signal that a constituent was missing, by means of the sequence preposition-preposition (e.g., “for while”. Secondly it provided a within-RC spill-over region.
Finally, the Short conditions were always preceded with a five-word embedding preamble so that the position of the critical region was ordinally matched across conditions. For the illustrated set in Table 1, the Short preamble was “The biologist was distressed that ...”.

Methods
Sentences were presented on a desktop PC using the Linger software package (Doug Rohde, MIT) in a self-paced word-by-word moving window paradigm (Just, Carpenter, & Woolley, 1982). Each trial began with a screen presenting a sentence in which the words were masked by dashes while spaces and punctuation remained intact. Each time the participant pressed the space bar, a word was revealed and the previous word was re-masked. A yes/no comprehension question appeared all at once on the screen after each sentence. The ‘f’ key was used for ‘yes’ and the ‘j’ key was used for ‘no’. On-screen feedback was provided for incorrect answers. Participants were instructed to read at a natural pace and answer the questions as accurately as possible. In all of the self-paced reading experiments presented here, participants were never informed that sentences would contain grammaticality errors. Order of presentation was randomized for each participant. 7 practice items were presented before the beginning of the experiment.

Analysis
The following analysis procedure was applied to all data sets reported in this paper so we describe it in detail here.

Accuracy data was analyzed by a logistic mixed-effects model. We estimated models using the *lmer* function from *lme4* (Bates, 2007) of the R software environment (R Development Core Team, 2007). Effects are reported for models that included significant
coefficients and significantly differed from a null model (i.e., one with no experimental coefficients, only intercepts).

Participants with accuracy less than 2.5 standard deviations of mean accuracy were excluded. In Experiment 1, one participant (S35) was removed from analysis.

Reading time data were treated for outliers by trimming all observations above the 90% percentile, by region and condition. A cutoff method was chosen for analysis, as it achieved greatest power among a variety of outlier treatments explored by Ratcliff (1993) for RT distributions, which are characteristically highly positively skewed. We chose a smaller cutoff than Ratcliff used in his simulations, however. The cutoff analysis was supplemented with an analysis in which the inverse transformation was applied to all RTs, also as recommended by Ratcliff (1993). No discrepancies were discovered between the two analyses, so only the cutoff analysis is reported.

For each region, the trimmed reading times were fit by a linear-mixed effect model using lmer. The region model included fixed effects for experimental factors and random effects for Item and Participant factors. Full models were estimated with the fixed effects nested under the random effects, though they were rarely better models than those without. Unless otherwise noted, we report effects from models with no nesting. P-values and 95% confidence intervals were determined by Markov Chain Monte Carlo sampling of the posterior distribution from the model with no nested experimental factors, using the pvals.fnc function in R (Baayen, 2007).
2.3 Results

Comprehension accuracy

Mean accuracy was 84%. Comprehension accuracy by condition is reported in Table 2. The only reliable contrast was for +CP-length dependencies, which showed 10% lower accuracy on average than Short dependencies (β: -0.8 logits, p < .05).

<table>
<thead>
<tr>
<th>Filler Category</th>
<th>Length</th>
<th>+PP</th>
<th>+CP</th>
</tr>
</thead>
<tbody>
<tr>
<td>DP</td>
<td>Short</td>
<td>90%</td>
<td>89%</td>
</tr>
<tr>
<td>PP</td>
<td>90%</td>
<td>90%</td>
<td>78%</td>
</tr>
<tr>
<td></td>
<td>90%</td>
<td>89%</td>
<td>79%</td>
</tr>
</tbody>
</table>

Table 2  **Experiment 1 Comprehension accuracy**
Average percentage correct over participants, with row, column and grand means. Standard error of the cell means across subjects is 4% for all conditions, except Mismatch:Long condition, in which it is 3%. N = 18.

Reading times

Reading time results are reported in Figure 1. Regions 14-16, for which a slow-down in DP conditions would count as evidence for active dependency completion, are annotated as “active filling” regions. Regions 18-20 are annotated as “gap-driven” regions.

In brief, evidence of active dependency construction was found for all dependency lengths, with some variation in timing and size of effect.

*Region 11-12 (Baseline).* In Region 11, there was a reliable effect of Length, with +PP-length conditions being read more slowly (28 ms, 95% C.I. [10, 46], p < .005). This contrast is unsurprising, since Region 11 is lexically identical across Short and +CP-length conditions, for which it constitutes the Subject head, but different for +PP-length conditions, for which it constitutes the head of the preposition’s NP complement. In Region 12, the adverb common to all Length conditions, there were no reliable differences.
Region 13 (Verb). There were was a reliable effect of length at the verb, such that +PP-length conditions were read more quickly than Short conditions (-24 ms, 95% C.I. [-48, -1], p < .05).

Region 14-16 (Direct Object; Critical Region). We performed two analyses: a pooled region analysis, and a word-by-word analysis. In the pooled analysis, there was a reliable effect of Filler, such that DP-filler conditions were read more slowly (11 ms, 95% C.I. [5, 17], p < .001).

As examination of Figure 1 reveals, this slow-down was present in all Length conditions, but there was variation in timing. In Region 14, the only reliable slow-down was for CP-length conditions (18 ms, 95% C.I. [1, 35], p < .05). In Region 15, there was a reliable effect of Filler (34 ms, 95% C.I. [15, 56], p < .005) for Short-length conditions, and a reliable interaction term for +CP-length conditions (-34 ms, 95% C.I. [-52,-8], p < .01). The interaction term for +PP-length conditions was marginal (-24 ms, 95% C.I. [-51,2], p < .10). CP-length conditions were also read more quickly overall in this region (-28 ms, 95% C.I. [-46, -9], p < .005.). In Region 16, there was only a reliable slow-down for DP-fillers for +PP-length conditions (21 ms, 95% C.I. [2, 39], p < .05).

Region 17 (Stranded preposition for DP conditions). Region 17, the stranded preposition, was only matched across DP-filler conditions. There were, however, length effects: both +PP- and +CP-length conditions were read more slowly than Short dependencies: +PP (14 ms, 95% C.I. [0, 28], p < .05), +CP (23 ms, 95% C.I. [9, 37], p <0.005).

Regions 18-20 (Post-gap regions). Regions 18-20 constituted the post-gap regions, for which we performed both a pooled analysis and a word-by-word analysis. In the pooled analysis, there were length effects, such that both PP- and CP-length conditions were read
more slowly than Short dependencies: PP (27 ms, 95% C.I. [16, 37], p < .001), CP (15 ms, 95% C.I. [4,26], p < .01). In addition, PP-filler conditions were read more slowly (12 ms, 95% C.I. [2, 23], p < .05), except for in +PP-length conditions, in which case they were read more quickly (-29 ms, 95% C.I. [-45, 15], p < .001).

In Region 18 the only reliable contrast was for +CP-length conditions, which were read more slowly (27 ms, [10, 45], p < .005), apparently continuing the trend from Region 17. In Region 19, there was a reliable effect of Filler for just +PP-length conditions, such that DP-filler conditions were read more slowly (40 ms, [12, 68], p < .01). In Region 20, there was a marginal tendency for DP-filler conditions to be read more quickly (14 ms, [-30, 1], p < .10), which was reliably reversed for +PP-length conditions (23 ms, [1, 45], p < .05).
Figure 1: Experiment 1 Reading Times

Closed symbol: DP filler phrase; open symbol: PP filler phrase

Sample sentence

[...] the chemicals (for) which [...]CP the technician [...]PP carefully12 prepared13 the14 clean15 tubes16 (for17) while18 wearing19 a20 mask21 might still become contaminated.
Each panels corresponds to a different Length condition. Error bars indicate standard error of the cell means. Reliable pair-wise differences indicated with the standard symbols—  

\[ p: \quad *** < .005 < ** < .01 < * < .05 < \bullet < .10 \]

2.4 Discussion

In this experiment we tested whether participants actively constructed the filler-gap dependency at all dependency lengths by contrasting DP fillers with PP fillers. DP sentences were expected to be read more slowly than PP sentences in the direct object, if the dependency was constructed actively. This filled-gap effect was found at all dependency lengths. Participants thus actively constructed the filler-gap dependency, even for long biclausal or PP-extended monoclusal dependencies. Our results thus support the claim of Frazier & Clifton (1989), that the active filler strategy is successive cyclic. However it offers greater resolution than that claim, because not only can we conclude that filler-gap dependency construction itself remains active across clause boundaries, but also that sensitivity to the filler’s grammatical category is retained when actively positing gaps.

The filled-gap effect was observed as a ‘punctate’ effect, whose onset varied across conditions. That is, the entire direct object region wasn’t rendered more difficult, but a particular word was. Interestingly this effect appears earliest for the +CP-length dependency: on the determiner of the direct object. It appeared in the middle word for short dependencies (an adjective or a noun), and on the last word for +PP-length dependencies (a noun). Because of the temporal resolution of the self-paced reading technique, we cannot conclude anything about the finer time-scale dynamics of the process. However in every condition pair the effect occur before the direct object, so we can still conclude the gap condition was posited before direct evidence of the missing constituent’s location. Beyond that, we can only speculate that the variation arises from the different
processing events that precede the verb. In other words, when the processor actually evaluates a hypothetical filler-gap dependency likely depends upon the status and difficulty of other interpretive or syntactic decisions made prior to the onset of the VP.

There was one unexpected effect, which was pronounced difficulty for DP filler phrases in the post-gap region of +PP-length sentences. This suggests that linking the gap and the filler is more difficult for dependencies in which the subject has been PP-modified. There is no particular reason to have supposed this would have been the case. It is possible that projecting a gapped PP, as in the case of DP-filler conditions, interferes with attaching an actual PP in the input. If these processes overlapped, then comprehension would be more difficult for DP-filler phrase sentences. This explanation is admittedly post hoc, and could have easily gone the other way (e.g., projecting a gap whose category is PP, as in the case of PP-filler conditions, would be more difficult, when there are other PPs in the sentence).

3 Experiment 2: Semantic anomaly detection

3.1 Rationale

In the second experiment we used another index of active dependency formation: the semantic anomaly effect (Garnsey, Tanenhaus & Chapman, 1989, Traxler & Pickering, 1996). Example (9) illustrates a semantic anomaly contrast. If filler were initially treated as the direct object of the verb ‘erect’, then the interpretation would describe a plausible outcome in (5a), but not in (9b).

(9)  (a) The **monument which** the pilgrim **erected** to appease the gods __ …
     (b) The **gods which** the pilgrim **erected** the monument to appease __ …

For this reason a reading time slowdown is expected on the verb or shortly thereafter.
We created contrasts like (9) for three different dependency lengths: a short single-clause dependency, a longer single-clause dependency, and a longer bi-clausal dependency. If comprehenders remained sensitive to the semantic details of the filler across the different dependency lengths, then we expected to observe a semantic anomaly effect either at the critical verb or shortly thereafter.

3.2 Materials and Methods

Participants were 36 native speakers of English from the University community, who received $10 for taking part in the experiment.

This experiment crossed the factors Filler Plausibility and Length in a $2 \times 3$ factorial design. Filler Category was either ‘DP’ or ‘PP’. Length was either ‘Short,’ ‘+PP’, ‘+CP’. The Length manipulation was identical to Experiment 1. Table 3 illustrates a full materials set. The category manipulation of the relativizer is indicated with bold face, the critical region with double underlining, and the gap position with an underscore.
Table 3  **Sample material set from Experiment 2**

<table>
<thead>
<tr>
<th>Length</th>
<th>Filler plausibility</th>
<th>Item Frame:</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Short</strong></td>
<td>Plausible</td>
<td>The <em>acorns</em> which the squirrels quickly <strong>crammed</strong> their small puffy cheeks with ___ before scurrying out of the park had fallen from a sickly oak.</td>
</tr>
<tr>
<td></td>
<td>Implausible</td>
<td>The <em>cats</em> which the squirrels quickly <strong>crammed</strong> their small puffy cheeks with ___ before scurrying out of the park were basking in the sun.</td>
</tr>
<tr>
<td><strong>+PP</strong></td>
<td>Plausible</td>
<td>The <em>acorns</em> which the squirrels with the bushy black tails quickly <strong>crammed</strong> their small puffy cheeks with ___ before scurrying out of the park had fallen from a sickly oak.</td>
</tr>
<tr>
<td></td>
<td>Implausible</td>
<td>The <em>cats</em> which the squirrels with the bushy black tails quickly <strong>crammed</strong> their small puffy cheeks with ___ before scurrying out of the park were basking in the sun.</td>
</tr>
<tr>
<td><strong>+CP</strong></td>
<td>Plausible</td>
<td>The <em>acorns</em> which the experienced naturalist concluded that the squirrels quickly <strong>crammed</strong> their small puffy cheeks with ___ before scurrying out of the park had fallen from a sickly oak.</td>
</tr>
<tr>
<td></td>
<td>Implausible</td>
<td>The <em>cats</em> which the experienced naturalist concluded that the squirrels quickly <strong>crammed</strong> their small puffy cheeks with ___ before scurrying out of the park were basking in the sun.</td>
</tr>
</tbody>
</table>

The verb is the critical region in this design because it is where the semantic fit of the filler can first be detected. However because such effects are prone to spilling-over in self-paced reading (Wagers & Phillips, 2009), the direct object region was made lengthy (4 words), so that evidence for the gap would be unavailable for some time after the verb. Consequently, a slowdown for implausible conditions either at the verb or in the direct object region could be considered an active dependency construction effect.

### 3.3 Results

**Comprehension accuracy**

Mean accuracy was 91%. One participant was excluded from future analysis for exceptionally low accuracy (more than 2.5 standard deviations lower than mean accuracy). Comprehension accuracy by condition is reported in Table 4. There was a 2% decrement in
accuracy for *Implausible* sentences (logit $\beta$: -0.19, $p < 0.05$). Both +PP and Short length conditions were answered more accurately than +CP-length conditions (respectively, $\beta$: 0.85, $p < 0.005$; $\beta$: 0.25, $p < 0.001$). No interaction between Length and Filler was reliable.

<table>
<thead>
<tr>
<th>Filler Category</th>
<th>Length</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Short</td>
</tr>
<tr>
<td><strong>Plausible</strong></td>
<td>91%</td>
</tr>
<tr>
<td><strong>Implausible</strong></td>
<td>88%</td>
</tr>
<tr>
<td></td>
<td>91%</td>
</tr>
</tbody>
</table>

**Table 4  Experiment 2 Comprehension accuracy**

Average percentage correct over participants, with row, column and grand means. Standard error of the cell means across subjects is 3% for +CP conditions, and *Implausible:Short*. For the three remaining conditions, it is 2%. N = 35.

**Reading times**

Reading time results are reported in Figure 2. Regions 12-16, for which a slow-down due to implausibility would count as evidence for active dependency completion, are annotated as “active filling” regions. Regions 18-20, for which a slow-down due to implausibility could be attributed to direct evidence for the location of a missing constituent, are annotated as “gap-driven” regions.

In brief, active filling was only robustly found in the Short Length conditions. In the active regions (Regions 12-16), there was a slow-down for implausible conditions, indicating sensitivity to the filler’s fit as an argument of the critical verb. However this slow-down was absent for CP-length conditions. For +PP-length conditions, it was smaller and reliable only in an analysis pooling over several regions. PP and CP Length conditions did show sensitivity in the gap-driven regions (Regions 19 and beyond). Overall this
pattern of results contrasts with the filled-gap effect in Experiment 1, which was detected for all length conditions.

*Regions 10-11 (Baseline).* In Region 10, +PP-length conditions were read more slowly (19 ms, 95% C.I. [3, 35], p < .05). This is not especially surprising, since lexically the materials varied at this position from the other Length condition. In Region 11, no effects were reliable in the full model.

*Region 12 (Verb).* This region shows two reliable contrasts: a slow-down for CP-length conditions (20 ms, 95% C.I. [1, 37], p < .05) and a slow-down for implausible fillers in Short-length conditions (29 ms, 95% C.I. [2, 54], p < .05).

*Regions 13-16 (Ground Argument).* We performed two analyses: a pooled region analysis, and a word-by-word analysis. In the pooled region analysis, the only reliable contrast was the Short-length/Filler interaction, reflecting a slow-down for implausible fillers in short dependencies (23 ms, 95% C.I. [12, 34], p < .001). There was a marginal plausibility effect for +PP-length fillers (10 ms, 95% C.I. [-1, 21], p < .10). In a post-hoc test that excluded the CP-length conditions, there was a reliable plausibility effect overall (10 ms, 95% C.I. [2, 19], p < .05) and a reliable interaction with Short-dependencies, such that Short/Implausible conditions were read even more slowly (13 ms, 95% C.I. [1, 25], p < .05).

Region 13 showed the same pattern as the pooled analysis: a reliable slow-down for implausible fillers in short dependencies (38 ms, 95% C.I. [15, 61], p < .001). No effects approached significance in Region 14. In Region 15, there was a reliable slow-down for implausible fillers in short dependencies (31 ms, 95% C.I. [5, 55], p < .05). A marginally reliable slow-down was also present for +PP-length dependencies (21 ms, 95% C.I. [-4, 46], p < .10). No effects were reliable in Region 16.
Region 17 (Gap-hosting preposition). In this region, there was a reliable effect of structure, such that Short-dependencies were read more quickly than other length conditions (-16 ms, 95% C.I. [-30, -2], p < .05).

Region 18-20 (Post-gap regions). In the pooled analysis, there as a slow-down for implausible fillers (19 ms, 95% C.I. [10, 29], p < .001). No effects were reliable in Region 18. In Region 18, there was a slow-down for implausible fillers (25 ms, 95% C.I. [7, 42], p < .01). In Region 19, the anomaly effect as also present (23 ms, 95% C.I. [6, 39], p < .01). There is a marginal interaction for +PP-length conditions, which goes in the opposite direction (-19 ms, 95% C.I. [-43, 3], p < .10).
Figure 2  Experiment 2 Reading times

Sample sentence

... x ... ]9-10 quickly11 crammed12 their13 small14 puffy15 cheeks16 with17 ___ before18 scurrying19 out20 [of the park ... ]20

Each panels corresponds to a different Length condition. Error bars indicate standard error of the cell means. Symbols indicate the result of a pair-wise test of filler plausibility—  \( p: \quad *** < .005 < ** < .01 < * < .05 < \bullet < .10 \)
3.4 Discussion

Across all dependency lengths comprehenders were ultimately sensitive to the semantic fit of the filler for the verb hosting its gap. However the onset of that sensitivity, as reflected in slower reading times for implausible verb-filler combinations, varied with dependency length. For long, bi-clausal dependencies, comprehenders only showed sensitivity once there was direct evidence in the input for the location of the missing constituent. This contrasted strikingly with short mono-clausal dependencies, where sensitivity arose immediately at the verb. Mono-clausal dependencies that were lengthened (temporally) by modifying the subject with a PP showed some sensitivity to the semantic fit of the filler prior to the gap region. However, the RT difference in the active region was reliable smaller for +PP-length dependencies than the short mono-clausal dependencies. This contrast for anomaly detection in short and +PP-lengthened dependency replicates a previous finding Wagers & Phillips (2009; Experiment 3).

This pattern of results can be interpreted in one of two ways: either (a) active dependency formation ceased by the time the most deeply embedded verb is reached in a bi-clausal dependency, or (b) the lexico-semantic features of the filler required to evaluate whether the verb’s selection restrictions were met did not participate in the active part of dependency construction for longer dependencies. The results of Experiment 1 rule out interpretation (a). In that experiment, the filled-gap effect was present in both mono-clausal and bi-clausal dependencies. Therefore, some version of interpretation (b) must be correct. It is important to observe, however, that is not simply the case that longer dependencies are just generally more taxing, and that the effect of semantic fit declines as length grows. In both long dependencies, the effect of semantic fit was reliable in the post-
gap region, and not different from short dependencies. Indeed, the most robust post-gap effect was found for +CP-length dependencies, which showed a total absence of active effect. Therefore the lexico-semantic features of the filler are not forgotten absolutely: in the active regions they are either inaccessible or not used to evaluate the dependency. The results suggest a trading relation between when an plausibility anomaly shows up with respect to the gap, and how long the dependency is.

We interpret this trading relation in the context of the previous experiment and propose that long-distance dependency formation reflects the integration of information in two states. The heart of active dependency formation is the act of predictive structure building: dependencies are projected forward in time by the comprehender, in advance of direct evidence for the gap location. However, we propose that the extent to which the dependency is initially elaborated or evaluated is limited by the information about the filler that is carried forward in time. In other words, the comprehender maintains what she can about the filler in an active representational state, and this is what guides initial dependency construction. This ‘active’ state can be identified with the focus of attention or capacity of concurrent processing (Broadbent, 1958, McElree, 2006, Jonides et al., 2008). Since this capacity is limited, a complete representation of the filler cannot be maintained; other processing events intervene, displacing some of the details of the filler. Therefore the initial dependency should be thought of as a provisional representation, whose details are filled out by retrieving the complete representation of the filler from memory, the process normally identified as reactivation (Nicol & Swinney, 1989).

Experiments 1 and 2 suggest that the category identity of the filler is maintained across all dependency lengths, whereas its semantic details are not. This makes sense, both
architecturally and functionally. From an architectural standpoint, the information that encodes category identity is briefer than the information about semantic features. This is because the category identity is a coarser, more general representation for which there are only a few possibilities. In contrast, individual semantic features (e.g., +animate, +concrete, etc.) are finer-grained, and there are many more possibilities to be encoded. From a functional standpoint, category information is grammatically more useful in evaluating the well-formedness of dependency. A sentence may express an unusual or implausible state of affairs, and yet be well-formed.

We flesh out this account in greater detail in the general discussion, but immediately it raises the question of what the appropriate level of ‘coarseness’ is for maintenance. In Experiment 3, we present a new index of active dependency formation, in which the information that must be maintained about the filler is not as general as category identity, but still more constrained than the semantic features.

4 Experiment 3: Subcategorization Match

4.1 Rationale

In the third experiment we devised a new index of active dependency formation, which we call subcategorization match. It is based upon the fact that arguments bearing particular thematic roles must occur as the complement of specific prepositions. For example, the verb ‘inherit’ requires a source argument. In English, source arguments are the complement of the preposition ‘from’ (10a). In contrast, verbs like ‘entrust’ require a goal argument. Goal arguments are the complement with the preposition ‘to’ (10b).

(10) (a) The orphan inherited the prayer book from/*to his uncle.
(b) The monk entrusted the prayer book *from/to his novice.

If we displace the Source/Goal argument, the restriction on preposition identity still holds.

(11) (a) The uncle from/*to whom the orphan inherited the prayer book
(b) The novice *from/to whom the monk entrusted the prayer book.

If comprehenders are sensitive to not just the category identity of a displaced PP, but also the particular preposition that heads the PP, then there should be a reading time slowdown when the displaced preposition is not appropriate for the kind of semantic argument the verb requires.

The subcategorization mismatch manipulation was tested as two different dependency lengths: short monoclausal dependencies and PP-extended monoclausal dependencies.

4.2 Materials and methods

Participants were 18 native speakers of English from the University community, who received $10 for taking part in the experiment.

This experiment crossed the factors Match and Length in a 2 × 2 factorial design. Match was either ‘Match’ or ‘Mismatch’, corresponding to the subcategorization match between verb and PP. Length was either ‘Short’ or ‘+PP’. As in Experiment 1, Short Length conditions separated filler and verb by a two-word subject and an adverb. +PP Length conditions attached a five-word PP modifier to the subject. Table 5 illustrates a full materials set. The Match manipulation is indicated with bold face, the critical region with double underlining, and the gap position with an underscore.
<table>
<thead>
<tr>
<th>Length</th>
<th>Match</th>
<th>Item Frame</th>
</tr>
</thead>
<tbody>
<tr>
<td>Short</td>
<td>Match</td>
<td>... <strong>to whom</strong> the secretary warily <strong>entrusted</strong> the confidential business correspondence __ after some hesitation ...</td>
</tr>
<tr>
<td></td>
<td>Mismatch</td>
<td>... <strong>from whom</strong> the secretary warily <strong>entrusted</strong> the confidential business correspondence __ after some hesitation ...</td>
</tr>
<tr>
<td>Long</td>
<td>Match</td>
<td>... <strong>to whom</strong> the secretary for the high-powered defense attorney warily <strong>entrusted</strong> the confidential business correspondence __ after some hesitation ...</td>
</tr>
<tr>
<td></td>
<td>Mismatch</td>
<td>... <strong>from whom</strong> the secretary for the high-powered defense attorney warily <strong>entrusted</strong> the confidential business correspondence __ after some hesitation ...</td>
</tr>
</tbody>
</table>

Table 5  Sample material set for Experiment 2

Items were balanced so that the “Match” preposition was ‘to’ in 12 sets and ‘from’ in 12 sets. Materials were distributed according to a Latin Square across four lists so that participants each read six sentences per condition. 72 filler sentences were included, adapted largely from Experiment 1. Some new distractors were devised that included pied-piping in contexts other than inside of a subject-attached relative clause. The purpose of this manipulation was to ensure that participants could not strategically identify the experimental targets.

Presentation and analysis details were as described in Experiment 1. Regions of interest were structurally aligned so that the VP words common to all conditions in an item set were analyzed word-for-word across conditions.

4.3 Results

Accuracy

Comprehension accuracy is reported in
Table 6. There was a 4% decrement in accuracy for long dependencies, and a 4% decrement for verbs with mismatching PP arguments. However none of these effects reached significance.

<table>
<thead>
<tr>
<th></th>
<th>Match</th>
<th>Mismatch</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Length</strong></td>
<td><strong>Short</strong></td>
<td><strong>Long</strong></td>
<td></td>
</tr>
<tr>
<td>Match</td>
<td>88%</td>
<td>84%</td>
<td>86%</td>
</tr>
<tr>
<td>Mismatch</td>
<td>84%</td>
<td>80%</td>
<td>82%</td>
</tr>
<tr>
<td></td>
<td>86%</td>
<td>82%</td>
<td>84%</td>
</tr>
</tbody>
</table>

**Table 6** **Experiment 3 Comprehension Accuracy**

Average percentage correct over participants, with row, column and grand means. Standard error of the cell means across subjects is 4% for all conditions, except Mismatch:Long condition, in which it is 3%. N = 18.

**Reading times.**

Reading time results are reported in Figure 3. The main finding from the reading time data is that there is only sensitivity to the verb-PP match in short dependency conditions. This sensitivity is almost entirely reflected by a large slow-down in Mismatch sentences compared to Match sentences, observed at the verb and in the immediately following region.
**Figure 3**  
**Experiment 3 Reading times**

*Sample sentence*

The courier to/from whom the secretary [for the high-powered defense attorney] ... warily_{12} entrusted_{13} the_{14} confidential_{15} business_{16} correspondence_{17} after_{18} some_{19} hesitation_{20}

Each panel corresponds to a different Length condition. Error bars indicate standard error of the cell means. Symbols indicate the result of a pair-wise test of filler plausibility— $p: *** < .005 < ** < .01 < * < .05 < * < .10$

**Region 12 (Adverb Baseline).** There were no reliable effects in this region.

**Region 13 (Critical verb).** In this region, there were two marginal effects in the overall model: a speed-up for Long conditions (-25 ms, 95% C.I. [-53, 3], $p < .10$) and an interaction of Length and Match (-50 ms, 95% C.I. [-105, 7], $p < .10$). The pairwise comparisons revealed the sources of these effects: for Matching conditions, there was no
effect of Length (Short: 418 ms ± 15 ms, Long: 418 ms ± 13 ms), and Short conditions alone showed a slow-down for Mismatching conditions (44 ms, 95% C.I. [3, 83], p < .05).

Regions 14-17 (Argument regions). In a pooled analysis of the four-word argument region, the only reliable effect was a slowdown for mismatching conditions (9 ms, 95% C.I. [2, 17], p < .05). Region 14 showed a pattern of effects similar to Region 13: an effect of Match (19 ms, 95% C.I. [3, 33], p < .05) and Length (-17 ms, 95% C.I. [-33, -3], p < .05). As in Region 13, there was no reliable difference of Length between Matching conditions, but Short conditions alone showed a slow-down for Mismatching conditions (29 ms, 95% C.I. [8, 51], p < .01). There were no reliable effects in Regions 15-17.

Regions 18-20 (Post-gap regions). There were no reliable effects in the pooled or word-by-word analyses.

4.4 Discussion
This experiment provided evidence that comprehenders were sensitive online to the subcategorization restriction that verbs place on a displaced PP-argument, but only when the distance between the displaced PP and the verb was short. When more than a two-word subject intervened between the PP filler and the verb, there was no indication in the reading times that participants detected a mismatch. This finding suggests that the information about the identity of the preposition is not well-preserved over a relatively short dependency.

Although the PP-identity manipulation is new, the logic is identical to existing experiments on active dependency completion. In the Short conditions, there is evidence that a mismatch PP is detected at the verb and in the successive region. The presence of a
slow-down in this region indicates that comprehenders are evaluating the displaced PP as a potential argument of the verb. We therefore conclude that comprehenders were actively completing the dependency.

An important difference between this experiment, and Experiment 2, should be highlighted. In Experiment 2, the semantic anomaly robustly led to a slow-down in active regions for Short sentences, but not for Long sentences. However, in Long sentences, there was eventually a slow-down, in the non-active, post-gap regions. In contrast, in the present experiment, there was no eventual slow-down for Long sentences. It may be that there was enough variability in the timing and strength of this effect at the individual level that we weren’t able to observe a ‘punctate’ response in the average reading times. Or it may be that the anomaly was never detected. In a follow-up speeded grammaticality test, we presented 32 participants with the sentences from Experiment 3. For Short sentences, participants accepted grammatical sentences 83% of the time, but rejected the ungrammatical sentences only 29% of the time. For Long sentences, participants accepted grammatical sentences 67% of the time, and rejected ungrammaticals 43% of the time. Overall sensitivity to this violation was low in speeded judgments (average d’: 0.3).

However, there were only main effects of Match (β: -.71 ± 0.49, p < .001) and Length (β: -.95 ± 0.49, p < .001), and no interaction (by mixed-effects logistic regression). We therefore do believe that individuals do not retain sensitivity to the lexical identity of the preposition in these sentences for a very long time. A likely reason is that the verb is so semantically constraining, that only information about the preposition’s complement is useful to retain. We discuss this possibility in greater detail in the General Discussion.
5  General discussion

5.1  Summary of results

In three experiments we tested the sensitivity of different indices of active dependency formation to extended dependency lengths. In Experiments 1 and 2, we tested the filled-gap effect and the plausibility effect (Experiment 2), both of which are commonly found in the literature on processing filler-gap dependencies (e.g., Stowe, 1986, Garnsey, Tanenhaus & Chapman, 1989, Traxler & Pickering, 1996, Lee, 2004). In Experiment 3, we tested the subcategorization match, which is novel. As Table 7 schematizes, how these indices differ is in specificity of the information needed about the filler phrase to distinguish the two sentences.

The filled-gap effect tested in Experiment 1 was apparent at all dependency lengths. The logic of the filled-gap manipulation we tested was as follows: if the parser posited a direct object for DP fillers, but not PP fillers, then there should be increased reading times in the direct object region for DP conditions compared to PP conditions. In that region, a reanalysis would be triggered for the DP filler sentences, but not the PP filler sentences. The fact that a filled-gap effect was apparent at all dependency lengths suggests that the parser actively completed the filler-gap dependency in all of those contexts, and that it projected specifically object gaps in the case of DP fillers. This result is consistent with a similar finding by Frazier & Clifton (1989).

In contrast to the filled-gap effect, the plausibility effect in Experiment 2 showed more variation. The logic of the plausibility effect was as follows: if the parser considered the filler an argument of the critical verb, then fillers that were semantically anomalous arguments would lead to increased reading times at the verb or in the direct object region.
In the active regions this effect was strong for short dependencies, but completely absent for long biclausal dependency. For long, PP-extended dependencies, it was present but weak. Of particular importance is the fact that a plausibility effect was present in the non-active, post-gap regions for the long conditions that showed weak or no effects in the active regions. Comprehenders therefore did not fail to notice the semantically anomalous arguments, but they did not notice them before the gap. We observed the same pattern of results in an earlier experiment (Wagers & Phillips, 2009, Experiment 3), which only included the PP-lengthened condition.

Finally, the subcategorization match effect in Experiment 3 showed a profile similar to the plausibility effect. The logic of this index was as follows: if the parser considered the PP filler an argument of the critical verb, then fillers whose functional head did not match the requirements of the verb would lead to increased reading times at the verb or in the direct object region preceding the gap. However such an effect was only apparent for Short dependencies. As was the case in Experiments 1 and 2, the active effect in short dependencies was large and occurred early in the regions of interest. As was also the case in Experiments 1 and 2, the effect was attenuated in a long, PP-extended dependency – in fact, the attenuation was total. Unlike Experiment 2, there was no post-gap detection of the anomaly. It is therefore unclear whether comprehenders ever recovered sensitivity to the identity of the filler’s functional head. A follow-up speeded grammaticality test indicated that end-of-sentence judgments were not very accurate; however, there was no interaction with length.
Table 7  Information required to distinguish critical sentences

The depth of information manipulated in the filler phrase differs among experiments. In Experiment 1, the information is coarse-grained: the category of the displaced phrase. In Experiment 2 is relatively fine-grained: semantic features of the lexical head of the filler are necessary. Experiment 3 represents an intermediate grain: the fillers are identical in category (PP), but differ in the lexical identity of the functional head (‘to’ v. ‘from’). The symbol ‘=’ indicates no difference for a given property.

5.2  Decision-making and memory in active dependency construction

We conclude that filler-gap dependencies are actively constructed regardless of dependency length. This would explain the pattern of results observed in Experiment 1, and by Frazier & Clifton (1989): even long, biclausal filler-gap dependencies are formed actively. What must be accounted for, then, is why longer dependencies failed to show any effects stemming from properties of the filler phrase that were more specific than syntactic category. The likely explanation, we argue, is that certain information is immediately accessible to the processor to make parsing decisions, while other information must be retrieved. Findings from a variety of cognitive tasks indicate that memory is partitioned into a highly capacity limited focal state (Cowan, 2001, Garavan, 1998, Jonides et al., 2008, McElree, et al., 1989, 2000, 2006, Verhaegen, et al., 2004) and a virtually unlimited non-focal state. Information in the focal state can be incorporated into on-going processing
directly. However, information that has been displaced from this state must be reincorporated by retrieval operations, a finding supported by speed-accuracy trade-off time-course studies of list memory (McElree & Dosher, 1989; McElree, 1998; McElree, 2006) and, more recently, studies of subject-verb dependencies in language comprehension (McElree et al., 2003).

Interpreted in view of this architecture, our data suggest that the comprehension system completes filler-gap dependencies based on two complementary processes: a “leading” process, which projects structure ahead of the input based on a limited amount of information in the focus of attention; and a “lagging” process which involves the retrieval of the full filler phrase. The longevity of the filled-gap effect derives from the fact that active dependency formation can be driven by a very minimal amount information – i.e., the information that a filler of category X exists – and that this small amount of information can survive during the course of the sentence, even when there are unrelated processing events. The filled-gap effect stems from just the minimal amount of information necessary to drive active filling, whereas the other tests depend on fuller syntactic and semantic details. In Experiments 2 and 3, when the filler-gap dependency was initially formed in long dependencies, lexically-specific information about the filler phrase was not available that would lead the comprehender to notice any anomalies. Once the comprehender retrieved the filler, though, the acceptability of the dependency could be evaluated.

The proposed interplay of predictive, leading processes and retrospective, lagging processes can be thought of as a straightforward hybridization of the two leading mechanisms of filler-gap processing: a maintenance mechanism, as in Wanner & Maratsos’ (1978) HOLD CELL hypothesis, and Frazier (1987)’s related idea that the filler is somehow
not 'inert'; and a reactivation mechanism, proposed on the basis of cross-modal lexical priming and recognition tasks (Bever & McElree, 1989; Nicol & Swinney, 1989; Nicol, Fodor & Swinney, 1994). Moreover, the sequence of events we propose above seems to be reflected rather directly by the dissociation of the plausibility effect in Experiment 2: the effect appeared on the verb for short dependencies, but downstream from the gap for longer dependencies. Immediately the question arises why retrieval does not occur immediately, rather than after position where the gap site is confirmed. A likely possibility is simply that other linguistic relations are being constructed simultaneously: after projecting the gap position, and not having it disconfirmed in Experiments 2-3, the processor shifts its resources to interpret the direct object. Data from probe recognition studies are relevant here, but are notoriously mixed (see McKoon & Ratcliff, 1994, Nicol, Fodor & Swinney, 1994). Among studies that have found evidence for reactivation, there is Nicol & Swinney (1989), who report that, in a cross-modal lexical priming task, facilitation (in RTs) for a visually-presented, semantic associate of the filler is found in the position immediately following the verb. On the other hand, in a synonym judgment task, McElree (2001) obtained the greatest facilitation (in accuracy) several words downstream, after which the gap had been unambiguously located. It is important to note that in both of these cases, the gap was in direct object position. In our case, the gap was in an oblique position. Further research is therefore needed to test the generality of the results.
5.3 The capacity of concurrent processing and underspecification

The amount of information that can be concurrently processed is clearly limited (Broadbent, 1958), but what the exact nature of these limitations are for linguistic structure remains poorly understood. For list memory, it is clear that focal attention does not apply to merely to the last percept, and that chunking and task expectations play an important role in determining what is maintained in focal attention (McElree, 2006). An idea with deep roots in psycholinguistics is that the capacity bottleneck delimits an expression into its major constituents, a notion which motivated the ‘click’ experiments in 1960s (see Fodor, Bever & Garrett, 1974, for discussion). More recently, in the ACT-R model of sentence processing (Lewis & Vasishth, 2005), it is a maximal projection (i.e., an XP) that can be maintained in the ‘active’ buffer.

What these ideas have in common is that the capacity bottleneck functionally carves up expressions into contiguous extents of structure. Put another way, the span of concurrent processing limits the breadth of information than can be represented at any given moment.

An alternative view, and one with which our proposal for processing filler-gap dependencies is consistent, is that limits on concurrent processing exert their influence primarily on the depth of information that can be represented at any given moment. Instead of maintaining all the features of a given constituent with full precision, the comprehender may choose to discard some features from focal attention, in order to accommodate information from other constituents. For filler-gap dependencies, maintaining a fully articulated encoding of the filler phrase while processing intermediate portions of the sentence may simply be impossible. But it may nonetheless be feasible to
maintain a pared-down representation of the filler-phrase, with just enough information to make the most crucial parsing decision, i.e. where to posit a gap.

Limits on the depth of representation can be thought of as a kind of underspecification, a concept which has received renewed attention in psycholinguistics, (Frisson, 2008, Sanford & Sturt, 2004, Weinberg, 1993). By underspecifying most constituents in the syntactic context, the comprehender frees up capacity to encode and process incoming information, while nonetheless keeping at the ready some amount of global context upon which to base parsing decisions. The availability of this global context would help overcome a major limitation for breadth-based segmentation as implemented in a content-addressable memory. Recovering displaced information requires retrieval from a content-addressable memory. While content-addressable memories are excellent architectures for recovering encodings on the basis of the information they contain, they are less well-adapted for retrieval based on relations between separate encodings in memory. C-command is such a relation, and one that regulates a wide variety of grammatical processes. Enforcing c-command relations is thus problematic when a constituent that is not currently maintained must be retrieved. However, under the present view, distant constituents aren’t retrieved in toto, but more information is retrieved about them, i.e., to reintegrate unspecified feature values.

What determines how much information can be maintained remains speculative. We conjecture that the major limitation is on the binding of specific lexical material to positions in the structure, an idea supported by recent computational modelling (Van der Velde & de Kamps, 2006, cf. Phillips & Wagers, 2006). In comparison to the large number of words that can instantiate a given structure, the individual structural motifs of a
comprehender’s language are relatively limited and well-practiced. Kintsch and Ericsson (1995) have argued that processing in many cognitive domains is rapid, because domain-specific knowledge determines the form of optimal retrieval cues for solving a given task. In our case, the domain-specific knowledge comes in the form of the abstract structures that bind together the novel expressions a comprehender must be prepared to interpret. However, future research is needed to broaden the extent of the data and more rigorously pinpoint the mechanisms we have proposed.
6 References


