# Short-term memory retrievals and expectation in on-line sentence comprehension: the effects of recent linguistic context

by

Brian D. Bartek

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Psychology)
in The University of Michigan
2011

Doctoral Committee:

Professor Richard L. Lewis, Chair

Professor Julie E. Boland

Professor David E. Meyer

Professor Shravan Vasishth

## ACKNOWLEDGEMENTS

I owe a great deal to several people. My parents Dave and Mary made sure I was in a position to focus on my work throughout my postgraduate education. My fiancee, Kary, kept me sane every way she knew how— commiserating about the trials of research, encouraging me, and sharing her love with me. Roisin, Nina, Susan, Mike and Justin kept me connected to the real world and showed me all the value of having reliable, deeply considerate friends. Without any of these people, reaching this point in my education would have been a great deal more challenging.

I also owe thanks to the faculty of the Cognition and Cognitive Neuroscience area in theUniversity of Michigan Psychology Department. I always felt that the faculty rallied around their students. While I didn't have frequent contact with everyone, it was always evident that those around me had a personal interest in seeing every student develop and come into their own as an academician. In particular, though, I owe an immense debt to Rick Lewis. He set an example as a scrupulous, diligent and careful scientist, and his patient guidance helped me persevere even when the disparity between my own efforts and his was embarrassingly obvious.

Finally, I would like to thank my gorgeous dog, Lemon. She is always happy to see me, her motives are always transparent, she is a skillful communicator, and her tail wags at a comical angle.

To all of these people, and the aforementioned canine: Thank you so much.

# TABLE OF CONTENTS

# LIST OF FIGURES

**Figure**

# LIST OF TABLES

**Table**

# CHAPTER I

# Introduction

Describing the role of short-term memory in language comprehension has been a major goal of cognitive and psycholinguistic research for several decades. This is largely because limitations on the ability to maintain linguistic representations in memory seem to impose a natural limit on one's ability to piece together the elements of a sentence— which arrive not all at once, but serially over the span of several seconds.

Past models, including Baddeley's highly influential multi-store model (Baddeley, 1986), focused primarily on how much would "fit" in working memory. The prevailing metaphor for short-term memory[1] was that of a fixed-capacity buffer entirely separate from long-term memory, and research tended to focus on variations in how performance suffered when individuals were asked to store more than their memory buffer would hold (Just & Carpenter, 1992; Daneman & Carpenter, 1980; Logie, Della Sala, Laiacona, Chalmers, & Wynn, 1996).

These models have been usurped in recent years by others that differ in several ways. One of the principal differences is that the new models focus less on lim-

[1] I will use short-term memory in the way that "working memory" is currently used conventionally, to describe a subset of representations in memory that have recently been activated and have some residual level of activation, rather than in the sense that Atkinson and Shiffrin (1968) use it. Below I describe how Atkinson and Shiffrin's conception of short-term memory as a separate memory store differs from contemporary theories that tend to use the term working-memory instead of short-term memory.

itations to the storage capacity of short-term memory and more on the dynamic processes through which information enters short-term memory, how it is accessed in short-term recall, and how accessibility degrades. This dissertation is primarily concerned with memory *retrieval*, and will focus on how the accessibility of an item in short-term memory can be degraded between its most recent retrieval and a subsequent retrieval, and whether linguistic expectation can mitigate these effects at a critical point of retrieval. To be explicit, I will discuss two sources of difficulty in the retrieval process: similarity-based interference (Gordon, Hendrick, & Johnson, 2001; Van Dyke & McElree, 2006; Lewis & Vasishth, 2005) ) and time-based decay (Gibson, 2000; Lewis & Vasishth, 2005) (of activation on the target item's representation). The emphasis of this the following experiments rests heavily on similarity-based interference.

Chapter III includes four experiments aimed at strengthening the empirical basis for a behavioral result often linked to short-term memory restrictions: the so-called locality effect. The essence of the locality effect is that dependency integration becomes more difficult as the distance (in time or amount of linguistic material) between a head and dependent increases. For instance, the subject-verb relation between *nurse* and *supervised* should be integrated more quickly in *The nurse supervised the administrator*, where no words intervene, compared to *the nurse who was from the clinic supervised the administrator*.

If locality effects reflect difficulty in completing memory processes, integrating *any* dependency should become harder as more linguistic material (or more time) is inserted between the head and dependent. Establishing dependencies like the subject-verb relation just illustrated is required in any grammatical sentence, regardless of its syntactic complexity of the characteristics of the words within it.

Many theories of parsing, notably including memory-based models, assume in various degrees of explicitness, that locality effects generalize to any dependency relation. However, as I describe in Chapter III, there is surprisingly little evidence that locality effects generalize across languages, syntactic constructions, or even from offline measures like sentence complexity ratings to on-line measures like self-paced reading and eyetracking. I will present four experiments— two in self-paced reading and two eyetracking studies— extending the results of a previous self-paced reading study, which tested locality contrasts in both simple and syntactically complex sentences (Grodner & Gibson, 2005). That study yielded null findings when testing locality effects in simple sentences, highlighting the need for definitive evidence regarding the ubiquity of locality effects. I will present evidence for locality effects even in syntactically simple sentences in eyetracking as well as in self-paced reading when lexical processing difficulty is reduced. These experiments find locality effects in both simple and complex sentences where they they are predicted by memory-based theories of parsing, but not by existing implementations of experience-based theories (which predict difficulty with unexpected input) or other accounts that attribute the locality effect to complexities inherent in sentences requiring argument movement.

Chapter IV will examine the interplay between similarity-based interference and intra-sentential expectations. The theoretical motivation for these experiments lies in the fact that both retrieval interference and expectations are hypothesized to have behavioral effects by modulating how quickly previously mentioned referents can be accessed and integrated into an ongoing parse. Expectations are thought to speed-up lexical access at the point of retrieval by pre-activating a lexical representation (Ehrlich & Rayner, 1981; Schustack, Ehrlich, & Rayner, 1987; Rayner & Well, 1996), while interference slows retrieval by simultaneously making the tar-

get representation less active and making similar distractors more active. Whether expectations have any impact on interference-resolution, or vice versa, is not clear. The experiment in Chapter IV will be the first experimental test of possible interactions between similarity-based interference and linguistic expectation. I present an English language eyetracking experiment (Experiment 5) that gives an empirical basis for building models of parsing that can encompass both expectation effects and interference effects. In that experiment, the distance of the critical argument-verb dependency is kept constant to insure that the behavioral effects we observe do not reflect the impact of activation decay.

## Structure of the dissertation

The remainder of the dissertation is structured as follows: First, I will provide a brief review of working memory retrieval effects in parsing. Then I will set the stage for the experiments of Chapter III by marking their place in theoretical context. A brief review of expectation effects will follow, tying together several strands of evidence under the unifying model of surprisal; then I will present new evidence from an eyetracking experiment testing the interaction of retrieval interference and expectation-based facilitation in sentence processing.

# CHAPTER II

# Theoretical Background

## Short-term memory in language processing

Sentence comprehension requires establishing many relations between current and past linguistic input. Because words are presented serially in both speech and text, even simple structure like the dependency between a modifier (e.g., *tall*) and a noun (*tree*) requires short-term maintenance of linguistic representations (in this case, maintaining the modifier so it can be interpreted as the modifier of *tree*). The necessity of short-term memory for such basic operations in parsing has inspired a long tradition of empirical work in psycholinguistics.

For many years, short-term memory's involvement in language comprehension was in some way influenced by Baddeley's tri-partite model. This model consisted of a "visuo-spatial sketchpad" to hold visual information, a phonological loop that could retain approximately two seconds of auditory information through active rehearsal (Baddeley, Thomson, & Buchanan, 1975), and a separate central executive that monitors and controls the contents of the verbal and visual buffers. Baddeley's model was also one of several multi-store models inspired by neuropsychological evidence that suggested short-term memory and long-term memory were completely separable systems, each with their own representations (Shallice & Warrington, 1970; Vallar &

Papagno, 2002; Baddeley & Warrington, 1970).

The Baddeley model impacted language research by offering an explanation as to why phonological similarity caused forgetting in verbal tasks, and why longer words in lists reduced verbal recall span (Logie et al., 1996). However, the theoretical foundation of multi-store models was weakened by convergent experimental evidence suggesting that the behavioral and neuroimaging data from studies of short-term and long-term memory could be reconciled in a model that assumes only one set of long-term representations (Jonides et al., 2008), with a recently retrieved subset of those representations entering short-term memory by virtue of elevated activation.

### *Current memory models in language processing*

In addition to viewing memory as one unitary resource, contemporary theories no longer emphasize the role of dedicated buffers responsible for storing input from each sensory modality. This shift was motivated by evidence from a number of studies indicating that short-term representations across different modalities were subject to the same set of systemic constraints, even though various underlying representations may have been distributed across different perceptual systems (Jonides et al., 2008).[2]

With attention shifting away from these two architectural claims— modality-specific buffers and separable systems for long-term vs̈hort-term memory— the dominant paradigm in memory research was supplanted by a new class of contemporary memory models. The new paradigm is characterized by several architectural features, including

(a) fast, content-addressed retrieval, (b) a very limited focus of attention, (c) a focus on domain-general constraints on encoding, storage, and retrieval, (d) little

[2]The authors of Jonides et al. (2008) do point out that, while contemporary memory theories do not posit modality-specific handling of information in memory, those theories do admit modality-specific processing of low-level sensory input in posterior regions of the brain.

focus on modality-specific buffers, and (e) a unified memory store.

Theories that emphasize the importance of retrieval processes build upon evidence from speed-accuracy tradeoff (SAT) studies showing that retrievals can be executed quickly enough to support rapid on-line comprehension (McElree, Foraker, & Dyer, 2003), and that the contents of short-term memory can be searched in parallel using specific stimulus features as cues to identify a retrieval target (McElree, 2000). Similarity-based interference occurs when more than one item in memory matches a retrieval cue. Distractors that were processed before the target item cause *proactive* interference, while distractors processed after the target cause *retroactive interference*. Additional retrieval cues must be used to discriminate the target from all competitors. By hypothesis, resolving interference takes time, and it becomes less accurate when distractors match the retrieval target on multiple features. The processing cost of resolving interference should be reflected in increased reading times or higher error rates at the point in the sentence that triggers the retrieval.

There is some theoretical dispute over what types of features are used as cues for retrieval. Under one model (Lewis & Vasishth, 2005), a limited number of specified syntactic cues are used to identify retrieval targets, including number (singular or plural), gender, and syntactic category. That model is not theoretically limited to such a small set of cues. The authors' model stipulates only that interference occurs when the retrieval cues set by a word match the features of more than one recently processed word that might be retrieved. There is no stipulation that the features used as retrieval cues must be limited to syntactic cues of any particular type. They could conceivably be semantic sues or even (hypothetically) phonological or orthographic cues. Whatever features the reader or listener actually uses to recognize the appropriate retrieval target counts as a retrieval cue. Presumably, however, the cues

set by the reader/listener are constrained by the semantic and syntactic context of what has already appeared in a sentence or in prior discourse, excluding recently attended words that might incidentally share features that would not effectively discriminate between, for instance, an appropriate subject and an inappropriate subject for a verb. In other words, as Lewis and Vasishth (2005) claim, retrieval is a skilled process.

Other descriptions of similarity-based retrieval interference (Gordon, Hendrick, Johnson, & Lee, 2006; Van Dyke & McElree, 2006), built upon the Search of Associative Memory (SAM) model (Gillund & Shiffrin, 1984), do not explicitly restrict interference to a set of features that have been selected as relevant identifiers of a target. In many respects, these models are compatible with the ACT-R based model just described. The critical difference is that they remain agnostic about the role of a weighting parameter, $w$, that Gillund and Shiffrin (1984) included in their model. In effect, the $w$ parameter discriminated between relevant and irrelevant retrieval cues by assigning weights to each feature according to their importance, thereby accomplishing what the skilled reader does in the ACT-R model. Van Dyke and McElree (2006) and Gordon et al. (2006) do not clearly state whether this parameter plays any role in their models of interference. Depending on how they intend to treat this weighting factor, their models may or may not differ substantially from the description of retrieval outlined in the ACT-R model. If they intend to drop the weighting factor, then their models would predict that any item in short-term memory with a strong association to the retrieval-triggering word (like a verb) can compete with other items with similar features. As a result, interference would occur more frequently than in the ACT-R model. If, on the other hand, the weighting parameter is assumed to play a role, then not every associative link between a retrieval-triggering

word and another item in short-term memory will cause retrieval interference. Under the reasonable assumption that a skilled reader or listener would assign more weight to features that help identify the appropriate item for retrieval, the models of Van Dyke and McElree (2006) and Gordon et al. (2006) would become almost indistinguishable from the ACT-R model.

None of the extant research on retrieval interference in sentence comprehension tasks effectively discriminates between models that include feature weighting and those that exclude it. The experiment in Chapter IV will take a first step in this direction. In the meantime, there is ample evidence that retrieval interference does adversely affect comprehension.

There is evidence that retrieval interference results from shared syntactic cues (Van Dyke & Lewis, 2003; Van Dyke, 2007) and semantic cues (Ehrlich & Rayner, 1981; Schustack et al., 1987; Rayner & Well, 1996). I will focus, for now, on the evidence for lexical semantic interference because the research proposed in later sections involves semantic interference, and not syntactic interference.

Semantic interference effects have been found in self-paced reading (Gordon et al., 2001; Van Dyke & Lewis, 2003; Vasishth & Lewis, 2006; Van Dyke, 2007) and eye-tracking (Gordon et al., 2006; Van Dyke, 2007) experiments, including both within-sentence manipulations (Van Dyke & Lewis, 2003; Gordon et al., 2001) and tasks that employ an external memory load (Gordon, Hendrick, & Levine, 2002; Van Dyke & McElree, 2006). Gordon et al. (2001) found that the well-known difficulty contrast between object-relative clauses and (simpler) subject-relatives, shown in (1) was mediated by how much semantic overlap existed between two pre-verbal argument NPs. Semantic interference was found at verbs like *climbed* when both NPs were proper names or definite NPs, but not when the two NP arguments were of different

semantic types.

(1) a. The banker that (the barber/Ben) praised *climbed* the mountain.

b. The banker that praised (the barber/Ben) *climbed* the mountain.

In another paper, Gordon et al. (2002) found that a short list of noun phrases presented in a recall task before test sentences caused proactive interference with NPs of the same type (names or definitive NPs) in a sentence that followed. Reading times and comprehension question accuracy both reflected semantic interference effects when NPs in the memory list matched the type of NP in the following sentence. Van Dyke and McElree (2006) used a similar paradigm to extend these findings, using stimuli that varied specific retrieval cues within a sentence and a preceding memory list, rather than a categorical difference between NPs in different referential classes. As shown in (2), interference from a memory list of words like *table*, *sink* and *truck* varied with the identity of the verb (*fixed* or *sailed*). In this example, list words should not interfere with retrieval of the object, *boat* at *sailed* because tables, sinks and trucks are not plausible objects and therefore lack the semantic retrieval cues used to select *boat* from memory. Since the list words are all plausible objects of *fixed*, semantic interference should increase reading times at that verb. The key result was the interaction between memory load and sentence type, demonstrating that cue overlap between list items and the object of the sentence caused increased reading time at the verb.

(2) a. LIST: [ table sink truck ] *or* [ *no list* ]

b. INTERFERING It was the boat that the guy who lived by the sea *fixed* in two sunny days.

c. NON-INTERFERING It was the boat that the guy who lived by the sea

*sailed* in two sunny days.

Retrieval-based theories can also account for a range of other phenomena, including limits on our ability to comprehend center-embedded relative clauses and certain cases of ambiguity(Lewis, 1996) as well as slowdowns and erroneous grammaticality judgments in negative polarity-licensing (leading readers to accept sentences like, *A man who had no beard was ever happy*)(Vasishth, Brüssow, Lewis, & Drenhaus, 2008).

Evidence for decay as an important determinant of comprehension difficulty is less compelling. In part, this is because studies targeting memory retrieval processes have unable to fully dissociate decay from interference effects (Berman, Jonides, & Lewis, 2009). Since decay is a simple function of how much time has passed since retrieving an item from memory, an ideal test might insert an interval with no stimulus or a distracting beep between a dependent and head, varying the duration of the interval to directly manipulate decay. The effect of decay on retrieving the pre-interval dependent could be measured at the post-interval head. Vasishth (2004) ran a similar experiment, inserting either an adjunct or silence before a verb in Hindi relative clauses. Reading times at the verb were longer after a silent interval than after an adjunct, suggesting that activation decay (rather than, in this case, the introduction of a new discourse referent) creates difficulty integrating a verb dependency.

### *Locality effects as an important prediction of memory-based theories*

Long-distance dependencies have long been regarded as important examples of how short-term memory is required for language comprehension. The intuition is that

linguistic dependencies, like argument-verb relations, are harder to integrate when the head and dependent are separated by intervening clausal material, compared to when the dependency can be resolved locally (as illustrated by the contrast in (3)). The increased difficulty created by splicing linguistic material between a dependent and head is called a *locality effect*.

(3) a. The nurse supervised the administrator.

b. The nurse who was from the clinic supervised the administrator.

This intuition inspires several theoretical proposals, from the Late Closure heuristic (Frazier, 1987) to the Active Filler hypothesis (Frazier & Clifton, 1989), and is formalized in Dependency Locality Theory (DLT) (Gibson, 1998). DLT predicts a monotonic increase in comprehension difficulty with every new discourse referent that is mentioned between a dependent and head. In DLT, however, the mechanistic underpinnings of the locality effect are not well-defined; the roles played by interference, decay, or other hypothetical properties of short-term memory, are left unspecified. Rather than hypothesizing about the internal structure of short-term memory, DLT follows many other extant memory theories in assuming that short-term memory function is governed by a static capacity limit (Just & Carpenter, 1992; Gibson, 1998), rather than constraints on the efficiency of retrieval processes (Lewis & Vasishth, 2005).

Although the mechanism underlying locality effects has not been explicated in most previous theories, Lewis and Vasishth (2005) demonstrated that locality effects are naturally predicted by activation decay. On the other hand, Lewis, Vasishth, and Van Dyke (2006) suggest that locality effects could also be the result from retroactive interference caused by referents introduced within the dependency. This is not

necessarily the same claim made by DLT, because under DLT the locality effect results from the number of new discourse referents introduced within the dependency, regardless of their semantic features or how they fit into surrounding syntactic structure.

If locality effects reflect difficulty accessing decayed targets, they should be observable in all languages, across a broad range of syntactic constructions. If they reflect retroactive interference from post-target distractors, they should vary in magnitude depending upon the extent of feature overlap between the target and subsequent distractors.

The next chapter describes how existing evidence for locality effects is linguistically and methodologically narrow— perhaps worryingly so, given the substantial role they have played in shaping parsing theory. That section will explain that on-line observations of locality effects have been restricted to syntactically complex structures involving argument movement, and that these observations fail to fully support the hypothesis that locality effects result from memory limitations. I will follow-up by describing four completed experiments that search for distance effects that are independent of structural complexity, independent of lexical processing difficulty, and independent of known sources of retrieval interference. The results of these experiments show that, when there is no obvious source of retroactive interference from material in the long-distance dependency, locality effects still exist.

## Expectation in short-term memory

While short-term memory's role in comprehension can partially be understood as looking *back* to unify past input with new input, there is abundant evidence for expectation-based processes that look *forward* to facilitate processing of new input as soon as it is perceived (or even earlier). These processes exploit the comprehender's knowledge of distributional frequencies in their language, at many levels, allowing some phase of processing (most likely recognition, as I describe later) to proceed faster for highly expected input than for less expected input.

I will refer to the speed advantage gained through these processes as *expectation effects* —not assuming that there exists a conscious or controlled search of memory to calculate features of upcoming input, or that any kind of short-term, integrative representation can be constructed on the basis of input that has yet to be confirmed by a written or spoken word— but to capture the idea that language comprehension is in some way biased to process some words more easily than others by the comprehender's implicit knowledge of how a sentence might unfold.

Two types of representation must be accounted for in theories of expectation in comprehension. There are stable representations in long-term memory that include lexical entries and their associated features (like physical properties, semantic categories, and arguably syntactic information); and there are short-term, constructed representations that bind together recently processed representations into a common thematic or propositional structure. Parsing requires taking advantage of both types of representations. Recognizing a word requires mapping a percept to stable, long term orthographic and phonological representations that allow the word's meaning to be retrieved from long-term memory. Parsing a sentence requires integrating several words to create a proposition that is greater than the sum of its parts. The integrated,

ongoing parse is more than a juxtaposition of retrieved lexical entries, however; it includes thematic content that must be composed on the fly, then maintained and updated in short-term memory. A subset of experience-based theories that I will call "expectation-based" theories make predictions about how probabilistic knowledge gained through learning can be applied through anticipatory or "expectation-based" processing.

Most accounts of expectation-based facilitation in language comprehension hypothesize that predicted words are understood faster because their long-term lexical representations are activated by earlier parts of the sentence, either through direct lexical association with individual words, or through "top-down" activation by the short-term, contructed representations that represent syntactic and semantic relations at the discourse level.

A great deal of research has been devoted to former claim. These studies demonstrated that lexical recognition is faster following a semantically related prime than an unrelated word(e.g., Meyer, Schvaneveldt, & Ruddy, 1975) Semantic priming of this sort has been attributed to spreading activation from the cue word (also known as the priming word) "pre-activating" the target word. This type of lexical-lexical priming may have much in common with the phenomena that are targeted in Chapter IV of this thesis. For instance, a number of supra-lexical cues may activate a lexical entry in the same way that a prime word does, but therein lies the difference between what I will call "expectation" and what has classically been discussed in priming research: in Chapter IV, I will set aside cases where the priming context is a single word immediately preceding the target word, examining the influence of expectation only at points where an entire phrase before the critical word is identical across levels of expectation, making lexical-lexical priming a highly unlikely source

of facilitation. In these cases, a broader context including multiple words and/or relations must conspire to create an expectation.

There is an empirical basis for distinguishing between my operationalization of expectation and the classic semantic priming effect. Although Kintsch (1988) argued that "the discourse context is actually irrelevant to the priming effect" (p. 171), others have carefully delineated the differences between semantic priming and what have been variously referred to as "holistic", "contextual", "situational", and "inferential" factors that draw on composed semantic structures or the "gist" of a sentence to facilitate lexical processing ( see Sharkey and Sharkey (1992); Schustack et al. (1987), and (Rayner & Well, 1996) for reviews of the topic). Foss (1982) and Foss and Ross (1983) may have been the first to draw a distinction between associative (lexical – lexical) priming and discourse priming (which implies some top-down effect from more-composed representations to less-composed ones like words). Around the same time, Gough, Alford, and Holly-Wilcox (1981) found experimentally that even a single word intervening between a prime and target can disrupt associative priming effect, casting doubt on whether simple associative priming could be sustained over the course of a sentence. So-called discourse priming effects, on the other hand, can be sustained across several interrupting, unrelated items (e.g., Foss (1982). These studies bolster the popular claim that lexical facilitation can occur not only through direct associative priming, but as the result of top-down activation — i.e, by propagating activation from some level of propositional or syntactic representation down to the target word (and possibly numerous other, related candidate words); the latter phenomenon is what I am interested in examining.

Priming and expectation effects are still joined by the idea that facilitation results from pre-activating a word, whether from another lexical entry in the long-term men-

tal lexicon or from a set of short-term representations constructed through parsing the current sentence. The pre-activation view of expectation-based facilitation has been around for decades (e.g., Posner and Snyder (1975a, 1975b)), and is the default model in the literature on context effects. It is also the model that I will adopt for much of this paper.

If, in fact, lexical recognition is facilitated via the spread of activation across strongly associated representations, associative strength would vary depending on how often any associative link is actually used (Hebb, 1961). From this perspective, expectation-based facilitation can be seen as the language processor's way of exploiting implicit knowledge of statistical regularities or "distributional frequencies" in a language. Below, I outline a proposal concerning the nature of the relationship between distributional frequencies and processing facilitation. In the meantime, however, this brief description provides some initial basis for drawing together some of the evidence of expectation effects in comprehension.

So, what evidence is there that readers actually form and use linguistic "expectations"?

There are several lines of work demonstrating expectation-based facilitation at lexical, semantic and syntactic levels. They have used various means of estimating expectation, including corpus-based estimates of transitional probability; Cloze probability; and probability computed over features that are tagged in corpora. Each of these methods predicts processing facilitation when a new word is highly probable in the context of constraining prior input.

At the lexical level, S. MacDonald and Shillcock (2003a) (and S. MacDonald and Shillcock (2003b)) demonstrated that a single word can create expectancy for the word that follows. They showed that bi-gram probability, defined as the probability

of word *w* following word *w-1*, reliably predicted variance in fixation times. Several others have demonstrated semantic expectation-based facilitation on a target word. (Morris, 1994) found shorter first-fixation and gaze durations on "moustache" in sentences like, *The barber trimmed the moustache this morning* compared to *The person trimmed the moustache this morning*. Morris' stimuli varied semantic constraint without using an empirical measure of the critical word's probability in context; however, other studies have estimated the expectedness of a key word using some form of Cloze norming, in which participants read or hear a partial sentence and are asked to provide the next word.

Ehrlich and Rayner (1981) and Schustack et al. (1987) found that faster lexical naming, shorter reading times and higher skipping rates at a direct object were predicted by its Cloze probability. In a modified Cloze task asking participants to vocally produce the object after silent, self-paced reading of a passage, Schustack et al. (1987) also found that responses with high Cloze probability (across all participants) were produced faster than responses with lower Cloze probability. Several other studies have confirmed the basic findings of Ehrlich and Rayner (1981) and Schustack et al. (1987), demonstrating a reliable, negative correlation between the Cloze probability of a word, given its semantic and syntactic context, and online processing measures (Rayner & Well, 1996; Ashby & Rayner, 2005; Frisson, Rayner, & Pickering, 2005).

Expectations have also been hypothesized to cause facilitation in some recent investigations of long-distance dependencies. Konieczny (2000) found facilitation at a clause-final verb in German sentences when the verb was separated from its arguments by a large prepositional phrase rather than a short prepositional phrase. One explanation that has been proposed for this result, and a similar "anti-locality"

effect found by Jaeger, Fedorenko, and Gibson (2005) in English, is that additional clausal material preceding the verb sharpen the syntactic (and possible semantic) expectation for a verb.

Levy (2008) has shown that anti-locality effects are one phenomena that can be predicted by surprisal theory (Hale, 2001). Surprisal is a formal model that predicts local processing effects as a function of how unexpected or surprising a word is in a given context. Jurafsky (2002) expresses surprisal as:

$$Surprisal = -logP(w_i) \tag{2.1}$$

The conditional probability of a word, P($w_i$), can be written as:

$$P(w_i|w_0, ..., w_{i-1}) = \frac{P(w_0, ..., w_i)}{P(w_0, ..., w_{i-1})} \tag{2.2}$$

Surprisal, a measure used to estimate processing difficulty, is the negative log of this probability. According to surprisal, then, words that are unexpected have high surprisal, and are more difficult to process. As each word is added to a parse, the absolute likelihood of encountering the parsed string (the numerator in equation 2.2) becomes smaller. For example, the probability of encountering a string like, "The boy..." may be relatively high because, for various reasons, this type of string with an animate noun beginning the sentence is common in English. However frequently a reader might have seen a string like this, whatever partial parse is constructed to include the next word will necessarily have been seen less often. Thus, the absolute probability of the partial parse at "boy" will be larger than the whole string's probability of occurrence at the next word, regardless of the next word's characteristics. The next word's properties, including syntactic category, can be very important, though. If "The boy..." is followed by "and", the probability of encountering the

string may be less compared to the same string continued with "hit", because the adjacent subject-verb structure occurs more frequently than a conjunctive phrase at the beginning of the sentence.

From this perspective, difficulty can be understood as the product of ruling out some portion of possible continuations. When the possible continuations disconfirmed by a new word (e.g., "The boy" + *verb*) represent a large amount of probability mass, surprisal is high. At each word, the magnitude of change in the probability of the parser's current state corresponds to the change in difficulty. That ratio, measuring change in probability, will always be less than one because the probability of a parser state necessarily decreases with each new word. Even so, the relative magnitude of change (compared between two conditions of an experimental sentence) should correlate with the relative expectedness and the relative difficulty of processing each new word. In a similar way, the probability of a participant seeing any given string decreases, but he or she does not experience monotonically increasing difficulty as every sentence unfolds. This would mean, for instance, that one would be unable to gather meaningful Cloze ratings because they would vary exclusively with sentence length. Fortunately, participants do not experience difficulty strictly as a function of adding a word. Differences between sentences of equal length are measurable in Cloze ratings, just as differences in the magnitude of probability change can be measured in surprisal.

High surprisal values can be used to predict on-line behavioral indicators of difficulty, like increased fixation time or longer reaction time latencies in self-paced reading. Conversely, low surprisal values can be used to predict processing facilitation relative to an appropriate control. Levy (2008) showed that an Earley parser that computes surprisal over syntactic category tags in a corpus can predict the

anti-locality effects in Konieczny (2000) and Jaeger et al. (2005).

As Levy also points out, however, surprisal theory generalizes over any type of representation from which expectation might be constructed. Surprisal can be calculated over any cue or feature for which probabilities can be estimated. This makes surprisal a convenient framework for understanding all the aforementioned evidence of expectation effects. Bi-gram probability, for instance, can be viewed as a restricted application of surprisal, where the relevant unit of expectation is a word token (rather than syntactic category or another type of representation), and the probability of the current input word is conditioned upon a single lexical item rather than an entire sentence prefix.

Surprisal can also be used to describe Cloze predictability experiments. Levy (2008) posits, in fact, that surprisal could be approximated as something very close to negative log Cloze probability. This is convenient because Cloze probabilities are easy to obtain; they do not require massive, hand-annotated corpora. They are also face-valid indicators of lexical expectation, generated through the same probabilistic knowledge participants naturally use during production and comprehension. In Chapter IV I will measure surprisal at the lexical level using Cloze predictability.

### *Connecting surprisal and interference*

Experiments showing that low surprisal (or strong expectation) predicts localized processing facilitation is a particularly interesting counterpoint to the evidence linking processing difficulty to similarity-based interference. There is empirical evidence that interference slows comprehension, while expectation can facilitate comprehension; yet there is no existing model that incorporates both memory effects and surprisal. Before examining the relationship between interference and expectation, it

is worth considering whether extant empirical evidence unequivocally indicates two distinct mechanisms underlying them.

Surprisal effects can be dissociated from interference effects. The structures used by Konieczny (2000) and Jaeger et al. (2005) effectively accomplish this. In sentences with a clause-final verb following a subject with increasingly long PP-modification (see Figure 2.1), surprisal predicts that the verb *escorted* should be read faster in (c) than in (a). Konieczny (2000) and Jaeger et al. (2005) found anti-locality effects in these sentences, where there were no differences in the number of potentially interfering referents across conditions.

a. Er hat den Abgeordneten begleitet, und ...
   He has the delegate escorted, and ...
   "He escorted the delegate, and ..."
b. Er hat den Abgeordneten ans Rednerpult begleitet, und ...
   He has the delegate to_the lectern escorted, and ...
   "He escorted the delegate to the lectern, and ..."
c. Er hat den Abgeordneten an das große Rednerpult begleitet, und ...
   He has the delegate to the big lectern escorted, and ...
   "He escorted the delegate to the large lectern, and ..."

*Figure 2.1.* : An example sentence from Konieczny, 2000.

Similarly, surprisal cannot fully account for interference effects. The clearest dissociating evidence for interference comes from sentences that manipulate retrieval interference using an external list (Gordon et al., 2006; Van Dyke & McElree, 2006). In particular, Van Dyke and McElree (2006) found that while verb-reading times in *It was the boat that the guy who lived by the sea [ fixed / sailed ] in three sunny days* were equal with either *sailed* or *fixed* as the verb, but *fixed* was 38 milliseconds slower than *sailed* when participants first read (and were asked to remember) a list of

three fixable objects like *table - sink - truck*. While it is difficult to know what existing computational models of surprisal will predict (because implementing surprisal requires committing to a particular grammar formalism and parsing algorithm), it would be extremely counterintuitive, if not impossible, to predict this interaction under surprisal. Because the memory list has no syntactic relation to the sentence, the list words cannot contribute to a syntactic expectation for the verb. The only source of expectation that could explain differences between *fixed* or *sailed* would be semantic associations between the list words *table sink truck* and the verb. These associations cannot plausibly cause facilitation at *sailed* because tables, sinks and trucks cannot be sailed. They can, however, be fixed. For this reason, surprisal would predict no difference between *fixed* and *sailed* as a result of adding a memory list beforehand— or else it would predict the opposite of an interference effect: facilitation at *fixed* compared to *sailed*.

While extant studies show independent effects of interference and expectation at points of dependency resolution, there has been no empirical observation of how they jointly affect comprehension. This is unfortunate, because examining the interplay of interference-based difficulty and expectation-based interference can furnish answers to several important questions about expectation. To name a couple:

- Can semantic expectations for heads— and verbs in particular— compensate for retrieval difficulty caused by interference?
- Can semantic expectations exacerbate interference effects by keeping interfering items active alongside appropriate retrieval targets?
- Can pre-activation of a lexical item by surrounding semantic context cause it to be retrieved before the written or spoken word is perceived (Lau, 2009)?
- If highly expected words like heads can be pre-retrieved, can they be integrated

into dependency relations without being physically perceived?

Answering these questions would contribute substantially to refining existing accounts of expectation-based processing. Moreover, modeling the interplay of interference and expectations will help develop more complete models of activation dynamics in short-term memory. The experiment in Chapter IV is the first empirical attempt to address these questions directly by testing expectations and interference together.

*Summary*

The preceding sections described two theoretical goals for this dissertaion:

- Search for on-line locality effects where they cannot be explained by other sources of complexity, and
- Investigate the relationship between the cognitive processes underlying retrieval interference and semantic expectations.

Having summarized some background research on short-term memory effects and expectation effects, and having also put these objectives in theoretical context, I will now move on to more detailed descriptions of the experiments.

# CHAPTER III

## In search of on-line evidence of locality effects

### Motivation

I have already stated the argument that short-term memory is functionally integral to parsing, and that a long tradition of research has emerged from this observation. One of the most straightforward and theoretically influential empirical generalizations to emerge from this work is that the *locality* of linguistic relations, such as the subject-verb relation in (4) , is a primary determinant of the speed and efficacy of the short-term memory processes in parsing (Chomsky, 1965; Just & Carpenter, 1992; Gibson, 1998). More specifically, increasing the distance over which these relations must be computed degrades the underlying memory processes in some way. For example, the implication of this view for (4) is that the subject-verb relation in (4b) is more difficult to compute than the same relation in (4a).

(4) a. The manager unexpectedly quit her job yesterday.

b. The manager who the supervisor admired unexpectedly quit her job yesterday.

This theoretical view has been expressed most transparently in *Dependency Locality Theory* (DLT) (Gibson, 1998, 2000), which uses as a measure of locality the

number of new linguistic referents interposed between a dependent and its head. DLT claims that the degree of locality should be reflected in a continuous and monotonic way in on-line reading time measures, thus yielding testable empirical predictions. This general class of effects on reading times will be called *locality effects* for present purposes. While Experiments 1–4 present evidence that locality effects are consistent with memory-based parsing theories, the term "locality effects" will be used without intending to associate them exclusively with the details of DLT or any other specific parsing model. Locality effects are important and relevant to a very broad range of extant memory and parsing theories (see Lewis et al. (2006) or Gibson (2000) for a summary)—even those which do not have mechanisms in place to directly produce them.

The following experiments pursue three objectives. First, the case will be made that current empirical evidence for on-line locality effects is narrow both linguistically and methodologically, and perhaps surprisingly difficult to find under the assumption that locality is a ubiquitous factor in sentence processing. This argument raises the possibility that locality effects may be evident only in relatively complex structures whose difficulty may be traceable to independent factors. If this is the case, it has major implications for how these phenomena bear on theory development.

Given the key role that locality effects play in shaping current parsing theory, it is important to significantly broaden its base of empirical support, and this relates to the second and third aims of this chapter. The second aim is to extend locality investigations to include eyetracking measures. This chapter will show that eyetracking has advantages over self-paced reading (SPR) for investigating locality effects.

Furthermore, Experiments 1–4 adopt the approach of running parallel experiments with identical materials in both paradigms. This facilitates efforts to develop

detailed theories of the link between the underlying short-term memory processes and the control of eye-movements and button-presses, and therefore the relationship between SPR and eyetracking as empirical measures. Such a model is not defined here, but a substantial first step has been taken in this direction in Bartek, Lewis, Vasishth, and Smith (2011).

The third objective of these experiments is to demonstrate (possibly more subtle) locality effects using linguistic material that is, overall, significantly easier to process than materials that form the basis of existing locality demonstrations, thus providing stronger evidence for the claim that locality exerts pervasive and continuous effects on sentence processing.

The remainder of this chapter will first assess the current evidence for locality effects, and discuss its potential theoretical implications. Description of the design and results from four new experiments, which consist of two pairs of SPR and eyetracking experiments, will follow. Finally, I will discuss the theoretical and methodological implications of the results.

### *Assessing current empirical evidence for locality effects*

The existing empirical evidence for locality effects is surprisingly mixed. Locality effects have been found in studies of English sentences (as we summarize below), but *anti-locality* effects—faster processing in longer-distance dependency integration—have been found in head-final languages including German, Hindi and Japanese (e.g., Konieczny, 2000; Vasishth & Lewis, 2006; Vasishth, 2003; Nakatani & Gibson, 2008), as well as English (Jaeger et al., 2005). Although anti-locality effects place important constraints on psycholinguistic parsing theory—and it is important to assess theories of locality effects in their context—it remains possible that independent factors give

rise to both locality and anti-locality effects; they need not be mutually incompatible. The present experiments focus on obtaining a better understanding of the nature and extent of positive locality effects.[3]

Locality effects have been observed in both ambiguous and relatively unambiguous structures. In ambiguous structures, locality plays a role in both resolving ambiguities (Kimball, 1973; Frazier & Fodor, 1978; Grodner, Gibson, & Tunstall, 2002; Gibson, Pearlmutter, Canseco-Gonzales, & Hickock, 1996; Pearlmutter & Gibson, 2001; Gibson, Pearlmutter, & Torrens, 1999; Altmann, Nice, Garnham, & Henstra, 1998) and in garden path reanalysis (garden paths involving longer ambiguous regions are typically more difficult to recover from; Pritchett, 1992; Gibson, 1991; Van Dyke & Lewis, 2003; Ferreira & Henderson, 1991). While these results have yielded useful constraints on parsing theory (Lewis & Vasishth, 2005), the goal of these experiments is to understand and find evidence for on-line locality effects in (putatively) globally unambiguous structures. (The Discussion section will take up the issue of possible local ambiguity in our materials in some detail.)

### *Existing on-line locality effects are restricted to points of extraction*

Table 3.1 provides an overview of the existing experimental evidence for locality effects in relatively unambiguous structures. The evidence is restricted to English (a cross-linguistic gap that is not filled in by these experiments), and to points of extraction—more specifically, to relations conventionally analyzed as $\overline{A}$-movement (of an argument) from its canonical position (Mahajan, 1990). In particular, the evidence generated so far involves relative clauses that contain so-called "filler-gap"

[3] In other work, my colleagues and collaborators on Bartek et al. (2011) have outlined a theoretical model that provides an integrated explanation of both locality and anti-locality (Vasishth & Lewis, 2006; Lewis et al., 2006; Lewis & Vasishth, 2005)

Table 3.1:: Extant experimental evidence for locality effects in (relatively) unambiguous structures.

| Source | Linguistic structures | Methodology |
|---|---|---|
| Gibson (1998), Exp. 1 | subject- and object-relative clauses | self-paced reading |
| Grodner et al. (2002), Exps. 1 & 2 | reduced-relative ambiguities | self-paced reading |
| Gibson and Warren (2004), Exp. 1 | extraction across VP or NP | self-paced reading |
| Grodner and Gibson (2005), Exp. 1 & 2 | subject- and object-relative clauses | self-paced reading |
| Wu and Gibson (2008), Exp. 1 | subject- and object-relative clauses | self-paced reading |

dependencies (e.g., The *man* who the woman *liked*__), where the object has been displaced from its canonical position after the verb to the beginning of the sentence. It has been speculated in Grodner and Gibson (2005, p. 284) and elsewhere (Gibson, 2007) that $\overline{A}$-movement may be an important condition for the occurrence of locality effects.

Given this restricted evidential base, there are two plausible accounts for the locality effects that have been obtained experimentally. Locality effects may be a direct result of the degradation of memory representations between initial activation and subsequent retrieval for integration into a dependency, which would imply ubiquity of the effects. Alternatively, locality effects could reflect a source of difficulty unique to structures that require $\overline{A}$-movement, such as object-extracted relative clauses. [4] Prior experiments that could have determined if locality effects generalize beyond object relatives, and beyond movement, have yielded ambiguous results.

The nature of the existing evidence can be understood by considering three of the

[4] Although most theories of short-term memory in sentence processing do not distinguish the computational demands of movement and non-movement relations, there is a line of work that does make such a distinction, starting with the *Hold Hypothesis* in the augmented transition network (ATN) model of Wanner and Maratsos (1978), and continuing with the Grodzinsky (2000) theory of neural processes associated with syntactic movement operations.

experimental conditions in Grodner and Gibson (2005) Experiment 2 (underlining is used here to indicate the word at which the locality effects are predicted to be observed). Note that, in these sentences, $\overline{A}$ movement occurs when the object is moved from its base position (adjacent to the embedded verb) to the beginning of the sentence.

(5) *Embedded verb conditions from Grodner and Gibson (2005) Experiment 2*

a. The administrator who the nurse supervised scolded the medic while . . .

b. The administrator who the nurse from the clinic supervised scolded the medic while. . .

c. The administrator who the nurse who was from the clinic supervised scolded the medic while. . .

In all three structures in (5a), the region of interest is the embedded verb *supervised*, and the locality manipulation involves increasing the distance from the embedded verb to its subject (*the nurse*) and its extracted object (*the administrator*). In (5a), no material intervenes between the embedded verb and the subject; in (5b), a three word prepositional phrase (PP) intervenes; and in (5c), a five word relative clause (RC) intervenes. The structure of this design is shown schematically in (6). The top arrow denotes the relation between the verb and the relative pronoun *who* that mediates the object extraction, and the bottom arrow denotes the subject relation. The ∅ symbol denotes the null string (nothing interposed).

(6) *Structure of the embedded verb conditions from Grodner and Gibson (2005)*

The administrator who the nurse $\left\{ \begin{array}{c} \emptyset \\ \text{from the clinic} \\ \text{who was from the clinic} \end{array} \right\}$ supervised. . .

The assumption (as expressed, e.g., in DLT) is that the computation of these

dependency relations happens immediately at *supervised* by accessing short-term memory representations associated with the relativizing pronoun and the subject,[5] and that this computation takes longer as the input items that triggered the target representations become more distant. Thus, the straightforward prediction is that reading times at *supervised* should increase monotonically in the three conditions (nothing interposed, PP interposed, and RC interposed). This prediction is consistent with what Grodner and Gibson (2005) found in their Experiment 2 using self-paced reading, with the sharpest increase in reading times observed for the RC condition (we discuss the empirical results in more detail below). This manipulation has the attractive property that the specific verbs in the critical region and the head nouns of the target subject and object noun phrases are kept constant while changing the locality of the relations.

But the reliable locality effect observed in (5) may have been driven entirely by the sharp increase in reading times for condition (5c): a case of double center-embedding of relative clauses, an effect that can be explained in ways that have nothing to do with locality (e.g., similarity-based interference, Lewis & Vasishth, 2005). How can we be sure that the observed effects in (6) generalize beyond object extractions over embedded relative clauses? We can compare the effects in (6) above to three other conditions in Grodner and Gibson (2005):

(7) *Matrix verb conditions from Grodner and Gibson (2005) Experiment 2*

a. The nurse supervised the administrator while ...

b. The nurse from the clinic supervised the administrator while ...

c. The nurse who was from the clinic supervised the administrator while ...

[5]There are further important distinctions to be made here about the nature of these representations—whether they involve *predictions* of the verb (Lewis & Vasishth, 2005; Gibson, 2000) the degree to which they are semantic (Van Dyke, 2007), etc.—but these distinctions are not relevant for present purposes.

These three conditions test for locality effects at a matrix verb from which no arguments have been extracted; the only linguistic relation affected by locality is the subject relation. The structure of the main verb conditions is shown schematically in (8):

(8) *Structure of the matrix verb conditions from Grodner and Gibson (2005) Experiment 2*

The nurse { ∅ / from the clinic / who was from the clinic } supervised the administrator...

If a locality effect is observed at *supervised* in (8), this would provide evidence that dependencies that are *not* the result of $\overline{A}$-movement relations are also subject to locality effects.[6] In other words, the presence of such effects in both kinds of structures would mean that increasing locality increases the processing cost of resolving simple subject-verb dependencies as well as object extractions. Figure 3.2 (upper left) shows the readings times observed by Grodner and Gibson (2005) at the critical verb. (This figure also contains the reading times for the four experiments in this paper, but the reader should focus for now on the upper-left graph).

We can now ask whether these extant results help to extend the empirical base of locality effects beyond relative clauses. Unfortunately, they do not. Separate locality contrasts within the matrix verb condition were not reported in Grodner and Gibson (2005), but do not appear to be reliable. The contrast between the PP and no-interposition conditions in the embedded structures also was not reported, and also appears not to be reliable.[7] In short, it is quite possible that the locality

[6] For present purposes we remain neutral about the precise nature of the subject relation—under some accounts it may also involve movement from within the verb-phrase to an argument position outside it. Under any analysis, the subject-verb dependency here is qualitatively different from the extracted object dependency.

[7] The possibility of a spillover effect from the preceding was not taken into account in the Grodner and Gibson (2005) study; I address this in the analysis of the new data presented here.

effects are driven by independent sources of difficulty resulting from embedding the verb and from center-embedding the relative clauses.[8]

Despite the ambiguity attending the Grodner and Gibson (2005) results, I believe that the structure of their Experiment 2 is still a promising way, in principle, to explore locality effects, and its structure will be used for the four experiments presented here. But before moving on to the new experiments, it is worth briefly considering the implications of the narrow methodological base for investigating locality effects.

### *A concern about the existing self-paced reading evidence for locality*

Self-paced reading has the virtue of yielding a simple measure that is often sensitive to the fluctuating processing demands of incremental comprehension. But because each word (or phrase) disappears as soon as the reader presses a button, the stakes of each button press are high relative to moving the eyes forward in reading. If the reader encounters difficulty that would best be resolved by regressing to an earlier part of the sentence, for instance to find a particular argument, he or she has no recourse in self-paced reading but to try to remember or mentally rehearse what came before. Eye-movements could potentially leave an interpretable record of such recovery processes, but SPR cannot—except perhaps in significantly increased reading times.

This difference between SPR and eyetracking turns out to be crucial for interpreting SPR reading time data such as that in Grodner and Gibson (2005). The locality results observed by Grodner and Gibson (2005) are marked by an increase in

[8] Grodner and Gibson (2005) also reported a linear regression analysis of the relation between reading times and integration cost (the DLT locality metric), but that analysis does not provide independent evidence for possible locality effects in the simple conditions of (8), because it includes data points from all the conditions.

reading times for the most difficult condition (the doubly embedded relative clause, (5c)). It is therefore possible that these effects reflect recovery from failed argument-verb integration caused by the center-embedding. More specifically, the observed 125–150ms increase in reading time may not be due to longer integration or memory processes affected by locality, but primarily recovery processes—perhaps covert rehearsal—triggered by retrieval *failures*[9]. To anticipate one of the findings reported in this paper: the combined results of Experiments 1–4 provide support for this interpretation of existing SPR locality effects.

Why does it matter whether observed effects are associated with recovery or initial retrieval or integration? It matters for the purpose of building a cumulative quantitative base of results on which to build computational theories of the underlying memory processes. One should, in principle, be able to use the empirical results from reading studies along with our developing models of memory in parsing to converge on stable estimates of memory retrieval processing rates that may be meaningfully compared (and combined with) processing rate estimates obtained through other methodologies, such as speed-accuracy-tradeoff paradigms (McElree et al., 2003). Such quantitative integration is important not simply because we desire quantitative predictions but because it facilitates theoretical integration.

### *Overview of the empirical strategy and four experiments*

I will now provide a brief overview of our empirical strategy and describe how it is realized in the four new experiments that follow. The overall goal is to determine if it is possible to observe locality effects that are not subject to the critiques

[9]Note that this sense of recovery and reanalysis of prior material is different from the more common usage in the literature, which focuses on reanalysis of *misinterpreted local ambiguities* (e.g., Frazier & Rayner, 1982; Fodor & Ferreira, 1998). The assumption here is that there is a parsing failure grounded in a short-term memory retrieval failure, not a garden path in the conventional sense.

above. Ideally, this means observing locality effects at points of computing relations that do not involve $\overline{A}$ movement or interference between multiple arguments, and observing locality effects under conditions of relatively easy processing. These goals are achieved through the use of four empirical devices:

1. The six-condition structure of Grodner and Gibson (2005), outlined above in (5) and (7) was adopted for Experiments 1–4. In principal, this structure has the potential to reveal locality effects in the main clause conditions at points that do not involve extraction.

2. Parallel eyetracking and SPR versions of each experiment were conducted. The specific aims were to (a) provide potentially more sensitive measures of locality effects in easy, non-extraction structures; (b) distinguish between locality effects on early measures (if they exist) vs. late measures in the eye-movement record; and (c) provide a better understanding of the nature of locality effects observed in SPR by providing evidence bearing on the specific hypothesis above concerning the role of parsing failure and recovery in SPR.

4. The second set of experiments (3 and 4) introduce a new set of stimuli based on these structures but with content words drawn from a list of relatively short (three to six letter), high frequency words. The specific aims are to (a) increase the overall ease of processing and therefore provide an additional test of the hypothesis that locality effects might only be evident in the presence of other sources of processing difficulty; (b) decrease item-dependent variance related to the length and frequency of content words; and (c) increase the proportion of single fixations in the eye-movement record which might provide the best opportunity to observe the early manifestations of locality.

4. In the new set of stimuli, only inanimate nouns appeared in the extracted

object position. As described above, both the subject and extracted object in the original Grodner and Gibson (2005) materials were noun phrases referring to humans. Thus in addition to increasing locality, the embedding manipulation also potentially increased similarity-based interference.

The four experiments thus cross materials (original Grodner & Gibson stimuli and new stimuli) with method (SPR and eyetracking). Experiment 1 is SPR with the original Grodner & Gibson materials (a replication of their Experiment 2), Experiment 2 is eyetracking with the original materials, Experiment 3 is SPR with the new materials, and Experiment 4 is eyetracking with the new materials. For simplicity of presentation and analysis, complete analyses for each experiment are presented separately, but I also report a small number of key comparisons that test materials effects directly between Experiments 1 and 2, and 3 and 4.

## Experiment 1: Replication of Grodner & Gibson (2005) Exp. 2

### *Method*

A self-paced reading replication of Grodner and Gibson's (2005) Experiment 2 was run.

#### *Participants*

Forty-nine University of Michigan undergraduates participated for payment or for partial course credit. All participants were native English speakers with normal or corrected-normal vision, and were naïive to the purpose of the experiment.

#### *Stimuli*

Participants in Experiment 1 read thirty experimental sentences taken from Grodner and Gibson (2005) Experiment 2. Six versions of each item were used, as origi-

Table 3.2:: Examples sentences from the six conditions in Experiments 1 and 2; the critical verb is underlined.

| Condition | | Example |
|---|---|---|
| *Matrix* | *Unmodified* | The nurse supervised the administrator while... |
| | *PP-modified* | The nurse from the clinic supervised the administrator while... |
| | *RC-modified* | The nurse who was from the clinic supervised the administrator while... |
| *Embedded* | *Unmodified* | The administrator who the nurse supervised scolded the medic while... |
| | *PP-modified* | The administrator who the nurse from the clinic supervised scolded the medic while... |
| | *RC-modified* | The administrator who the nurse who was from the clinic supervised scolded the medic while... |

nally shown in (5) and (7), and repeated in Table 3.2 with condition labels.

For every item, the *matrix/unmodified* condition was a declarative sentence containing a transitive verb with human NP arguments. In the *matrix/PP-modified* condition the subject was modified with a prepositional phrase. In the *matrix/RC-modified* condition, a subject-modifying relative clause was made by placing the words *who was* at the beginning of the PP. In these three conditions, the object never undergoes movement.

The remaining three conditions were created by applying the same series of modifications (unmodified, PP-modified, RC-modified) to an adaptation of the core sentence. In all three conditions the object NP became the subject of the matrix clause (through $\overline{A}$-movement), and the rest of the sentence became an RC modifying that subject. A clausal connective always followed the matrix object.

Thirty experimental sentences were created and assigned to lists with a Latin square design. Forty-eight fillers and sixty-four sentences from unrelated experiments completed each list. Experimental trials never appeared consecutively, and no verbs or arguments were re-used.

The dependent measure is reading times at the first verb (e.g., *supervised*), which always occupied the same underlined position as in the examples in Table 3.2. This was where the dependency initiated by the first argument (*nurse* in the first three conditions or *administrator* in the last three conditions) was resolved. In the first three conditions this verb was in the matrix clause, so these conditions will be called the *matrix verb conditions*. In the last three conditions, the same verb was in an embedded clause, so these will be called the *embedded verb conditions*. In all conditions, the verb integrated with the same arguments across the sentence.

*Procedure*

Participants were seated with their eyes approximately twenty inches in front of a 17-inch Apple LCD display. After reading instructions, they read twenty practice sentences in the moving-window SPR paradigm, each followed by a comprehension question. Participants then began experimental trials.

In the moving-window paradigm, a series of dashes appeared wherever a word would appear for the current sentence. Participants pressed the spacebar to reveal the first word. Subsequent spacebar presses revealed the next word while replacing the prior word with dashes. Some sentences were long enough to require a second line of text, but in all cases the line break occurred after the critical verb.

Pressing the spacebar after the final word of a sentence removed the sentence from the screen and displayed a comprehension question. Participants responded *yes* to the question by pressing *f* on the keyboard or *no* by pressing *j*. If they answered correctly, "correct!" was displayed briefly; "incorrect" was briefly displayed if they answered incorrectly. Each press of the spacebar during sentence presentation was used as a reaction time measure for the text that had just been displayed.

*Statistical techniques used in the analysis*

Data analysis was carried out using linear mixed models (LMMs) (Bates & Sarkar, 2007) available as the package `lme4` in the R programming environment (R Development Core Team, 2006). In the analyses, participants and items were treated as random intercepts (sometimes referred to as random effects) and the contrasts (discussed below) as the fixed factors (or fixed effects). The effect of each contrast was derived by computing 95% highest posterior density (HPD) intervals for the coefficient estimates. Compared to conventionally used confidence intervals, the HPD interval is easier to interpret since it demarcates a range within which the population coefficient is expected to lie; this is how the 95% confidence interval is usually (incorrectly) interpreted. For details on how the HPD intervals are computed, see Gelman and Hill (2007); for an accessible description of posterior density estimates, see Kruschke (2010).

Following Grodner and Gibson (2005), analyses included all reading times within three standard deviations of the condition-mean reading time. (Less than 1% of the data were affected by this procedure.) Reaction time data from the critical verb in every experiment were log-transformed to correct for the typical positively skewed distributions observed with reaction times, yielding approximately normal distributions.

Two sets of five orthogonal contrasts across the six conditions were run in separate iterations of a linear mixed model that included both subject and item as crossed random factors. The key theoretical contrasts of interest in these sets are specified in Table 3.3. Contrasts were normalized to make the contrast coefficients in our models directly interpretable as estimated mean differences between the two groups

Table 3.3:: Two sets of contrasts used in the linear mixed models to analyze reading times from Experiments 1–4. Set 2 was a full matrix of five orthogonal contrasts, but only the theoretically interesting and non-redundant contrasts are shown here.

| | | Matrix | | | Embedded | | |
|---|---|---|---|---|---|---|---|
| | Contrast | ∅ | pp | rc | ∅ | pp | rc |
| Set 1 | Embedding effect (overall) | -0.50 | -0.50 | -0.50 | 0.50 | 0.50 | 0.50 |
| | Local vs. non-local (matrix) | -0.67 | 0.33 | 0.33 | 0.00 | 0.00 | 0.00 |
| | PP vs. RC (matrix) | 0.00 | -0.50 | 0.50 | 0.00 | 0.00 | 0.00 |
| | Local vs. non-local (embedded) | 0.00 | 0.00 | 0.00 | -0.67 | 0.33 | 0.33 |
| | PP vs. RC (embedded) | 0.00 | 0.00 | 0.00 | 0.00 | -0.50 | 0.50 |
| Set 2 | Locality × embedding interaction | 0.75 | -0.38 | -0.38 | -0.75 | 0.38 | 0.38 |
| | Modification type × embedding interaction | 0.00 | -0.50 | 0.50 | 0.00 | 0.50 | -0.50 |

represented by the contrast.[10] I refer to the difference between the means of the three matrix conditions and the three embedded conditions as the *embedding* effect (the first contrast in Table 3.3. I refer to the difference between the local (no modification) condition and the mean of the non-local conditions (the PP and RC modifications) as the *locality* effect, and specify two such effects, one for the matrix conditions (the second contrast in Table 3.3) and one for the embedded conditions (the fourth contrast in Table 3.3). The difference between these two locality effects is the *locality* by *embedding* interaction (the sixth contrast specified in Table 3.3). Similarly, I specify contrasts testing the difference between the two kinds of non-locality (PP and RC modification), separately for the matrix and embedded conditions (the third and fifth contrasts in Table 3.3). The difference between these two modification contrasts is the *modification type* by *embedding* interaction (the last contrast in Table 3.3).

*Spillover*

Although the critical verb was identical across conditions, the immediately preceding region was different in the unmodified (local) vs. modified (non-local) condi-

[10]Each contrast vector was normalized by dividing it by the difference between the positive and negative coefficients coding the two groups. For example, to normalize the vector [ −2 1 1 0 0 0 ], we divide it by the difference between the positive coefficient 1 and the negative coefficient −2, or $1 - (-2) = 3$. The normalized vector is thus [ $-\frac{2}{3}$ $\frac{1}{3}$ $\frac{1}{3}$ 0 0 0 ].

tions, so spillover is a possible contributing factor to estimates of the two locality contrasts. I adapted the statistical control for spillover used by Vasishth and Lewis (2006) as follows. In the analysis of data from self-paced reading experiments (1 and 3), reading time from the prior region, as well as the length and frequency of the word in the prior region, were included in the model. The final form of the models used for all analyses can be seen below.

(9) **Model 1 :** log( reading time ) ={contrasts 1 – 5} + length(current word) + frequency(current word) + length(previous word) + frequency(previous word) + spillover + random(subjects) + random(items) + error

(10) **Model 2 :** log( reading time ) ={contrasts 6 – 10} + length(current word) + frequency(current word) + length(previous word) + frequency(previous word) + spillover + random(subjects) + random(items) + error

*Results*

Two items were removed because they were improperly designed.[11] This left twenty-eight experimental items.

*Question accuracy*

Participants answered 74% of all trials correctly. Participants who answered fewer than 70% of the comprehension questions correctly were removed from analysis. Ten participants were excluded by this procedure, leaving thirty-nine participants' data to be analyzed.

[11] One item was ungrammatical because it was missing the matrix verb in the object-extracted sentences. The other item contained an intransitive verb in the critical position. Both design errors were present in the original Grodner and Gibson study.

*Word length and frequency*

The critical verb region does not vary from condition to condition, but we can potentially obtain tighter estimates of the contrast coefficients by explicitly modeling the effect of word length and frequency. The results reported for this experiment, and for Experiments 2–4, are from linear mixed models that include length and log frequency as covariates.

*Overview of the results figures*

Before describing the results of Experiment 1, an overview of Figures 3.2, 3.3 and 3.4 is appropriate. These figures systematically depict the results of all the experiments in this paper (as well as Grodner and Gibson (2005) Experiment 2).

Reading times in milliseconds at the critical verb are presented in Figure 3.2. Each separate panel in this figure depicts the reading times (and standard errors) across the six conditions. The three panels in the top row display SPR results (Experiments 1 and 3 and Grodner and Gibson (2005) Experiment 2) alongside the Total Fixation Times from the eyetracking experiments (Experiments 2 and 4). Data obtained from the original Grodner and Gibson (2005) materials (Experiments 1 and 2) are depicted with black lines; data obtained from the new materials (Experiments 3 and 4) are depicted with grey lines. As described in more detail below, the second row of panels in Figure 3.2 depicts the early eyetracking measures, and the last row depicts the late measures. The scale on the y-axis is always consistent across a row in the figure, but note that the early eyetracking measures are plotted on a different scale.

Rather than report the details of the statistical analyses in-line in the text, results of the tests are summarized graphically by plotting the mixed effect models' point

estimates of the contrasts as well as the surrounding 95% HPD intervals. The *locality* and *modification* contrasts within the matrix and embedded conditions (described above) are plotted in Figure 3.3. The *embedding* effect and its two associated interactions are plotted separately in Figure 3.4. The layout of both Figures 3.3 and 3.4 corresponds to the reading time panels in Figure 3.2.

The coefficient estimates depicted in Figures 3.3 and 3.4 are contrasts on the log-transformed reading times (normalized as described above) and so may be directly interpreted as differences on the log scale, or as multiplicative effects on the original untransformed scale. As in Figure 3.2, effects obtained with the original Grodner and Gibson (2005) materials are plotted in black lines, and effects obtained with the new materials are plotted in grey lines. The HPD intervals that include zero (and therefore fail to reach conventional levels of significance) are plotted as dotted lines; intervals corresponding to conventionally significant effects are plotted as solid lines.

*Results*

Analyses were conducted first using all trials, then again excluding trials on which the comprehension question was answered incorrectly. Because none of the analyses were affected by excluding incorrect trials, the reported analyses include all trials.

*Locality effects (see middle panel, top row of Figure 3.3).* There was an effect of locality in the embedded verb conditions but not in the matrix verb conditions; i.e., the non-local conditions (where the critical verb and its subject were separated by a PP or RC) were read more slowly than the local condition (where the subject and critical verb were adjacent, or local), but this effect was only reliable in the embedded conditions. In the embedded conditions, critical verbs in the RC condition were read more slowly than critical verbs in the PP condition, but this was not true in the

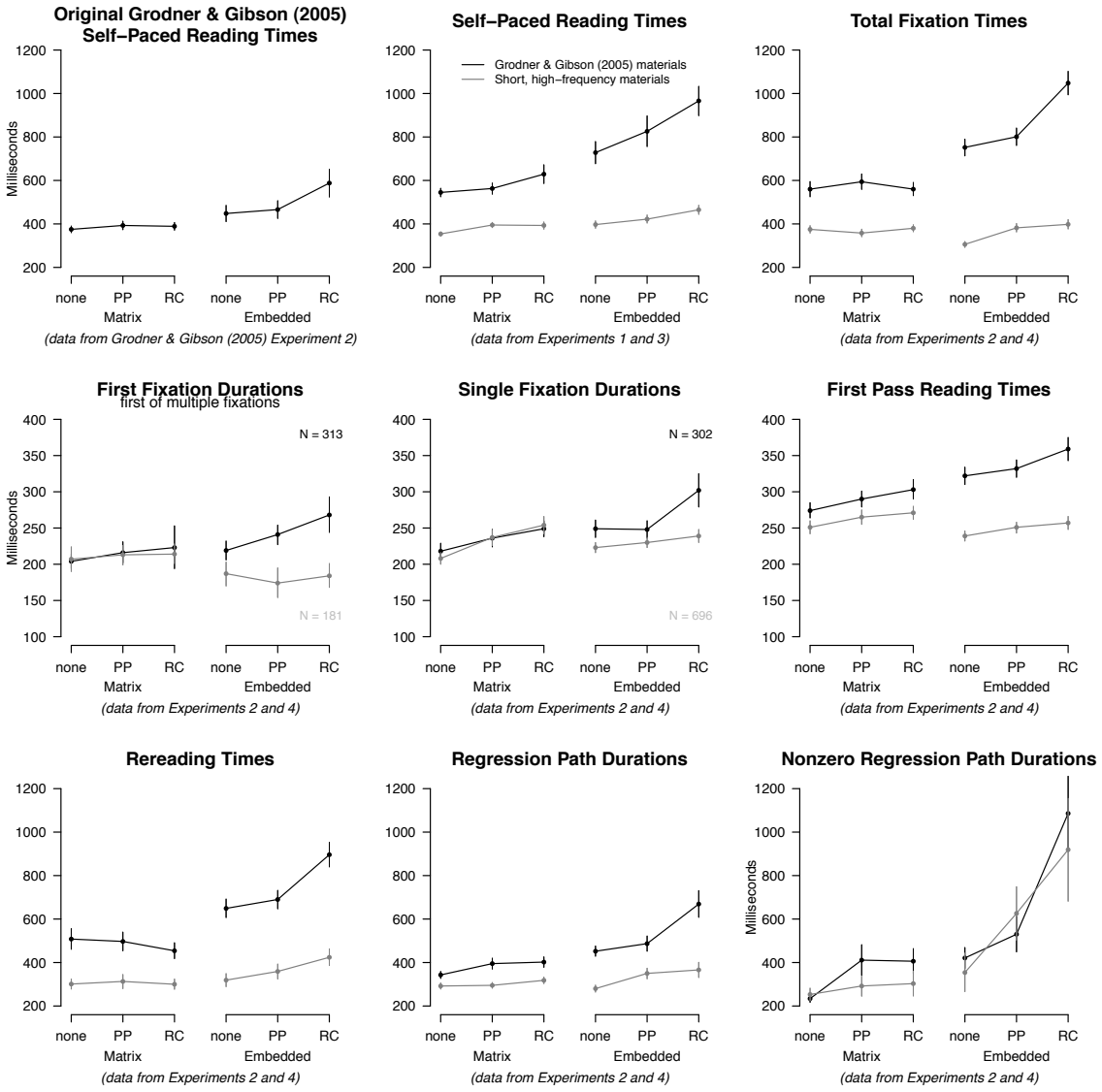


*Figure 3.2.* : Reading time measures from Experiments 1–4 and the original Grodner & Gibson (2005) self-paced reading study. Error bars are one standard error around condition means. Black lines indicate data collected using the Grodner & Gibson materials; grey lines indicate data collected using the materials composed of short, high-frequency words. The top row shows self-paced reading times from the Grodner & Gibson study (top left), self-paced reading times from Experiments 1 and 3 (top middle), and total fixation times from eyetracking Experiments 2 and 4 (top right). The middle row show the early eyetracking measures, and the bottom row shows the late eyetracking measures. Note that the scale for the early measures has a smaller range.

matrix conditions.

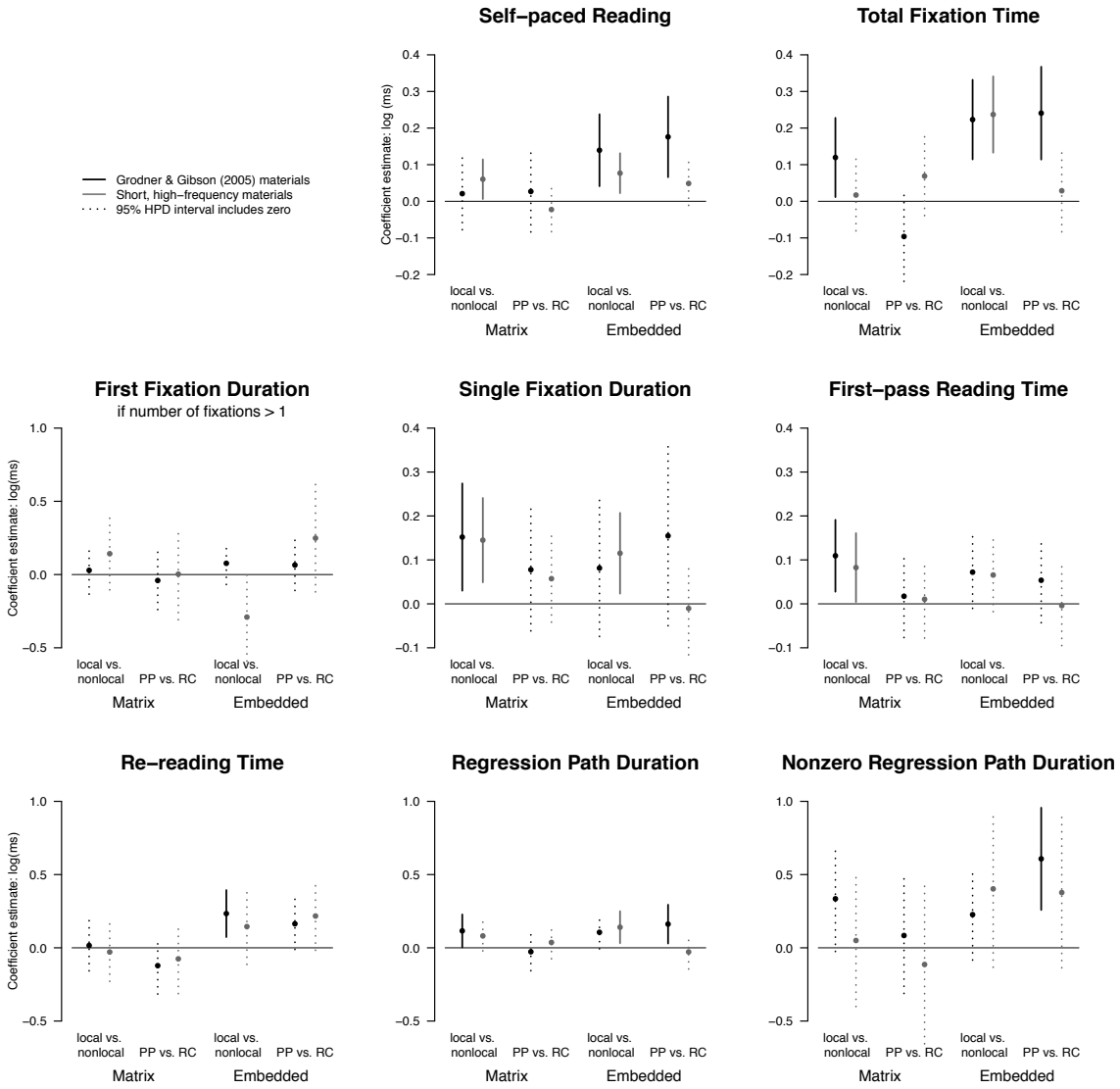


*Figure 3.3.* : HPD (highest posterior density) intervals for the locality contrasts in Table 3.3 for Experiments 1–4. Black lines indicate results obtained from data collected using the Grodner & Gibson materials, grey lines indicate results obtained from data collected using the materials composed of short, high-frequency words. HPD intervals that do not include zero, indicating a conventionally reliable non-zero coefficient estimate for the contrast, appear as solid lines.

*Embedding effect and interactions (see middle panel, top row of Figure 3.4).* Reading times at the critical verb were reliably slower overall in the embedded verb con-

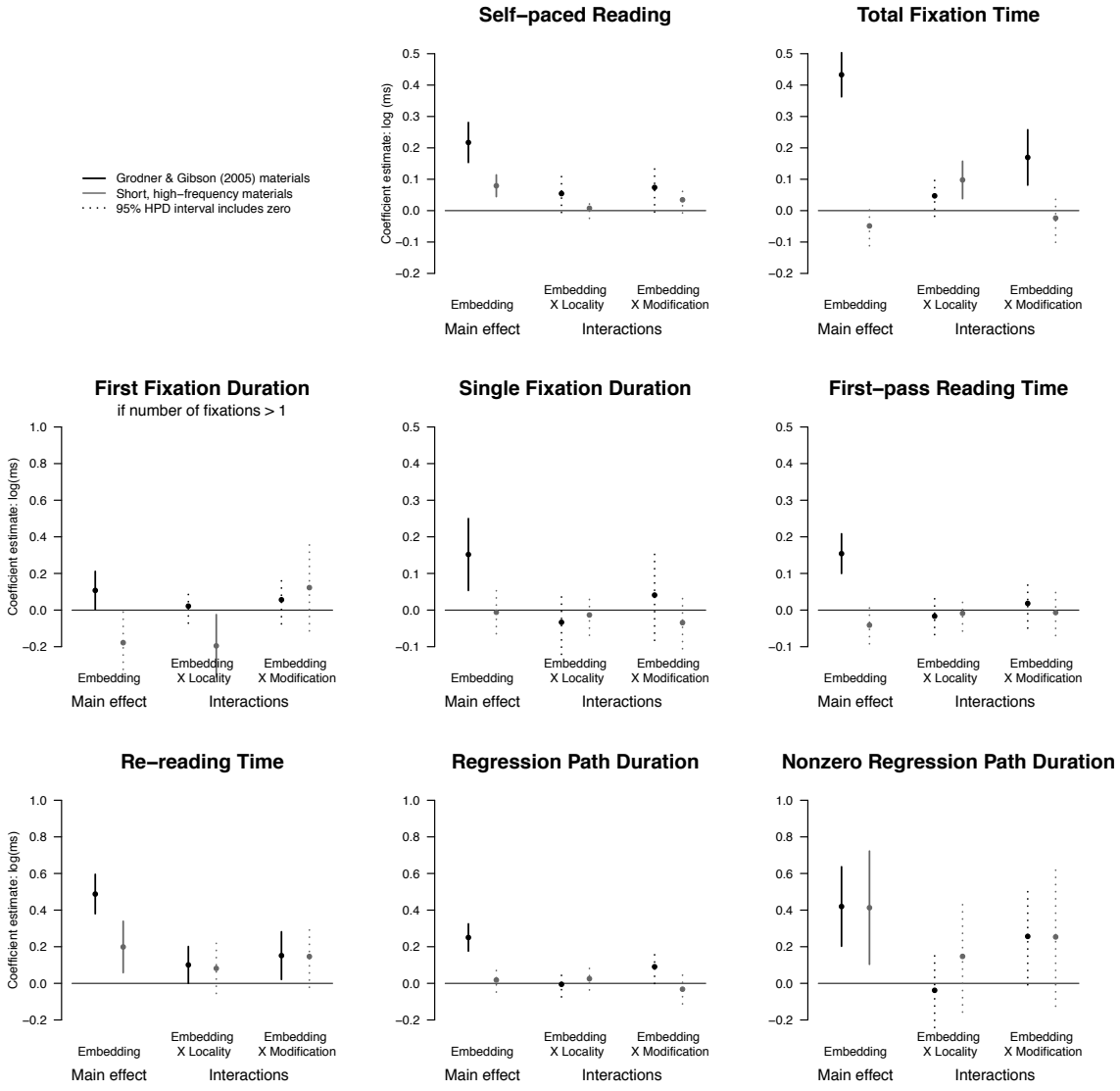


*Figure 3.4.* : HPD (highest posterior density) intervals for the embedding contrast and interaction contrasts in Table 3.3 for Experiments 1–4. Black lines indicate results obtained from data collected using the Grodner & Gibson materials, grey lines indicate results obtained from data collected using the materials composed of short, high-frequency words. HPD intervals that do not include zero, indicating a conventionally reliable non-zero coefficient estimate for the contrast, appear as solid lines.

ditions than in the matrix verb conditions. The *locality* effect was larger in the embedded verb conditions, and the difference between PP and RC modification was also larger in the embedded conditions; though these differences were only marginally reliable.

### *Discussion*

Experiment 1 replicated the basic pattern observed in Grodner and Gibson (2005). There was a locality effect in the embedded verb conditions, but not in the matrix verb conditions. The interaction of locality and embedding was marginally significant.

These results are thus ambiguous concerning the nature of locality effects in the ways detailed above in the analysis of the Grodner and Gibson (2005) results. The observed locality effects in the embedded verb conditions may be directly related to the increased distance between the subject and verb, but they could also be explained by retrieval interference between the two relative pronouns (*who*) in the embedded RC conditions, by interference between the object (*administrator*) and subject (*nurse*), or by other sources of difficulty related to center-embedding and object-relative extraction. The matrix conditions do not help to disambiguate the results of the embedded conditions. No locality effects were found in these simpler sentences. It is of course possible that locality effects are present but harder to detect in the simpler sentences due to other sources of variance in the materials or methodological limitations of self-paced reading

For present purposes, Experiment 1 serves the dual role of providing further motivation for the eyetracking and materials manipulations of Experiments 2–4, and providing an SPR baseline for the Grodner and Gibson (2005) materials in the same

participant population used in the subsequent experiments. Further discussion of Experiment 1's results is deferred for the time being, so they can be interpreted in the context of the results of the remaining experiments.

## Experiment 2: Eyetracking version of Experiment 1

Experiment 2 was an eyetracking version of Experiment 1 (and Grodner and Gibson (2005) Experiment 2).

### *Methods*

#### *Participants*

Forty-seven University of Michigan undergraduates participated for partial course credit or for payment.

#### *Apparatus*

Fixation time measures were gathered from both eyes using an SMI (SensoMotoric Instruments) Eyelink I head-mounted eye-tracker running at a 250 Hz sampling rate. Data from the right eye was used for all analyses.

#### *Stimuli*

The stimuli for this study were the same as Experiment 1. The same two items were removed from analysis due to design problems.

### *Procedure*

Participants were seated with their eyes twenty inches in front of a 17-inch CRT computer monitor, and the eye-tracker was fitted to their head. After the eye-tracker was calibrated using Eyelink-I software, participants began the first of twenty practice trials. Participants fixated a cross in the middle of the screen before every trial

to allow the experimenter to verify the calibration of the tracker. As soon as the experimenter observed stable fixation on the fixation cross, he pressed a button to replace the central cross with an identical one at the left edge of the screen. The entire sentence for the trial was presented as soon as the participant made a stable fixation on this fixation cross. Fixation data was gathered continuously throughout each trial. When the participant finished reading the sentence, he pressed the spacebar and a comprehension question appeared, and the participant proceeded as in Experiment 1.

## *Results*

### *Question accuracy*

Four participants were excluded from analyses for answering fewer than 70% of the comprehension questions correctly. The remaining participants averaged 80% accuracy on the comprehension questions for this experiment.

### *Reading time measures and covariates*

Definitions of the eye movement measures used in the analysis of Experiments 2 and 4 are given in Table 3.4. Note that our definition of First Fixation Duration excludes single fixations: it is the duration of the first fixation of multiple fixations, but we retain the shorter label for convenience. Linear mixed models were constructed for each measure using the contrasts given in Table 3.3; as described in detail above, Figures 3.3 and 3.4 show the contrast estimates and associated HPD intervals.

### *Spillover*

Last-pass reading time from the word immediately before the critical verb was used to model spillover. (See Table 3.4 for a definition of last-pass reading time). The

Table 3.4:: Definitions of the eyetracking measures used in the analysis of Experiments 2 and 4.

| Measure | Definition |
|---|---|
| *First Fixation Duration* | Time between the initial landing in a region and the beginning of the first saccade out of the region; excludes trials where only one fixation was made. |
| *Single Fixation Duration* | Time spent fixating a region when only one fixation was made therein. |
| *First-Pass Reading Time* | The summed duration of all fixations made within a region before exiting to the right or left. |
| *Regression Path Duration* | The sum of all fixations within a region $n$ and in any regions to its left before fixating to the right of $n$. |
| *Non-zero Regression Path Duration* | Identical to Regression Path Duration, but Non-Zero Regression Path Duration excludes cases where no regressions occurred. |
| *Re-reading Time* | The sum of all fixations in a region excluding first-pass reading time. Re-reading analyses include zero-millisecond re-reading times. |
| *Last-Pass Reading Time* | The sum of all fixations in the last run of fixations within a region. |
| *Total Fixation Time* | The sum of all fixations within a region during a trial. |

length and frequency of the preceding word were also used as covariates. Spillover was modeled for Single Fixation Duration, First Fixation Duration, and First-Pass Reading Time in all the results we report. Last-Pass Reading Times from the previous word accounted for a near-significant amount of variance in First Fixation Duration—which suggests that measuring spillover this way may be reasonable.

*Reading times*

Analyses were conducted with and without incorrect trials. Because excluding incorrect trials did not change any results, we report analyses over all trials.

*Locality effects (Figure 3.3).* There were locality effects in the matrix verb conditions in two first-pass measures—Single Fixation Duration and First-Pass Reading

Time (but not First Fixation Duration)—as well as Total Fixation Time. The embedded verb conditions showed a locality effect only in Total Fixation Time and in later, regression based measures. More specifically, there was a locality effect in Re-Reading Time, and a marginal effect in Regression Path Duration and Non-zero Regression Path Duration.

No difference was found between PP- and RC-modification in the matrix verb conditions. In the embedded verb conditions, critical verbs in RC sentences were slower than in PP sentences in Regression Path Duration, Non-zero Regression Path Duration and Total Fixation Time.

*Embedding effect and interactions (Figure 3.4).* Reading at the embedded verb was slower than the matrix verb in all measures. The locality effect differed between the matrix and embedded verb conditions only in Re-reading Time. More specifically, it was larger in the embedded verb conditions (see the *locality by embedding* interaction in Figure 3.4). Additionally, the difference between PP and RC modification was greater in the embedded verb conditions in Total Fixation Time and all the later measures (Re-Reading Time, Regression Path Duration and Nonzero Regression Path Duration).

### *Discussion*

Consistent with prior studies that have paired SPR and eyetracking (e.g., (Ferreira & Clifton, 1986; Ferreira & Henderson, 1990; Kennison, 2002; Trueswell, Tanenhaus, & Kello, 1993), Total Fixation Time (and Re-Reading Time) yielded times similar to SPR, in both qualitative pattern and absolute value. This relationship was most evident in the embedded verb conditions, where both SPR and Total Fixation Time (and Re-Reading Time) monotonically increased with increased subject-verb

distance, with a large increase in the most complex condition, the embedded relative clause sentences.

The most interesting results from Experiment 2 concern locality patterns in both early and late fixations. The first such result is the presence of locality effects in the simpler matrix conditions in the earlier measures. This can be appreciated by inspection of the middle row of Figure 3.2, which reveals a consistent monotonic increase in times across the matrix conditions for Single Fixation Duration and First-Pass Reading Times. The PP vs. RC contrast was not reliable for the matrix condition, but there was a consistent trend of greater reading times in the RC conditions across all the early measures.

The second interesting result from Experiment 2 is that only later measures (Re-Reading, Regression Path Duration and Non-zero Regression Path Duration) mirror the most salient result of the self-paced reading experiment: a sharp increase in reading times in the most complex doubly-embedded condition.

The locality effect was not reliable in every eye movement measure. While Total Fixation Time showed a locality effect for both the matrix and the embedded conditions, there were differences between the matrix and embedded conditions in other measures. For the matrix verb conditions, there was a locality effect in first-pass measures (Single Fixation Duration and First-Pass Reading Time). For the embedded verb conditions, there was a locality effect only in Re-Reading Time.

There were no reliable locality effects found in First Fixation Duration (in either Experiment 2 or 4). This is consistent with the locality effects found in Single Fixations and First-Pass Reading Times *not* being driven by spillover from the previous word.

*Interim summary and motivations for Experiments 3 & 4*

Experiment 1 replicated the results of Grodner and Gibson (2005), and provided a baseline for evaluating the relationship between SPR and eyetracking measures. The results of Experiment 2 are important for two reasons. First, the effects observed in the simple matrix conditions in Experiment 2 provide the first on-line evidence of locality effects in non-extraction structures, suggesting that locality effects are *not* restricted to complex movement structures, and that they do not rely on interference between possible retrieval targets or between multiple relative pronouns.

Second, for the more complex embedded conditions, the locality effect found in self-paced reading appears in *regressive* eye-movements to and possibly from the critical verb, not in first-pass fixation durations. The following tentative hypothesis may explain this finding. First-pass measures may reflect, in part, the duration of short-term memory retrievals that underlie successful integration, while later measures reflect recovery processes that occur when argument retrieval cannot be completed on time (i.e., before a programmed saccade must be executed). In the current materials, these retrieval failures in the most difficult of the embedded conditions may be a result of the combined effect of locality and similarity-based interference as described above. Experiment 4 offers further data relevant to assessing this hypothesis. In the general discussion, I will consider this claim in light of evidence from all four experiments. For now, I note that SPR times do not distinguish between recovery processes that show up in regressions and other processes that are reflected in first-pass measures.

The primary goal for both Experiments 3 and 4 was to increase our ability to detect locality effects across the conditions, and especially in the early eye-movement measures. The strategy adopted toward this end was to minimize overall compre-

hension difficulty, especially the difficulty associated with the embedded conditions.

Two lexical changes to the materials were made to accomplish these aims while maintaining the structure of the six conditions:

1. All words prior to and including the critical verb were restricted to short (3–6 letters), high-frequency (greater than fifty occurrences per million) words (Table 3.6).

2. The object of the critical verb, which was always an animate, human referent in the original materials, was made uniformly inanimate in the new materials. This change was expected to make processing of the embedded conditions easier in two ways. First, inanimate referents in the object position should reduce retrieval interference at the verb. Second, using inanimate referents as object may ease processing at the verb by biasing the reader towards an object-relative reading.

This manipulation to increase the bias toward the object relative reading is important because experience-based parsing theories predict local comprehension difficulty at points where new input signals a relatively unlikely continuation of the sentence (see Gennari and MacDonald (2008) for a summary). In particular, the constraint-satisfaction account of Gennari and MacDonald (2008, 2009) predicts difficulty in the embedded structures of our Experiments 1 and 2 on this basis. These studies demonstrate that object relatives beginning with an animate head noun like *administrator* are difficult to comprehend because the parser learns that structures other than object relatives are more likely to follow in such contexts (such as passives, e.g., *The administrator who the nurse was supervised by ...*). Encountering the verb *supervised* rules out more likely parses in favor of the unexpected object relative. Thus, the verb creates difficulty by violating the parser's implicit expectations. However, object relatives are frequently produced in sentences where an inanimate head noun fills the object role (Gennari & MacDonald, 2008), and there is evidence that these

constructions are nearly as easy to process as subject-relative clauses (Traxler, Morris, & Seely, 2002).

## Experiment 3: Testing locality effects using self-paced reading with short, high-frequency words

Experiment 3 was a replication of Experiment 1 using a new set of materials composed from a set of short, high frequency words. The motivations for this manipulation were detailed above.

### *Methods*

#### *Participants*

49 University of Michigan undergraduate students participated for partial course credit or for payment.

#### *Stimuli*

Thirty experimental sentences were created for use in a self-paced reading experiment (Experiment 3) and a parallel eyetracking experiment (Experiment 4). The syntactic structure of all sentences was identical to Experiments 1 and 2, and Grodner and Gibson (2005) Experiment 2, but content words were restricted to 3–6 letter words that had a frequency higher than fifty occurrences per-million-words in the First Release of the American National Corpus.[12] A comparison of the relevant lexical properties of the new and old materials is given in Table 3.6.

Table 3.5 gives examples of the materials. Items were assigned to lists using a Latin square design. Experimental items never appeared consecutively, and no arguments or argument modifiers were used more than once.

[12] http://www.americannationalcorpus.org/FirstRelease/

Table 3.5:: Example sentences from the six conditions for Experiments 3 and 4. The critical verb is underlined.

| Condition | | Example |
|---|---|---|
| *Matrix* | *Unmodified* | The child played the sports that were hard to master. |
| | *PP-modified* | The child from the school played the sports that were hard to master. |
| | *RC-modified* | The child who was from the school played the sports that were hard to master. |
| *Embedded* | *Unmodified* | The sports that the child played were hard to master. |
| | *PP-modified* | The sports that the child from the school played were hard to master. |
| | *RC-modified* | The sports that the child who was from the school played were hard to master. |

Table 3.6:: Lexical properties of each set of materials, through the critical verb position. The new materials for Experiments 3 & 4 included plural forms of content words, not including the verb, whose singular forms met all length and frequency criteria. Statistics for those content words were computed for the plural forms the participants saw. Frequency counts displayed are occurrences per-million-words in the American National Corpus.

| | Critical verb | | Content words | |
|---|---|---|---|---|
| | Exps. 1 & 2 | Exps. 3 & 4 | Exps. 1 & 2 | Exps. 3 & 4 |
| Median length | 8.0 | 4.0 | 7.00 | 5.00 |
| *Std. deviation* | 1.6 | .91 | 2.56 | .97 |
| Median frequency | 5.0 | 112.0 | 12.50 | 77.0 |
| *Std. deviation* | 13.2 | 166.3 | 53.10 | 88.78 |

*Plausibility norming*

In these materials locality is manipulated via nominal modifications that unavoidably change the semantic content of the sentences. To control for possible plausibility effects that may be confounded with the locality manipulations, I conducted a separate norming study with 57 participants from the same population who did not participate in the reading experiments themselves. Participants read each experimental item at one level of subject-modification, distributed randomly among 54 filler sentences, and rated plausibility on a 5-point Likert scale. Table 3.7 provides

Table 3.7:: Mean plausibility ratings on a 5-point scale for each level of subject-modification used in the new materials for experiments 3 and 4.

| Modification | Example | Mean rating |
|---|---|---|
| none | The child played sports ... | 4.29 |
| PP | The child from the school played sports ... | 3.55 |
| RC | The child who was from the school played sports ... | 3.9 |

the mean ratings for each level of modification.

To test whether dependency locality predicted plausibility ratings, a linear mixed model including two orthogonal locality contrasts was run. One contrast tested the unmodified-subject condition against both types of subject modification; the other tested PP modification against RC modification. Both contrasts were significant (HPD: local vsn̈on-local ( -0.45,-0.56 ); PP vs. RC ( 0.21, 0.45). Although there are plausibility differences, they are relatively small and we control for their effects on reading times in all the subsequent analyses by including item-level plausibility predictors in the mixed-effect models. None of the results reported below were affected by the inclusion of plausibility as a predictor.

*Procedure*

The procedure was identical to Experiment 1. Participants pressed the space-bar on a keyboard to advance through each sentence, and then answered a comprehension question about the sentence.

## *Results*

*Question accuracy*

Participants responded more accurately to comprehension questions in the second experiment, averaging 92% accuracy across all trials, suggesting that the lexical

manipulation succeeded in reducing overall difficulty. As in Experiment 1, participants failing to meet a 70% accuracy criterion were excluded from analysis. This disqualified one participant. Data from the remaining forty-eight participants were analyzed. One item was removed from analysis because it was displayed with words missing. Another item was removed because the critical verb did not meet the word frequency criterion; a third was removed because the sentence was missing its subject. The remaining 27 items were analyzed.

*Reading Times*

The self-paced reading times at the critical verb are presented graphically in Figure 3.2 (top row, middle panel, grey lines), and HPD intervals corresponding to the seven contrasts of interest are presented in Figure 3.3 and Figure 3.4.

*Locality effects (Figure 3.3, top row, middle panel, grey lines).* There was a locality effect in both the matrix and embedded verb conditions: reading times at the critical verb were longer in the non-local conditions than the local conditions. There were no reliable differences due to modification (the PP vs. RC contrast). The RC and PP contrasts were larger in the original materials than the new materials. This was established by a linear mixed model combining the data from the two SPR experiments that included a contrast coding the interaction of materials set and the embedding effect (contrast estimate = -0.056, HPD ($min$,$max$)= (-0.107, 0.005)).

*Embedding effect and interactions (Figure 3.4, top row, middle panel, grey lines).* Embedded verbs were read more slowly overall than matrix verbs. There were no reliable interactions, and unlike Experiment 1, these interactions did not approach conventional significance.

The embedding effect found in Experiments 1 and 2 appeared to be reduced, sug-

gesting that replacing the object with an inanimate noun phrase made the embedded verb sentences easier to comprehend. However, this cross-experiment difference in the embedding effect, tested by a contrast coding the interaction of materials set and verb embedding, showed no reliable difference between the SPR experiments (coefficient estimate=0.007; HPD interval = (-0.02, 0.03).

*Discussion*

The most important result of Experiment 3 is the locality effect in the matrix verb conditions. Using short, high-frequency words, locality effects were detected where they were not apparent (in SPR) in Experiment 1. The joint analysis of Experiment 1 and 3 also provide evidence suggesting that locality may interact with overall processing difficulty—here manipulated by lexical processing difficulty.

The empirical goals of this study were thus met: the materials change produced faster overall reading times *and* made it possible to detect a locality effect in the matrix condition. Furthermore, the size of the locality effect in both the matrix and embedded clause condition is comparable. The evidence from Experiment 3 thus supports the tentative conclusion we advanced in Experiment 2: locality effects exist outside of $\overline{A}$-movement and may be detected under conditions of relatively rapid and easy comprehension. Finally, the effects in Experiment 3 cannot be explained by the relative rarity of object-extracted structures with an animate, discourse-new direct object (because these sentences used inanimate objects).

## Experiment 4: Eyetracking version of Experiment 3

Experiment 4 was an eyetracking version of Experiment 3. Using shorter lexical items has the further advantage in eyetracking of reducing the number of fixations on individual words (Brysbaert & Vitu, 1998; Rayner, 1979), which should increase

the number of data points available to analyse as Single Fixations.

## *Methods*

### *Participants.*

Forty-five University of Michigan undergraduates participated for partial course credit or for payment.

### *Stimuli.*

The stimuli were identical to Experiment 3.

### *Procedure*

The procedure was identical to Experiment 2. Participants read each sentence and then answered a yes-or-no comprehension question about the sentence. Eye movement data were collected.

## *Results*

### *Question accuracy*

Participants averaged 92% accuracy across all conditions. All participants met the minimum accuracy criterion of 70%.

### *Reading times*

The same eye-movement measures used in the analysis of Experiment 2 were used to analyze Experiment 4 data, and these measures are plotted as solid grey lines along side the Experiment 2 results in Figure 3.2. The same seven contrasts in Table 3.3 were analyzed using linear mixed models with the same structure as Experiment 2, including covariates for length and frequency of the verb and the preceding word. The contrast estimates and HPD intervals are shown in Figure 3.3 and Figure 3.4.

*Locality effects (Figure 3.3).* There was a locality effect for the matrix verb conditions in the first-pass measures: Single Fixation Duration, and First-Pass Reading Time. In the embedded verb conditions, there was a locality effect in Single Fixation Duration, Regression Path Duration and Total Fixation Time.

Reading times for PP and RC sentences did not differ in any measure for the matrix verb or embedded verb conditions.

*Embedding effect and interactions (Figure 3.4).* Embedding the verb led to increases in Re-Reading Time and Non-zero Regression Path Duration.

There was only one reliable interaction: The locality effect was smaller in the embedded verb conditions than the matrix verb conditions in First Fixation Duration.[13]

A comparison between the two eyetracking experiments showed a smaller embedding effect in the new materials in all measures but Single Fixation Duration and Non-zero Regression Path Duration (HPD(*min*, *max*): First Fixation (0.04, 0.22); First-Pass Reading (0.05, 0.13); Regression Path (0.06, 0.16); Re-Reading (0.15, 0.32); Total Fixation Time: (0.18, 0.28)).

### *Discussion of Experiment 4*

There are three key results from Experiment 4. First, there were locality effects in the matrix verb conditions, as there were in Experiments 2 and 3. As one can see in Figure 3.2, there was a consistent increase in reading times (denoted by the grey lines) from local (no modification) to non-local (PP and RC-modification) in the Matrix condition across all the measures except First Fixation Duration and

[13]In fact, First Fixations show an anti-locality trend in the embedded verb conditions, although this trend is difficult to interpret in light of Total Fixation Time, which shows a larger locality effect for the embedded verb conditions than the matrix verb conditions.

Re-Reading Time.

Second, in contrast to Experiment 2, a locality effect for the embedded conditions emerged in an early measure (Single Fixation).

Third, and perhaps most striking, the main effect of embedding was eliminated in the early measures and was reliable only in Re-Reading Time and Non-zero Regression Path Duration. One possibility is that the embedding effects obtained in this experiment reflect only regressions triggered by retrieval failure.

One aspect of the data pattern in Experiment 4 remains surprising: the absence of a locality effect in Total Fixation Time for the matrix verb conditions. However, this negative result should not necessarily be taken to mean that subject-verb integration is unaffected by locality in the matrix verb conditions, because there were reliable locality effects in Single Fixation Duration and First-Pass Reading Times. Rather, the absence of a locality effect in Total Fixation Time appears to be a function of the high variance and null-locality effect in the re-reading measures, which contribute to the Total Fixation measure.

## Discussion of the locality experiments

Locality effects are important because they potentially inform us about the the short-term memory processes that underlie the on-line computation of linguistic relations in language comprehension. But as argued in the Introduction, the evidence for locality overall is surprisingly mixed, and the existing on-line evidence is both linguistically and methodologically narrow, while at the same time admitting alternative explanations that do not involve mechanisms affected by locality.

The four experiments presented in this paper were intended to broaden the evidential base and provide new insights into locality and its empirical manifestation.

In the remainder of this discussion section, I review the main conclusions, consider alternative explanations, and outline a theoretical model of how locality effects might arise as features of adaptive policies for controlling eye-movements and button-presses in reading.

### *The ubiquity and nature of locality effects*

There are three main conclusions that we draw from Experiments 1–4 concerning the extent and nature of locality effects. These conclusions represent tentative answers to the motivating questions in the Introduction.

1. Locality effects may indeed be ubiquitous: they emerge not only in the computation of relatively difficult embedded structures involving $\overline{A}$ movement (as replicated in Experiment 1), but can be detected in the computation of relatively simple subject-verb relations (as shown for the matrix conditions in Experiments 2–4). Experiment 1 replicated an earlier null finding for the matrix conditions, but Experiments 2–4 consistently showed that locality effects may be detected in those structures using eyetracking (Experiments 2 and 4) and using lexical items designed to ease overall processing.

2. The locality effects obtained in the present experiments appear to be robust against spillover effects and plausibility differences. Locality effects emerged in both the matrix and the embedded verb conditions when lexical properties and reading times from the pre-critical word were included in the model. Furthermore, locality effects were not evident in First Fixation Duration, where spillover effects would be expected, and where they were in fact observed. Including item-level plausibility for Experiments 3 and 4 in the analysis models did not alter the estimates of the locality effects.

3. The largest and most robust effects of locality previously observed in SPR correspond well with the pattern observed in rereading and regression measures in the eyetracking record. This is consistent with our hypothesis that the long SPR times correspond to recovery from short-term retrieval failures during parsing—the effects are large in SPR in part because they include time to recover from failure.

### *Alternative explanations*

I briefly consider here two possible alternative explanations for the observed locality effects: local ambiguity and experience-based accounts.

#### *Local ambiguity explanations*

In some of the items in the matrix conditions, there is a temporary attachment ambiguity at the critical verb: the verb may be parsed as either the main verb or the beginning of a reduced relative clause (as in *The child (from the school/who was from the school) played by his friends as a fool . . .*). Could this local ambiguity give rise to the locality effects found in our experiments?

Local ambiguity is unlikely to be the source of the locality effects for two reasons. Consider first how the ambiguity might in principle give rise to the effect. In animate-subject contexts such as these, there is an overwhelming bias for a main verb continuation (M. C. MacDonald, Perlmutter, & Seidenberg, 1994). The post-nominal modifications could thus give rise to a locality effect if they made the relative clause continuation more likely, producing either greater competition times for a single-path parser or longer reading times associated with pursuing the relative clause structure for a ranked parallel parser. But such post-nominal modifications make the onset of the matrix verb *more likely*, not less likely (Levy, 2008). Put another way, shorter subject phrases are more likely than longer ones (a point we take up again below

when considering experience-based approaches).

Second, the ambiguity in question rests on a morphological ambiguity between the active and past-participle form of the verb—an ambiguity that is present in twenty-three of the items in Experiments 3 and 4 (such as *played/played*) but not in seven of the items (such as *wrote/written*). When we analyze the effect of morphological ambiguity in a linear mixed model, we find no interaction between morphological ambiguity and locality.[14]

*Experience-based explanations*

It is also incumbent upon me to consider how two prominent experience-based theories might account for the observed effects: the *Production-Distribution-Comprehension (PDC) Theory* of Gennari and MacDonald (2009), and the *surprisal* metric of Hale (2001) (Levy 2008 noticed the relevance of this metric for locality and anti-locality effects). The central claim of PDC is that pressures on the production mechanism create distributional regularities in natural language, and comprehension performance is shaped by exposure to these distributional regularities. Thus, a mechanism that created a preference for producing short phrases might result in sentences with the non-local conditions being less probable, and more costly to parse, than the unmodified matrix or embedded condition baseline. The locality effects here are in principle consistent with this account, but it is presently not specified in enough detail to make clear predictions concerning the direction of the effects.

To see why, it is useful to consider an existing experience-based parsing account that is both consistent with the overall PDC theory, and is specified in enough detail to make on-line processing predictions: surprisal (Hale, 2001). Under the surprisal account, a contextual manipulation will make reading time on a word increase to

[14] A table of coefficient estimates and their HPD intervals is included in the appendices.

the extent that the manipulation makes the word less likely[15]—a clear and natural assumption of the effect of the probabilistic encoding of experience on reading time that is consistent with PDC. For the materials in the experiments presented here, locality effects would be expected if the post-nominal modification—increasing the length of the subject noun phrase—makes the matrix verb less likely. Working out the precise predictions of surprisal depends upon assumptions about grammar and parsing algorithm, but at least one implementation of surprisal has been shown to predict exactly the opposite pattern (Levy, 2008). The reason is simply that longer noun phrases are less likely than shorter ones, and so the longer the noun phrase, the more likely the matrix verb is to appear. In addition of the present findings from English, there is also evidence from German which appears to be inconsistent with the predictions of expectation-based accounts (Vasishth & Drenhaus, 2011).

The point of considering PDC and surprisal together here is not to argue that experience-based theories are unable to account for the observed effects, but simply to demonstrate that, even under the very plausible assumption that we have more experience with shorter rather than longer phrases, additional processing assumptions are required to generate specific reading time predictions that flow from this assumption. And at least one experience-based processing account (surprisal) has been instantiated in a way that does not make the correct predictions for the materials in Experiments 1–4.

Experiments 1–4 mark the start of a substantial empirical effort that will be complemented by a substantial modeling effort. Whatever theoretical developments may arise from that effort— and whatever developments ensue under any approach to incremental processing— the evidence from the four experiments presented here is

[15]See Hale (2001) and Levy (2008) for the precise mathematical formulation of surprisal, which we need not appeal to here.

relevant because it suggests that locality effects may indeed be a ubiquitous feature of human sentence comprehension.

# CHAPTER IV

# The interplay of expectation effects and retrieval interference

## Motivation

The results of Chapter III provide important new evidence of locality effects that are consistent with the notion of activation decay, but not with a range of alternative explanations. These results are critical evidence that dependent-head distance is a basic determinant of comprehension difficulty, and they confirm that distance— possibly as a proxy for decay— should be taken into account by models of parsing difficulty. Chapter IV of this thesis pursues another angle on understanding the role of working memory in parsing. These experiments tested how retrieval interference and semantic expectations (built over the course of a sentence) interact.

A variety of studies mentioned in Chapter II have shown speed-ups in comprehension when linguistic context strongly constrains semantic properties or other dimensions of upcoming input. Evidence for comprehension slow-downs due to retrieval interference was also presented. Surprisingly— despite substantial bases of research surrounding both expectation-based facilitation and similarity-based interference— the relationship between them has largely been left unspecified in models of parsing.

Unfortunately, Experiments 1–4 do not offer insight to the interaction between

retrieval interference and expectation, even if lexical frequency effects are construed as a type of expectation effect. There appears to be an interaction in which the difference between embedded PP and embedded RC sentences shrinks due to the consistently high lexical frequency and short lexical length in the new materials of Experiments 3 and 4. However, the new materials differed from the old materials in another crucial way: an inanimate noun was always used as the object of the embedded clause (which preceded the embedded subject). Even though the substitution of the embedded object may have decreased surprisal throughout the new sentences, interference between the embedded subject (of the critical verb) and the embedded object was, by hypothesis, reduced. Because the new set of materials in Experiments 3 and 4 confounded changes to interference and surprisal, we can't conclude that the embedded PP-RC difference shrank in the new materials because of an interaction between expectation and retrieval interference. It is equally plausible that the PP-vs-RC difference was reduced because of decreased retrieval interference, and that the embedding effect went away because changing the embedded object had an orthogonal effect on surprisal. Since Experiments 1–4 do not provide clear insight to the interplay between retrieval interference and expectation, another experiment was designed to examine whether and how they interact. Several hypotheses concerning the interplay between interference and interaction will be tested, and I will review each of them in turn; but first I will describe the experiment's design, to make it easier to interpret each hypothesis and evaluate the experiment that follows.

## Design

This experiment used sentences consisting of a main clause (e.g., *The notorious student heard that ...* ) and a sentential complement (*the unprepared student in the*

*difficult class failed the final exam ...*). Four versions of an example sentence are shown in Table 4.8.

Semantic expectation and retrieval interference were both varied within the subject-verb dependency in the sentential complement. Mean Cloze completion scores for each sentence were used as a measurement of expectation for the verb. Interference was operationalized through a ratio of semantic fit scores, in which a distracter NP's semantic fit with a verb was divided by the semantic fit of the verb's actual subject. This semantic-fit ratio was used as a predictor of retrieval interference. The sentences were designed to cover a range of values for (a) strength of expectation for the verb (see Experiment 5.1), (b) proportions of subject–verb "fit" measured by dividing the distracter subject's fit by the distracter subject's fit (see Experiment 5.2), and (c) similarity between the target and distracter noun phrases in a phrase-similarity-rating task (see Experiment 5.3). Each of these variables was measured empirically in three separate auxiliary studies described below. Later in this chapter, results will be plotted by splitting each variable at the median value, artificially creating four discrete groups for expository purposes; but all analyses were conducted with continuous predictors.

Expectation and each of the interference predictors were observed at the embedded verb (e.g., *failed*) and surrounding regions. Reading times were analyzed at the verb because it requires retrieval of the embedded subject to integrate the subject-verb dependency, and because it is a convenient point to measure the semantic expectations generated throughout the main clause and complement.

Semantic expectation was manipulated at the adjective modifying the embedded subject (*unprepared student vs. bright person*). In the first two examples of Table 4.8, the modifier strongly predicts the verb in conjunction with the subject

Table 4.8:: Four versions of an example sentence from the eyetracking experiment. These four versions are for illustrative purposes only, and do not indicate a definition of four discrete experimental conditions. Expectation and interference predictors were modeled as continuous predictors of reading time at the embedded verb.

| Expectation | Interference | Sentence |
|---|---|---|
| *strong* | *high* | The **notorious slacker** heard that the **unprepared student** in the difficult class failed the final exam and never re-took it. |
| | *low* | The **bright person** heard that the **unprepared student** in the difficult class failed the final exam and never re-took it. |
| *weak* | *high* | The **smart woman** heard that the **bright person** in the difficult class failed the final exam and never re-took it. |
| | *low* | The **unprepared student** heard that the **bright person** in the difficult class failed the final exam and never re-took it. |

noun and the subsequent prepositional phrase— for instance, *the **unprepared** student in the difficult class(failed).* In lines three and four, the probability of the verb that appeared is reduced because other verbs satisfy the semantic constraints of the sentence. For instance, *the **bright** person in the difficult class* could plausibly be followed by *failed*, but several other verbs like *excelled*, *learned* or *passed* may be equally (or more) probable continuations.

Both measures of retrieval interference (within the embedded subject-verb dependency) varied with changes to a two-word subject NP in the main clause. The first line in the table contains both a target subject (*unprepared student*) and a distracter subject(*notorious slacker* that semantically fit well as the subject of the critical verb. Similarity between these two NPs was measured by asking participants how similar the two phrases were to each other (Experiment 5.3). The other predictor of interference was derived from ratings of how well each NP fit as the subject of the embedded subject. For instance, the features that make the unprepared student in

the example sentences likely to fail an exam in a difficult class can easily apply to a notorious slacker as well; this makes the slacker a suitable distracter subject. Conversely, the distracter subject in the second line (*bright person*) shows poor semantic fit as a subject of the critical verb. The ratio of semantic fit ratings for both subjects (notorious slacker rating / unprepared student rating) became one predictor of interference. Scores on this variable are greater than one when the distracter is a better subject than the target NP, and less than one when the target is a better subject than the distracter NP. Higher scores predict greater interference effects at the embedded verb.

The distracter subject and target subject shared some features in all sentences. Notably, both were invariably adjective-modified, animate nouns. This was constant across all versions of a sentence. Changing other semantic features of the distracter subject therefore still increased distracter–target similarity in the higher-interference sentences as compared to the lower-interference sentences.

The syntactic structure of the sentences— including the distance over which subject-verb integration occurred— was kept constant across all sentences. Besides ruling out dependency locality as a factor in reading time differences, this means that the conditional probability of a verb appearing after the embedded PP cannot explain differences in the critical region. Controlling syntactic expectations for a verb created a more precise measure of *semantic* expectations for the verb that appears in the sentence.

Constructing a large number of items that could induce an expectation for the embedded verb restricted the range of usable verbs. As a result, about half of the verbs were ditransitive while the rest were transitive or optionally transitive. The consequence of this is that the three-word phrase that ended in the embedded verb's

object were not always identical. Those three words were either [PREPOSITION + ARTICLE + NOUN] or [ARTICLE + ADJECTIVE + NOUN], with the noun in the third post-verb region being the object of the critical verb. Exactly half of the items (17) had each ending type.

### *Predictions*

Extant theories suggest several possible outcomes from this experiment. One possibility is that expectations facilitate early, lexical processing of the current word by pre-activating its representation in long-term memory, whereas retrieval interference (between items in short-term memory) affects a later, syntactic integration stage. For example, (Levy, 2008) suggests that a two-factor model (incorporating surprisal and retrieval difficulty) might be the correct one. He does not explicitly argue for an additive effect, but the simplest implementation of that idea would be an architecture in which retrieval interference and expectation-based effects do not interact, for instance, a model in which expectation affects early, lexical processing and interference affects a later, syntactic processing stage.

There does exist some empirical evidence for this claim (Vasishth & Drenhaus, 2011), and it certainly has some face validity. Retrieval interference affects selection between words that have already been processed, with some residual level of activation after having been attended. The effect of expectation putatively unfolds in the sub-threshold activation of a word in long-term memory that has not yet been encountered (and may, in fact, never be encountered). Since retrieval of both the predicted word and its dependent is most likely to be triggered by fixating the predicted word, and retrieval of the dependent is presumably contingent upon recognizing the predicted word and setting retrieval cues, the predicted timeline of expectation ef-

fects and interference effects is consistent with eye-tracking studies that have found expectation-based effects in early, first-pass measures (e.g., Schustack et al. (1987)) and interference effects in later re-reading and regression-based measures (Gordon et al., 2006).

This simplifying assumption only holds, however, under the additional condition that the lexical entries that compete for retrieval have not been kept active throughout the sentence via higher-level integrative processes that contribute to the construction and maintenance of linguistic expectation. This assumption is not clearly supported. In fact, it seems highly plausible— especially in the case of similar arguments of a verb— that the activation of the interfering items and the building expectation for an upcoming word would interact in some way. However, most models of parsing do not predict any such interaction.

Since the aim of this experiment is to test the very assumption that expectation and interference have orthogonal effects on comprehension, a set of specific predictions must be derived. Under a simple additive model, strong expectation for the verb should cause shortened reading times as early as the first fixation if it affects lexical access, and possible also in later measures as higher-level processes see a percolating effect of this facilitation. The central prediction of this model is that, even if expectation-based facilitation and interference-related slowdowns are reflected in the same fixation measure, no interaction would be predicted at the point of retrieval or lexical access, both of which occur at the embedded verb.

Since the additive model predicts a null effect, a more detailed exposition of this prediction may help evaluate it more meaningfully. Fleshing out how interference and expectation might individually exert an effect on reading times may help toward this end.

The predictions for retrieval interference can be derived from a parsing model Lewis and Vasishth (2005) built upon the ACT-R architecture (Anderson et al., 2004) that models cognition in many tasks. This is by no means the only model of retrieval processes, but I use it as a representative of contemporary models because it provides the clearest, most concrete hypotheses for the present purposes and it shares many architectural assumptions with current domain-general working-memory theory (viz., unitary memory; a limited focus of attention; and parallel, content-addressed access to memory) (Jonides et al., 2008). Unlike other candidate models like the competitive inhibition model of Vosse and Kempen (2000), the Lewis and Vasishth (2005) model (henceforth LV05) also is capable of making word-by-word predictions of reading times.

In the LV05 model, there is no *structural* division between the durable representations associated with long-term memory and the privileged, more quickly accessible short-term representations associated with working memory. Instead, working memory is defined as a subset of long-term memory that is more highly activated than the rest. The focus of attention is claimed to be very small– including only the stimulus currently being attended.

Grammatical knowledge is represented as procedural knowledge in the form of production rules, while the lexicon is represented in declarative memory as a set of features bundled together in 'chunks'. A word's baseline activation[16] increases sharply when it is retrieved, and decays exponentially until it is re-activated by another retrieval.

Retrieval is modeled by the execution of several production rules, shown in Figure 4.5. The processing at "failed" in Example 11 illustrate how retrievals serve

[16]Words are retrieved as chunks from a unitary memory store, with each chunk containing some syntactic information like argument structure in addition to the lexical entry itself.

dependency integration.

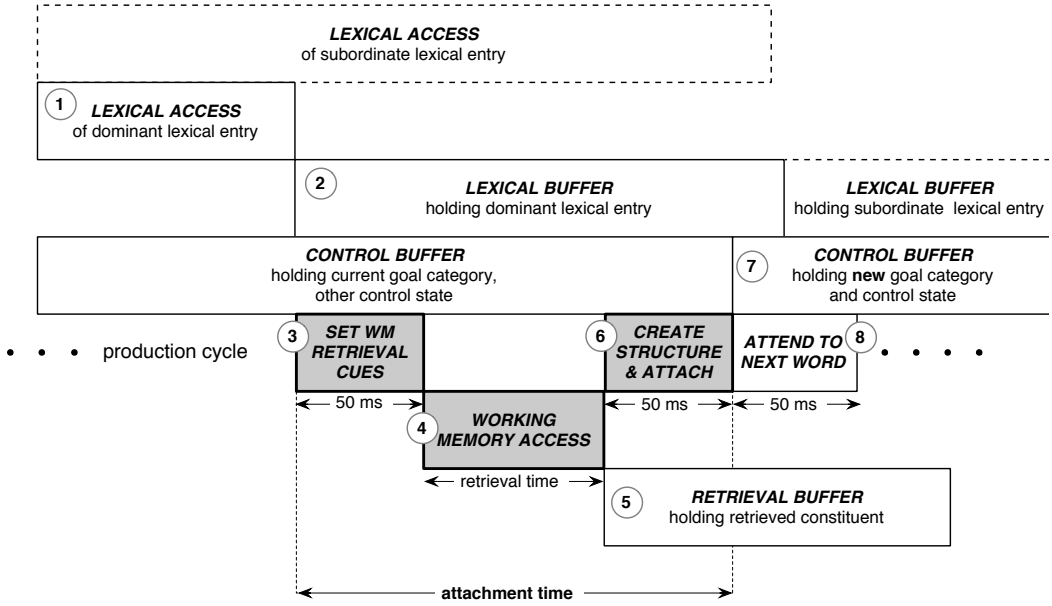


*Figure 4.5.* : An illustration of the time course of retrieval proposed by Lewis and Vasishth (2005), reprinted from that paper.

(11) The notorious slacker heard that the unprepared student in the difficult class **failed** the final exam ...

Once the reader has retrieved "class", a syntactic category goal is set through a production rule that embodies the parser's grammatical knowledge. In this case the goal that occupies the control buffer is to open a VP and attach it to the partial parse. When attention is shifted to the verb, the syntactic goal remains in the control buffer while "failed" is retrieved and occupies the lexical buffer.[17] Given the conditions that (a) a verb has been retrieved and (b) the syntactic goal is to construct a VP, a production rule fires that sets retrieval cues for a noun to serve as its subject. Retrieval cues will search all of memory in parallel for a properly

[17]The lexical buffer does not exist in the canonical version of ACT-R, but was added by Lewis and Vasishth.

inflected noun looking for a verb. Activation from the verb chunk will be distributed to any item in memory that matches these retrieval cues. The NP that is ultimately retrieved and placed in the retrieval buffer will be the one that has the highest level of activation. "The unprepared student" is likely to win this competition because it was recently retrieved, it has an unfulfilled syntactic goal of integrating as an argument of a verb. Retrieval interference occurs because similar items in memory may also match some of the retrieval cues. An NP that was recently retrieved *and* matches retrieval cues, like "notorious slacker", may have an activation level very close to the target NP (unprepared slacker), and if their activation levels are similar enough, signal noise may cause the distracter NP to be retrieved in error.

The LV05 model predicts that retrieval interference may affect the same fixation measures as expectation, but should be less apparent in first fixation where lexical retrieval is thought to be reflected, since the subject-retrieval that is manipulated in this experiment depends upon first retrieving and recognizing the verb. To the extent that retrieval interference causes retrieval errors, participants may also reliably answer comprehension questions incorrectly more often after sentences where interference is high. By the same token, strong expectation might cause increased accuracy to the extent that discourse-level processing is facilitated. Critically, the interference model in LV05 does not predict an interaction with expectation.

The retrieval interference model in LV05 predicts no interaction based on the premise that lexical access of the verb occurs in a separate processing stage, earlier than the syntactic integration stage that includes the retrieval of the subject. Going beyond the retrieval interference component of that model, two other predictions can be derived from memory theory.

First: Lewis et al. (2006) posit that difficulty retrieving a dependent might result

from encoding interference at the retrieval target. In Example 11, similarity between "notorious slacker" and the target subject "unprepared student" would cause some degree of difficulty or error in encoding the target. At the verb, retrieval would be difficult because features of the target NP may be confused with features of the distracter NP because the target was not correctly encoded. Encoding interference does not specifically predict an interaction with expectation at the critical verb. It does, however, suggest that interference effects might be traced to the target subject, upstream from the verb. Examining reading times there could clarify whether the expected interference effects at the verb are the typical type of retrieval interference described earlier, or a byproduct of encoding interference.

Second: While LV05 does not predict any expectation-by-interference interaction, that model also includes a decay component. Decay could, conceivably, interact with expectation in the following way: Assuming that readers take longer to complete lexical processing at the critical verb when it is relatively unexpected, the subject that must be retrieved may decay enough during the longer lexical processing of the verb to make retrieval more difficult. This would create an over-additive interaction in which interference effects were larger for less expected verbs. These larger interference effects could potentially be observed in any fixation measure, although the LV05 model of retrieval suggests that they might be more easily observed in measures later than the first fixation.

The preceding hypotheses regarding expectation and retrieval all place the locus of reading-time effects at the embedded verb. Lau (2009)has posed a very different hypothesis. She posited that sufficiently strong expectation for the verb could arise as early as its subject phrase, and the expectation would trigger the attachment of a verb slot to the existing parse. The payoff, downstream, is that the parser

should only need to do a quick check of bottom-up input to confirm its prediction of the verb. Expectation-based facilitation would then occur because the subject-verb dependency had already been integrated, obviating the need to perform an interference-prone retrieval of the embedded subject. The LV05 model also posits that a syntactic expectation is formed at the subject, but it does not predict that the verb itself is retrieved (in order to integrate the subject-verb dependency). A syntactic expectation could plausibly be maintained in the control state of the parser; however, since Lau (2009) predicts lexical retrieval of the verb and specifies no subsequent retrieval events, there is no mechanism to allow the verb to remain in the privileged spotlight of focal attention throughout the words intervening between the subject and verb.

One reason Lau's hypothesis is interesting in the present discussion is the fundamental difference between Lau's model and others that assume a small focus of attention and gradual degradation of recently retrieved representations. More importantly, Lau also makes a novel prediction that strong linguistic expectations can circumvent retrievals.[18] The consequence of this is the prediction that strong expectations could eliminate interference effects at important points of retrieval such as verbs. To my knowledge, this is the only explicit prediction of an expectation-by-interference interaction in the literature.

The following eyetracking experiment tests the simple additivity model as well as Lau's prediction, exploring whether semantic expectancy and retrieval interference make use of the same processing resources. Two supporting studies were conducted to gather empirical estimates of how strongly the critical verb was predicted by the preceding context (Experiment 5.1) and how much a distracter subject-NP matches

[18]It is not clear how encoding interference might be affected by early construction of the verb-argument structure in Lau's hypothesis.

likely retrieval cues set at that verb (Experiment 5.2). An additional supporting study gathered ratings of how similar the designated distracter- and target-NPs were to each other, when presented with no context (Experiment 5.3). Finally, an alternative predictor of interference was derived from the results of the semantic fit study. It measured the semantic fit of the distracter as a proportion of the rated semantic fit of the target subject, or [ distracter / subject ]. This semantic fit-ratio variable and the raw similarity variable were both used to predict reading times in several regions. Thus, in addition to testing the interaction between expectation and semantic fit, this experiment explores the differences between two methods of estimating interference effects.

## Methods

Choosing a valid measure is essential to drawing theoretical conclusions about the role of "expectation" or "predictability" in comprehension. I have argued that it is appropriate to operationalize expectation as the forward probability of a word— that is, the probability of a given word occurring as the continuation of a given prefix. There are, however, numerous ways to estimate probability. Extracting forward-probability statistics from corpora tagged with part-of-speech or other features has been a popular method for decades. This approach yields the transitional probability of a word occurring after one specific word (bi-gram probability), a two-word phrase (tri-gram probability) or a string of arbitrary length. This is also the approach that has been used to test the predictions of surprisal (Boston, Hale, Kliegl, Patil, & Vasishth, 2008). The disadvantage of this approach is that the researcher must make several choices about which features should be taken into account when calculating a word's probability. In order to make detailed predictions with surprisal, one also

must choose a grammar architecture and parsing algorithm to implement the theory.

On this count, getting readers to produce a word after a prefixed string has substantial advantages. While we may not know what features human readers are using to generate a completion, they are likely to reflect the actual predictive strategy of the parser during reading— at least more so than data from a tagged corpus. This is why the present experiment estimates of the target verb's semantic predictability empirically, using a Cloze completion procedure.

### *Participants and stimuli*

Thirty-seven undergraduates from the University of Michigan participated for partial course credit. Two subjects were excluded because they answered fewer than 70% of the comprehension questions correctly. Five-point Likert responses to the questions, "To what degree did you feel mentally tired during the experiment?" and, "How many hours of sleep did you get last night" were then used to identify participants whose fatigue may have impacted their fixation patterns. Participants who responded above the group median rating for mental fatigue *and* reported sleeping less than four hours the previous night were excluded. This disqualified only one subject. Five other subjects were excluded because they did not complete the post-survey questionnaire.

From the set of 40 items that were constructed, eight were excluded for a variety of reasons. Three were excluded because participants answered the associated comprehension question correctly less than 70% of the time. Two were excluded because they were missing words immediately before the critical verb. One was excluded because it ended at the third spillover region, which was included in the analyses. The remaining twenty-nine participants and thirty-four items were analyzed.

*Procedure*

Fixation times were recorded from the participant's right eye for every word in these sentences, using an SMI Eyelink I head-mounted eye-tracker.Sentences were assigned to a Latin square list and presented one-per-trial in the center of a computer screen. Experimental sentences were mixed with 59 filler sentences, and no two experimental sentences appeared consecutively. Some sentences were long enough to require wrapping to a second line of text, but the line break always occurred after the second word past the verb, or later.

To indicate they had finished reading and understanding the sentence, they pressed the space bar on a keyboard. A yes/no question probing comprehension of the preceding sentence then appeared. Comprehension questions probed comprehension of the embedded verb, the target subject and the distracter subject equally. Participants responded using the keys "f" and "j" on the keyboard.

The embedded verb (e.g., *failed*) was the critical region for all analyses, since it requires retrieval of the embedded subject and because it is a convenient point to measure the semantic expectations generated throughout the main clause and complement. The previous word, the noun that concludes the embedded PP, was also analyzed to detect possible parafoveal preview effects. Spillover effects were examined in the three words following the critical verb.

*Experiment 5.1: Cloze norming at the critical verb*

Participants in the Cloze study completed one version of each experimental sentence, truncated just before the critical verb. Each sentence was displayed in the center of a computer screen, with four blank lines displayed below it. Participants were instructed to use the first line to type the first word they think of to continue

the sentence. They were instructed to type three other reasonable continuations on the remaining three lines. Each response was visible and editable until the participant moved to the next blank line by pressing 'Enter'. After all four blank lines were filled, the trial screen was cleared and a transitional screen was displayed, to allow the participant to rest briefly. In the center of the transitional screen was printed, "Press space bar for the next sentence."

Cloze materials included four versions of each sentence from the main experiment, truncated immediately before the critical verb in the complement clause (for instance, *The night guard reported that the sneaky thief in the darkened museum ...* ). Example 12 shows an example sentence from a high-interference, strong expectation sentence. Participants saw only one version of each item.

(12) a. The notorious slacker said that the unprepared student ...

Cloze scores for each sentence were computed by taking the percentage of trials in which the pre-selected target verb was given as the first response.[19] The mean score for each sentence was calculated across subjects, standardized and used as a continuous predictor of expectation effects.

*Experiment 5.2: Semantic fit strength: distracter subject fit / target subject fit*

Retrieval cue overlap between the target and distracter subjects was assessed with a simple procedure Gordon et al. (2002) (and subsequently Van Dyke and McElree (2006)) used to measure the degree of semantic fit between a three-word list of potential subjects and two verbs that appeared in their experiment. A typical display is shown in Figure 4.6. All text was displayed in white courier font against a black background.

[19]Results for each version of the sentences appear in the appendix.

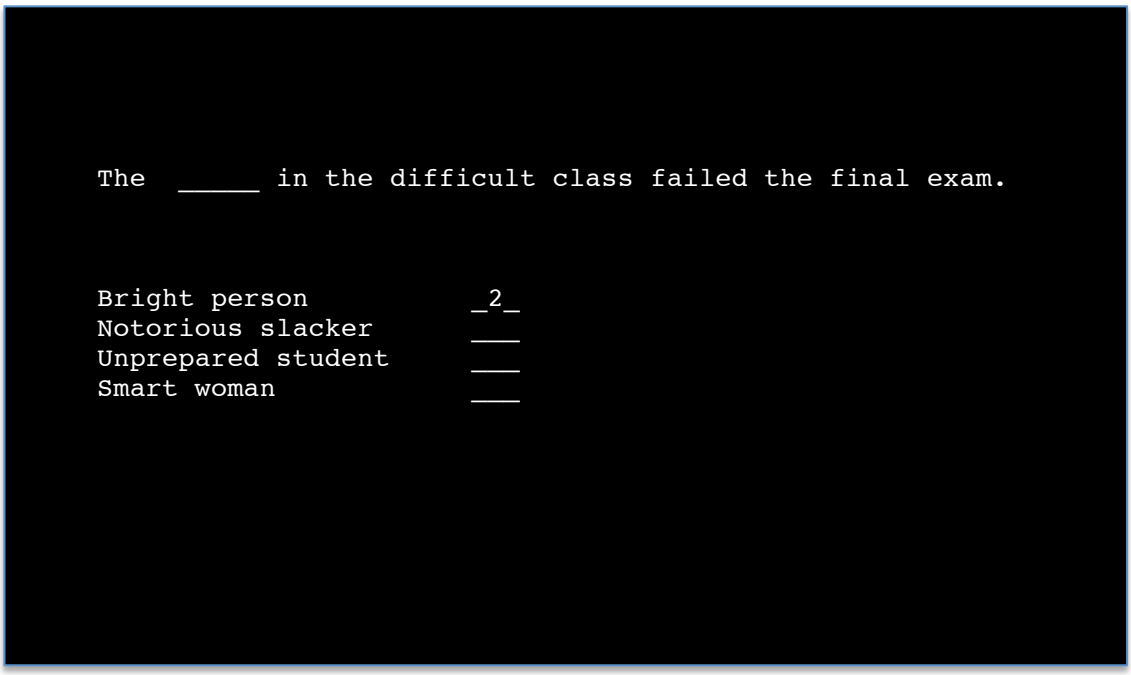


*Figure 4.6.* : A sample screen from the semantic fit rating experiment.

Participants viewed a sentence with a blank line where the subject noun-phrase was to appear in the experiment. A list of noun phrases was displayed below the sentence on the same screen, and participants were asked to rate how well each one would fit in the blank spot (at the subject position) of that sentence. The protocol was intended to classify potential subject NPs based on how well they might match retrieval cues at the verb, using semantic fit as a proxy for retrieval match. In Van Dyke and McElree (2006), this approach predicted interference-related slowdowns insofar as noun phrases with similarly high semantic fit caused difficulty when they both appeared in the same trial. The past success of this method led me to adopt the same procedure to measure the potential for interference in the present study. As shown in Figure 4.6, participants viewed sentential complements from the experimental materials, with the subject removed. Below the sentence,

participants saw four NPs, including both versions of the complement subject as well as both versions of the distractor subject that appears in the main clause of the full experimental materials, and rated each between 1 (very, very poor) and 9 (very, very good) as possible subjects of the sentence. The increased range on this scale was intended to encourage participants to make finer distinctions between the appropriateness of each subject NP. High ratings for the distracter noun phrase were taken to indicate that they should match retrieval cues intended to retrieve the embedded subject at the critical verb. Each trial could be ended by pressing the Enter key after the fourth response; otherwise it would time-out after thirty seconds.

In order to make the semantic fit ratings more analogous to the raw similarity judgments, the fit ratings were transformed. Scores from the similarity task represent a judgment about the semantic relationship between the two subject NPs, whereas the scores for each word in the semantic fit study represent a judgment about the fit between a single NP and a verb. A modified semantic fit score was created by taking the ratio of distracter fit to target fit. These new ratio scores were intended to remove a task-based difference in the comparison between the effects of retrieval-cue overlap and raw NP-similarity. [20]

*Experiment 5.3: Similarity between main clause (distracter) subject and target, embedded subject*

As a methodological complement to the semantic fit measure, a simple measure of similarity between subject NPs was also taken. On each trial, two of the four subject noun-phrases used in each item (shown in Example 4.6 were randomly paired side-by-side on a computer display. Participants were then asked simply to rate how

[20] All analyses were also conducted using the simple distracter-fit scores. The results were qualitatively identical to the results from the fit-ratio models, so they are not reported here.

similar the two phrases were, on a seven-point Likert scale from very dissimilar (1) to very similar (7). The average rating of similarity between the distracter and target subject for each experimental sentence was standardized and used as a continuous predictor of reading time.

The comparison between similarity ratings and semantic fit ratings will yield interesting results regardless of whether they predict exactly the same effects in reading times. If their predictive behavior is identical, the importance of a distinction between retrieval cue overlap and lexical similarity— which is critical to the behavior of the LV05 parsing model—may have to be revisited, at least in the context of gathering empirical estimates of similarity. If there is no predictive difference between paired-comparison and the multiple-response semantic fit paradigm, the pragmatic advantages and disadvantages of each task can be used to adjudicate which might be a more appropriate method for empirically estimating interference effects in the future.

Divergent results from these two methods could be even more interesting from a theoretical standpoint, because there is the potential for the tandem of predictors to reveal very distinct effects of retrieval cue overlap and semantic similarity on eye movements. This would bring some attention to the need for an empirically supported model of the types of features evaluated during retrieval— a critical, perhaps under-appreciated aspect of any model describing the computational underpinnings of retrieval. In the worst case, the effects predicted by similarity and semantic fit will differ in a way that is difficult to interpret. In this case, the present work still plays an important role, identifying a problem in need of empirical attention.

*Statistical techniques used in the analysis*

This experiment used LMER models similar to the ones described in Chapter III, again using the R statistical language. Examples 13 and 14 illustrate the structure of the LMER models used. Semantic fit and similarity were analyzed in separate models due to exceptionally high collinearity between them.

(13) **Semantic Fit Model :** log( reading time ) = *distracter / target fit ratio* * expectation + spillover + word length
+ word frequency + random(subjects) + random(items) + error

Cloze scores for each sentence were modeled as a continuous variable. Semantic fit between the distracter subject and the embedded verb, measured using the Van Dyke and McElree (2006) procedure, were used as a continuous predictor of retrieval interference. The effects of subject and item were modeled as partially-crossed random factors.

Mild collinearity between some predictors was removed by residualizing one of the collinear terms (see Example 15). These terms appear in italics in the example models. For example, because semantic fit ratings and Cloze scores were mildy correlated ($r$=.3), semantic fit was modeled as a function of Cloze scores, and the residuals— the variance in semantic fit ratings having partialed out covariance with Cloze scores— were used as a predictor in the model. Semantic fit was residualized against Cloze scores because the effect of interest is the degree to which distracter–verb fit affects reading times, beyond whatever effect expectation might have.

(14) **Similarity Model :** log( reading time ) = *distracter–target NP similarity* * expectation + spillover + word length
+ word frequency + subject(random) + item(random) + error

Raw similarity between the target and distracter NPs was measured using the phrase-comparison procedure described above, and was modeled as a continuous predictor. Similarity ratings were residualized against Cloze scores to eliminate multicollinearity between their coefficients.

(15) semantic fit = Cloze scores + word frequency + word length + item(random) + error

The effects of similarity were identical in the two models, so they will only be reported once, using coefficient estimates from the similarity model.

## Results

### *Comprehension question accuracy*

Participants answered 90% of all comprehension questions correctly. A logistic model was run to test whether semantic fit, similarity, or strength of expectation affected participants' ability to answer comprehension questions correctly after reading each sentence. Participants were reliably less accurate following high semantic fit sentences than low semantic fit sentences ($z = 2.76$, $p < .01$), showing that interference at the critical verb decreased accuracy. No other main effects or interactions were significant.

### *Reading times*

Results from the target subject, the critical embedded verb, the pre-verbal noun and a three-word spillover region after the verb were analyzed. Reading times from the verb are plotted here; all other results are included in the Appendix.The effects of word length, word frequency, and spillover are factored out to show the relationship

between expectation and each variable. These plots help to visualize the data conveniently, but they are artificial in the sense that they divide continuous measures of Cloze probability (of the verb), semantic-fit ratio and similarity into discrete groups. Recall that all three of these variables were modeled as continuous predictors in all analyses. The HPD interval plots display the results testing each variable as a continuous variable. Thus, the HPD intervals supersede the residualized reading plots in the case of an inconsistency.

*Target subject: embedded subject NP*

Reading times from the target subject are not plotted chiefly because the Lau (2009) model does not explicitly predict a slow-down at the embedded subject where she posits the subject-verb dependency can be predictively integrated. It is worth noting briefly, however, that there was no evidence of encoding interference in this region. There were no significant effects of expectation, interference, or their interaction.

*Pre-critical region: Noun*

The region before the critical, embedded verb was always a noun that ended the subject-modifying prepositional phrase (e.g., *class* in "... *in the class*"). At this point in the sentence, the reader must integrate the prepositional phrase with the subject that it modifies. They may also get some parafoveal preview of the verb from this word, depending upon how far right in the word they fixate.

Total fixation time at the pre-verb region was found to decrease as a function of increased semantic fit. This effect is difficult to interpret, but the lack of a similar effect in any of the first-pass measures suggests that this is not related to parafoveal processing of the verb or first-pass processing of the pre-verb region. In this context,

it is more plausible that total fixation times at the pre-verb region were affected by subsequent passes over the region that do not directly affect processing stages commonly linked to memory retrieval or the initial effects of expectation. It only appears in total fixation time, which aggregates over all runs of fixations on a word.[21]

There was no expectation effect in the pre-critical region, and no significant interactions between interference and expectation. High similarity between the distracter and target subject-NPs (but not a high semantic-fit ratio) resulted in longer total fixation times. This effect is also difficult to interpret, given its absence from first-pass measures.

*Critical region: embedded verb*

At the embedded verb, the dependency between the verb and the embedded subject can be resolved. The LV05 model predicts slowdowns in first-pass measures here due to retrieval interference as this dependency is integrated. Lau (2009) predicts no effects of similarity or semantic-fit ratio here— so long as there is evidence of expectation-based facilitation— because the subject–verb dependency should already have been resolved.

Results for the critical verb region appear in Figures 4.7 though 4.10. Figures 4.8 and 4.10 show the HPD intervals for the coefficients of each of the main coefficients and their interaction. Figures 4.7 and 4.9 show residual reading times for each fixation measure, with the effects of spillover, word length, word frequency and similarity regressed out to clearly illustrate the relationship between expectation and semantic fit. Table 4.9 and Table 4.10 show the mean

The effect of semantic fit was not significant at the verb. There was a marginally significant trend indicating longer single-fixations and longer first-pass reading times

[21] Plots of residual reading times and HPD intervals for this region appear in the Appendix.

Table 4.9:: Mean predictor values in each category for plots of semantic-fit ratio against expectation. Standard deviations are shown in parentheses.

| Mean Fit-ratio values | | |
|---|---|---|
| | High fit-ratio | Low fit-ratio |
| High Cloze | 1.52 (1.43) | 0.95 (0.9) |
| Low Cloze | 1.61 (1.11) | 1.25 (0.83) |

| Mean Cloze values | | |
|---|---|---|
| | High fit-ratio | Low fit-ratio |
| High Cloze | 44.58 (17.37) | 40.61 (12.82) |
| Low Cloze | 7.13 (7.24) | 5.91 (6.61) |

when the distracter NP was proportionally better than the target NP as the subject of the verb. The interaction with expectation did not approach significance in any measure.

High similarity between the distracter and target subjects caused longer first-fixations (see Figure 4.9). This could be interpreted as spillover from similarity-based difficulty that began at the previous region. However, the similarity effect was not significant in any first-pass measures at the pre-verb region. It was only significant in total fixation time, which also reflects regressive fixations from other regions.

Strong expectation for the verb caused shorter first-fixations and shorter total fixation times on the verb. The expectation trend was also marginally significant in first-pass reading time.

Strongly constraining semantic context has also been found to increase the probability of skipping the predicted word and decreasing the probability of regression from it (Rayner & Well, 1996). There was no effect of expectation on skipping rates (all HPDs include zero). This is not especially surprising because expectation has only been found to impact skipping rates in length-controlled stimuli where the

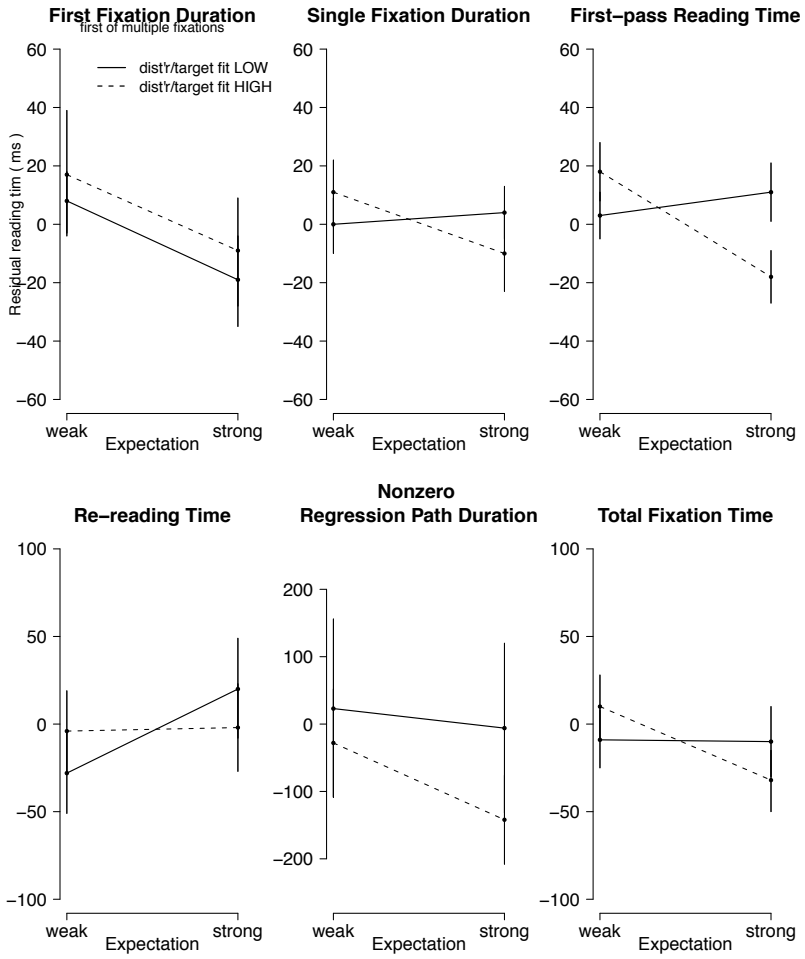


*Figure 4.7.* : CRITICAL VERB residual reading times: Cloze scores plotted against semantic-fit ratio.

predicted word was between four and six letters in length. The present study statistically modeled the variance due to word length, but did not experimentally control the length of the verb.

*Spillover region*

The expectation effect was also found in the three-word spillover region immediately following the verb. There were two types of phrase that followed the verb, depending on the type of verb. This results from natural limitations on the number

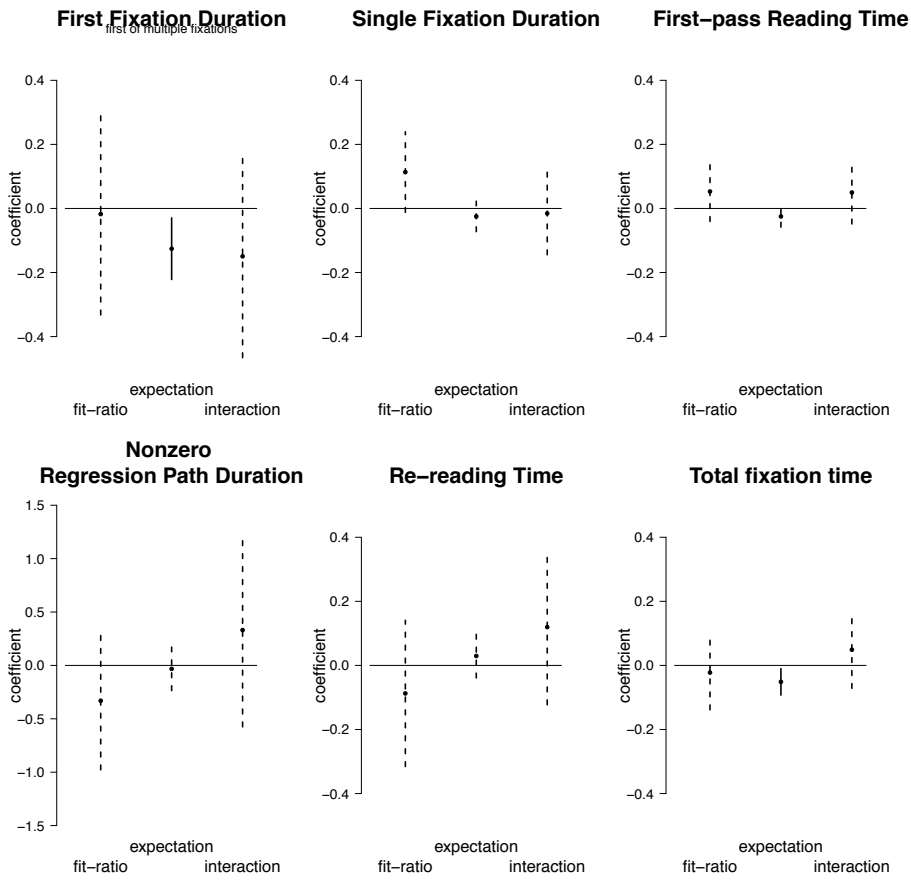


*Figure 4.8.* : CRITICAL VERB HPD intervals showing the estimated regression coefficients for Cloze scores, semantic-fit ratio, and their interaction.

of verbs that could be selected using the ability to induce a strong verb-expectation as a selection criterion.

Reading times for the spillover region were modeled using a two-level factor to represent the different ending types. Residualized reading times are displayed separately for each spillover ending-type, to help visualize the differences between them.

In the NP ending, sentences with strong expectation for the verb were read faster than weak-expectation sentences. This effect was significant in first-pass reading time, right-bounded reading time, and total fixation time. There was a trend in regression path durations indicating that regressions from the spillover region were

Table 4.10:: Mean predictor values in each category for plots of similarity against expectation. Standard deviations are shown in parentheses.

| Mean similarity values | | |
|---|---|---|
| | High similarity | Low similarity |
| High Cloze | 5.99 (0.49) | 2.98 (0.62) |
| Low Cloze | 6.03 (0.81) | 3.01(0.58) |
| **Mean Cloze values** | | |
| | High similarity | Low similarity |
| High Cloze | 44.69 (17.99) | 40.79 (12.33) |
| Low Cloze | 6.31 (7.11) | 6.71 (6.77) |

longer as the strength of the distracter NP's fit with the verb increased over that of the target NP. This trend consistent with a semantic fit effect was not significant. [22] No other effects were significant in the NP ending.

In the PP ending, there were no significant effects of expectation, semantic fit, or their interaction.

Residual reading times plotting similarity against expectation appear in the Appendix. The associated HPD intervals appear here. There was no effect of expectation or distracter-to-target NP similarity, and no interaction between them.

[22]Note that the regression-path residual reading times are plotted on a wider scale.

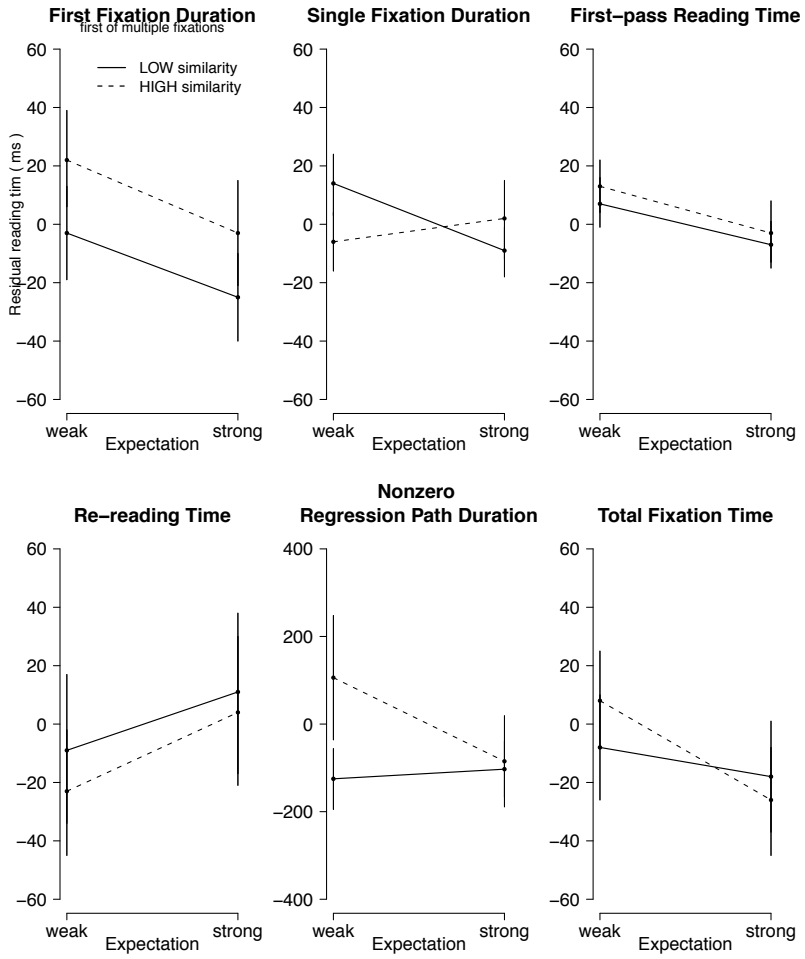


*Figure 4.9.* : CRITICAL VERB residual reading times: Cloze scores plotted against subject-NP similarity.

## Discussion

This experiment supports the simple additive model of retrieval interference and expectation-based facilitation and upholds the simplifying assumption that retrieval interference and expectation-based processes have separate, non-interactive effects on comprehension processes. There was no interaction between expectation for the verb and either semantic fit or distracter–target similarity. The absence of any interactions is not unequivocal evidence that no interaction exists; but the present results do support that hypothesis.

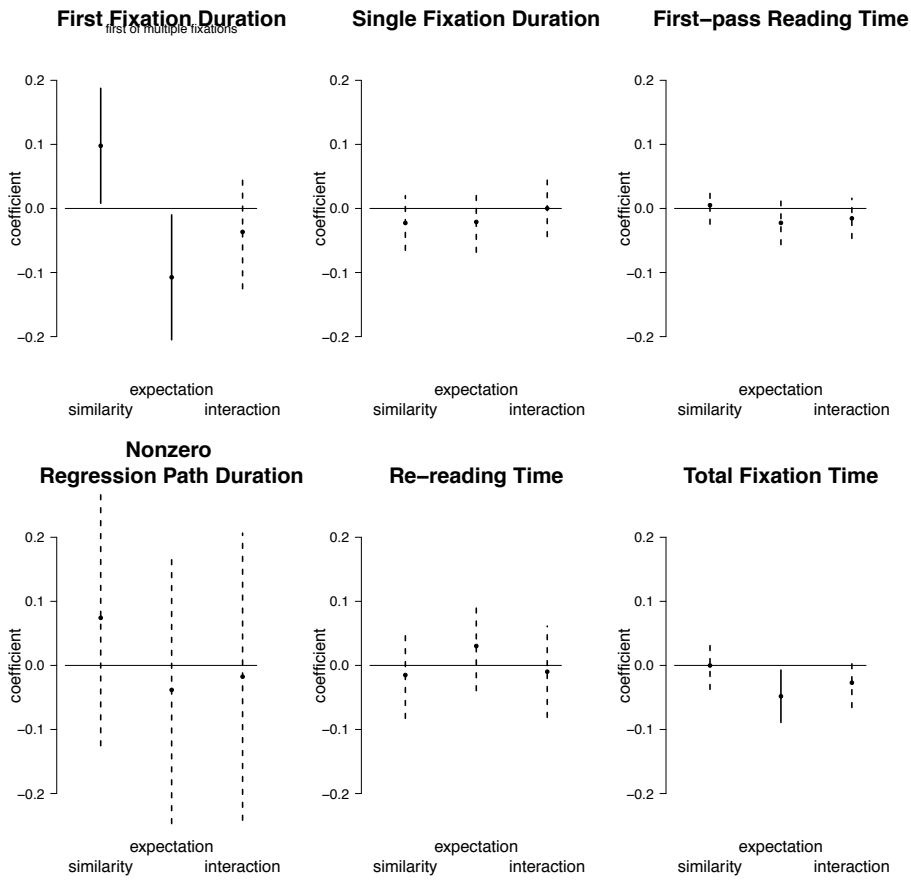


*Figure 4.10.* : CRITICAL VERB HPD intervals showing the estimated regression coefficients for Cloze scores, subject-NP similarity, and their interaction.

Induced expectation for the embedded verb caused shorter first-fixation times and total fixation times at the critical verb, and a matching trend was found in single-fixation duration and first-pass reading times. This effect also spilled over onto the following phrase when the verb was followed by an NP— but not when the following phrase was a PP.

The expectation effects in this experiment also support surprisal's prediction that difficulty parsing a word varies *continuously* with its probability of occurring after a given prefix. Expectation-based facilitation was found across a range of sentences that induced varying degrees of expectation for the verb (despite efforts to create two

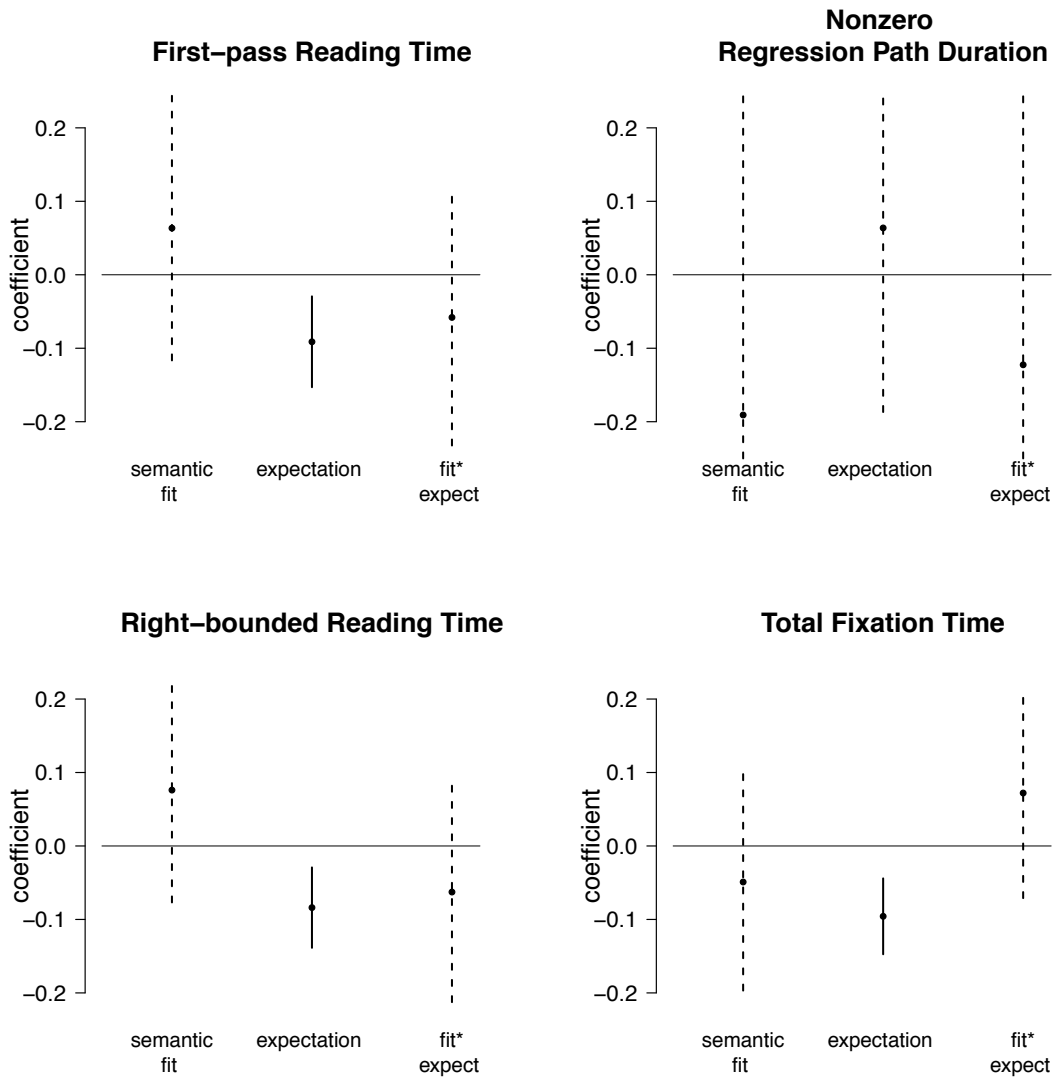


*Figure 4.11.* : NP-ENDING SPILLOVER region HPD intervals showing the estimated regression coefficients for the expectation effect, interference effect, and their interaction.

discrete levels of expectation in the materials). This raises the interesting question as to whether expectation effects could have been observed in previous studies that failed to find them with relatively less-extreme expectation contrasts, had expectation been treated as a continuously varying predictor (Hyona, 1993).

While there was no reliable on-line evidence of semantic fit effects, there were

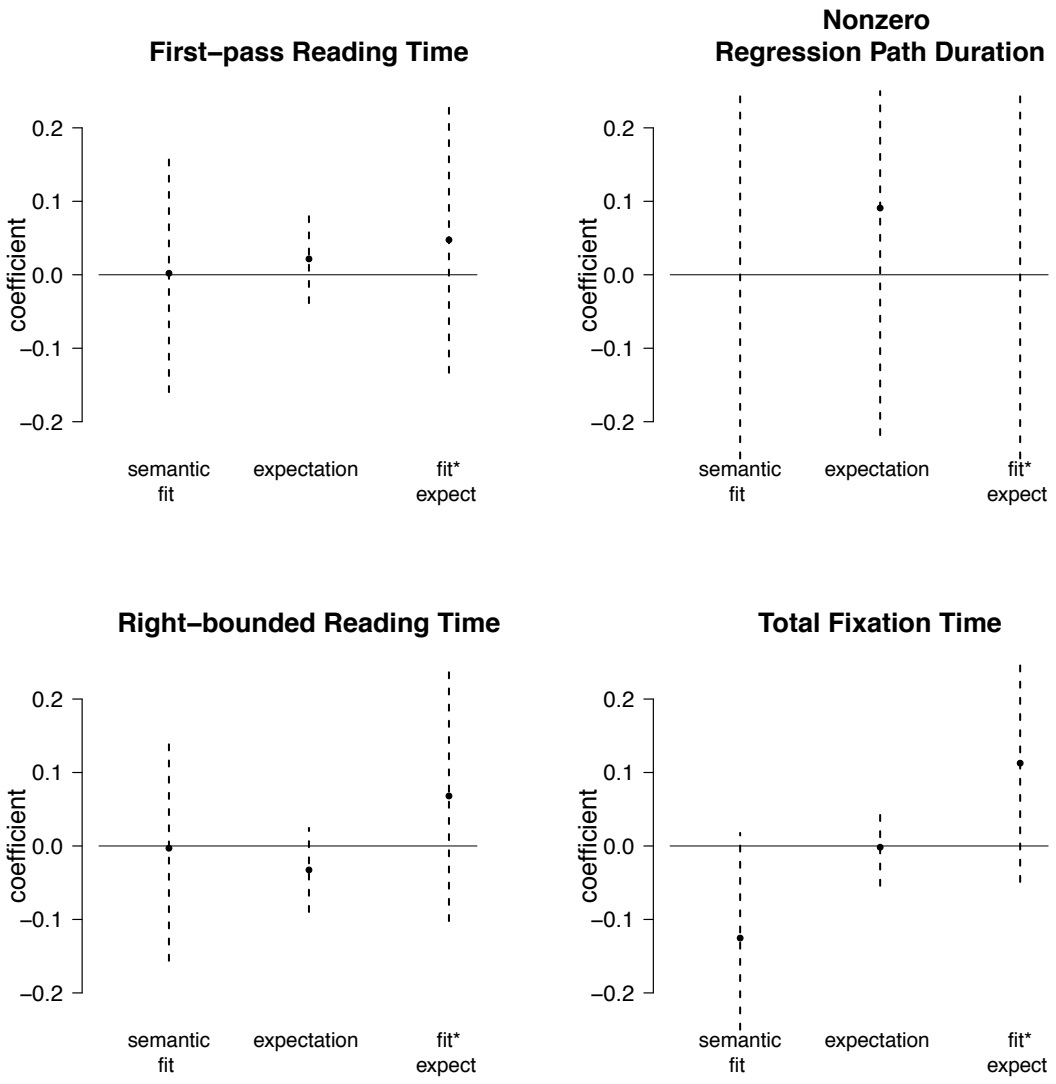


*Figure 4.12.* : PP-ENDING spillover HPD intervals showing the estimated regression coefficients for the expectation effect, interference effect, and their interaction.

marginally significant trends in the predicted direction in single-fixation duration and first-pass reading. Participants' accuracy on the comprehension questions additionally suggests that the rate of retrieval failure may have been elevated when semantic fit was high.[23] Increased similarity between the distracter and target NPs

[23]Results were qualitatively unchanged in all analyses using the simple distracter-fit scores reported by Van Dyke and McElree (2006).

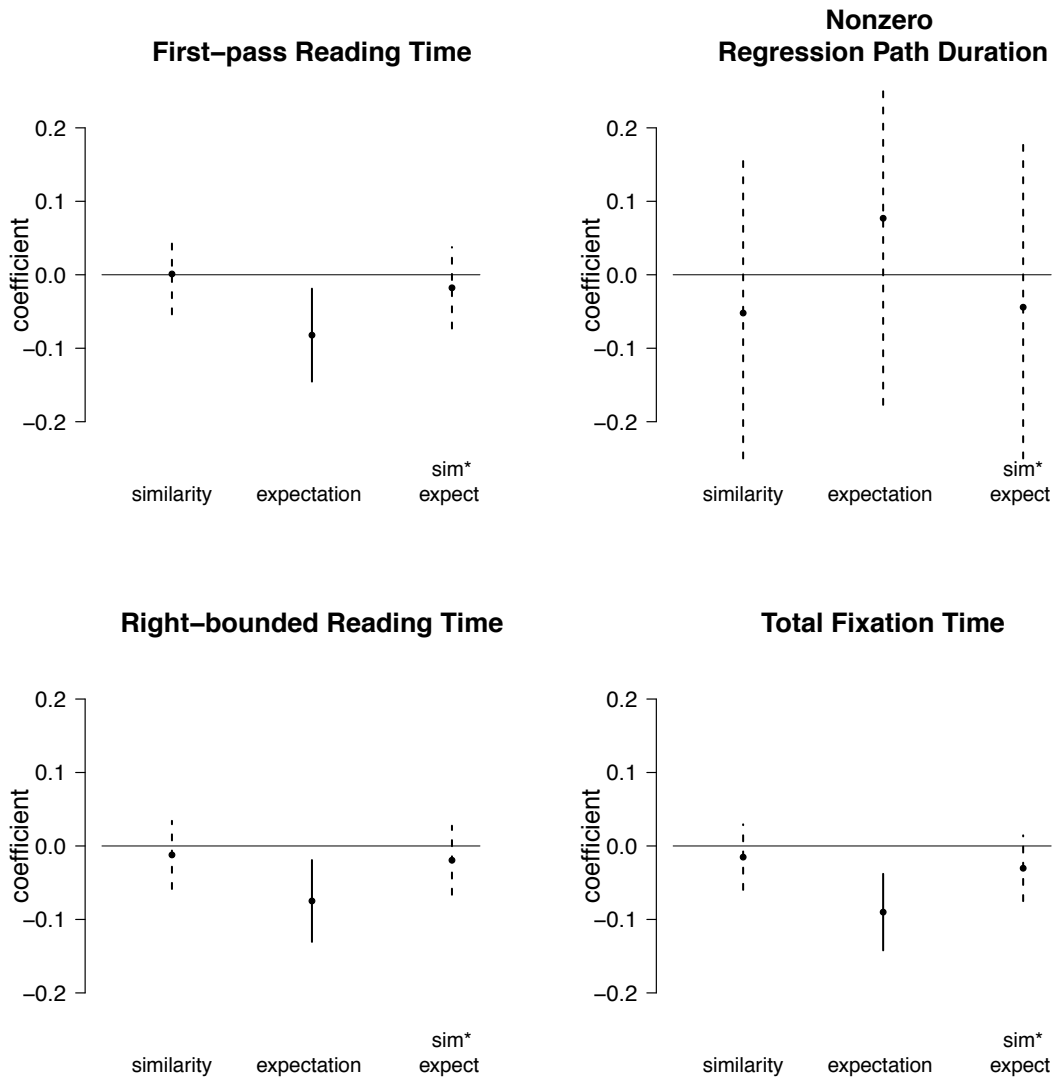


*Figure 4.13.* : NP-ENDING SPILLOVER region HPD intervals showing the estimated regression coefficients for the expectation effect, similarity between the target and distracter subjects, and their interaction.

caused significantly longer first-fixations at the verb. In fact, similarity had a significant effect only at the verb. I have argued that this is not a spillover effect because the effect was not found in first-pass measures at the region before the verb. There are additional reasons to believe that spillover cannot account for this effect, nor the early effect of expectation at the verb, namely: (a) The four words prior to the

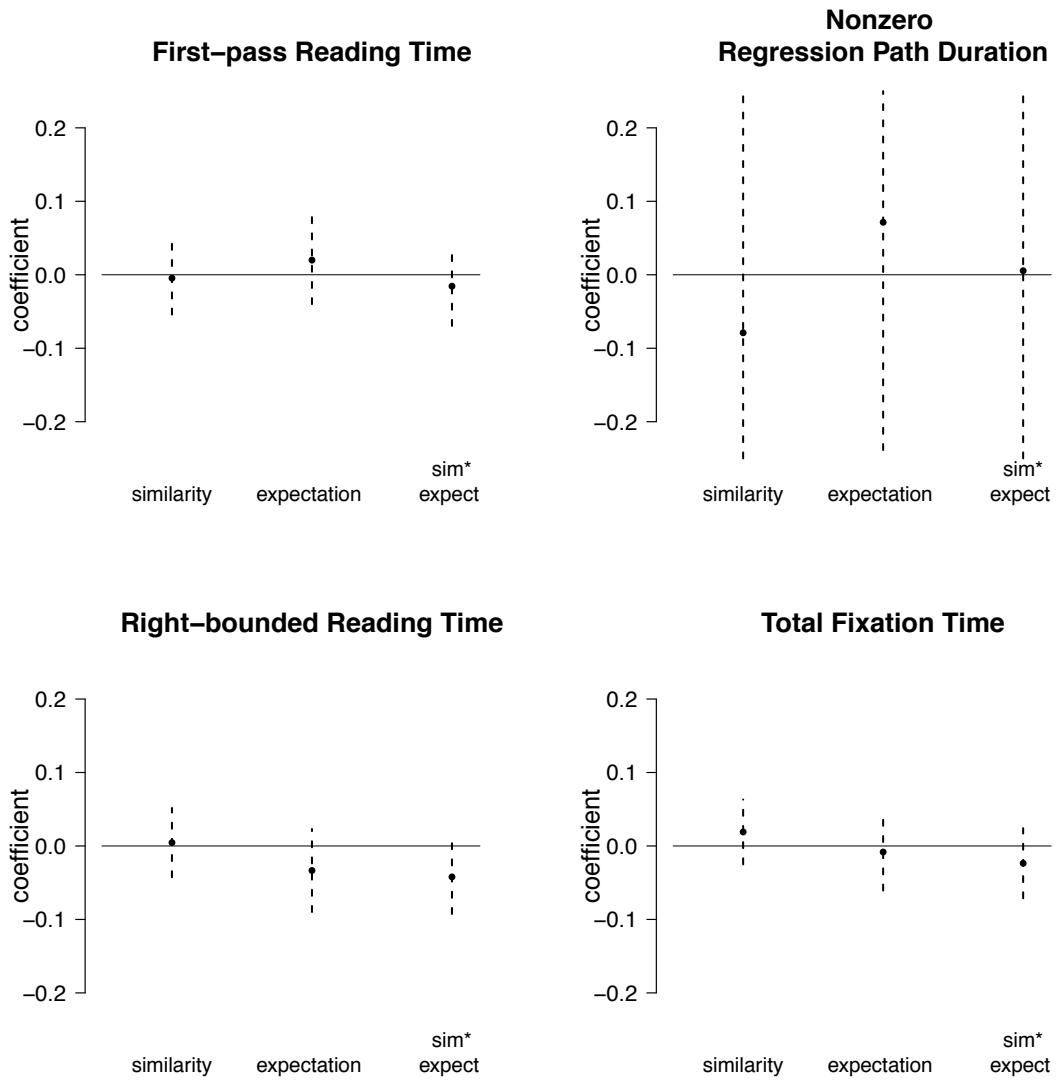


*Figure 4.14.* : PP-ENDING spillover HPD intervals showing the estimated regression coefficients for the expectation effect,similarity between the target and distracter subjects, and their interaction.

verb were constant across all versions of a sentence; and (b) because spillover was included as a predictor in the statistical analyses.

While the effect of similarity was different from the effect of semantic fit, both high similarity and a high target–distracter ratio of semantic fit predicted decrements in performance. High similarity between the distracter and target subject also reliably

increased reading times.

The similarity variable may still partially reflect retrieval cue overlap because if words share retrieval cues, they logically are more similar than if they did not have those cues in common. Even with salient retrieval cues included as a subset of the features contributing to similarity ratings, similarity's ability to predict interference effects in this experiment tentatively support to the broader conception of interference outlined by Gordon et al. (2006) and Van Dyke and McElree (2006). This point will be revisited in this chapter's Discussion.

At the same time, it must be acknowledged that the similarity effect could be an indicator of something other than retrieval interference. The similarity effect was seen in first-fixation, which is commonly taken to reflect lexical-access processing (Clifton, Staub, & Rayner, 2007). Since the verb must be recognized before its argument can be retrieved, an interference effect could only manifest in the first fixation on the verb if a large proportion of first-fixations were long enough to achieve both lexical access and retrieval of the subject, or if lexical access had already been achieved through parafoveal preview while fixating the previous word. If preview was the source of the similarity-related slowdown, however, an expectation effect would be predicted for first-pass measures in the pre-verb region. No such effect was found.

Reading times at the target subject also disconfirm the hypothesis that encoding interference would have caused interference effects at the verb. Neither high similarity nor a high semantic-fit ratio score (nor any interactions involving either variable) predicted longer reading times at the target subject.

Without engaging in ad-hoc speculation about a separate mechanism that might explain why similarity would have such an early impact on processing at the verb, the simplest hypothesis is that first-fixations were often long enough to reflect retrieval

processes.

The ambiguity attending the source of the similarity effect highlights an important area of uncertainty in the empirical estimation of interference effects: *How do similarity judgments and semantic fit judgments differ*? A conclusive treatment of this question exceeds the scope of this paper; but the eyetracking results of this experiment motivate the question by demonstrating that the processing correlates of semantic fit and similarity are not the same. In other words: Some discrepancy between semantic fit and similarity, as they have been measured here, must explain why a similarity effect was found, even though a significant semantic fit effect was not. So what is the discrepancy? Is one norming paradigm more sensitive than the other to the semantic features actually involved in on-line retrieval processes? Is either method a more valid predictor of interference effects than the other; or do they measure different sources of memory difficulty?

Intuitively, the semantic fit measure seems to have higher construct validity inasmuch as it directly asks participants whether each of two candidate arguments matches whatever features a verb might require of its argument. The paradigm has also proven predictive in previous experiments (Gordon et al., 2001; Van Dyke & McElree, 2006). But those experiments also made much stronger encoding and rehearsal demands of the participants while they read a sentence. In the Van Dyke and McElree study, three potentially interfering nouns were read in a list before reading a sentence, and participants were asked to (a) read the list aloud as many times as possible in three seconds, (b) retain the words in memory while reading, and (c) write the whole list in the correct serial order after reading the sentence.

It is plausible that the cues used by semantic fit raters to generate their responses correspond to the cues used online during retrieval, and the ratings failed to predict

interference effects in this experiment due to some other circumstance; but it is equally plausible that semantic fit ratings are an insensitive measure, and that extra efforts are a necessary condition to ensure thorough encoding (and likely rehearsal) of interfering words and induce detectable interference effects. Several other studies have found interference effects without requiring similar strategies (Lewis, 1996; Gordon et al., 2001; Gordon, Hendrick, & Johnson, 2004; Van Dyke & Lewis, 2003), but none have used the semantic fit rating paradigm to predict those effects. Instead, they targeted specific features that were known to be shared between interfering words in their materials.

Perhaps the best way forward, towards understanding how similarity and semantic fit judgments differ, is to focus on how the norming paradigms' task environments differ. One such difference was anticipated by this experiment, and an attempt was made to offset it by adapting the semantic fit scores into a ratio that expressed something about the relationship between subject NPs, like the similarity ratings do. Still, other important differences between the norming paradigms remain, which could affect the predictive validity and reliability of both norming paradigms.

One important aspect of the semantic fit paradigm is that participants view all possible subject nouns simultaneously on each trial. This might have the effect of compressing variance in responses for all referents except the extremely ill-fitting words, assuming that participants err towards accepting marginally-good subjects whenever possible. Despite using a wider rating scale (1–9 as opposed to 1–7 for the similarity experiment, Figure IV that the distribution of semantic fit ratings is very top-heavy. In contrast, similarity ratings follow a bimodal distribution. Greater variance in responses to the similarity rating task may represent a simple but adequate explanation for the difference between similarity and semantic fit as predictors

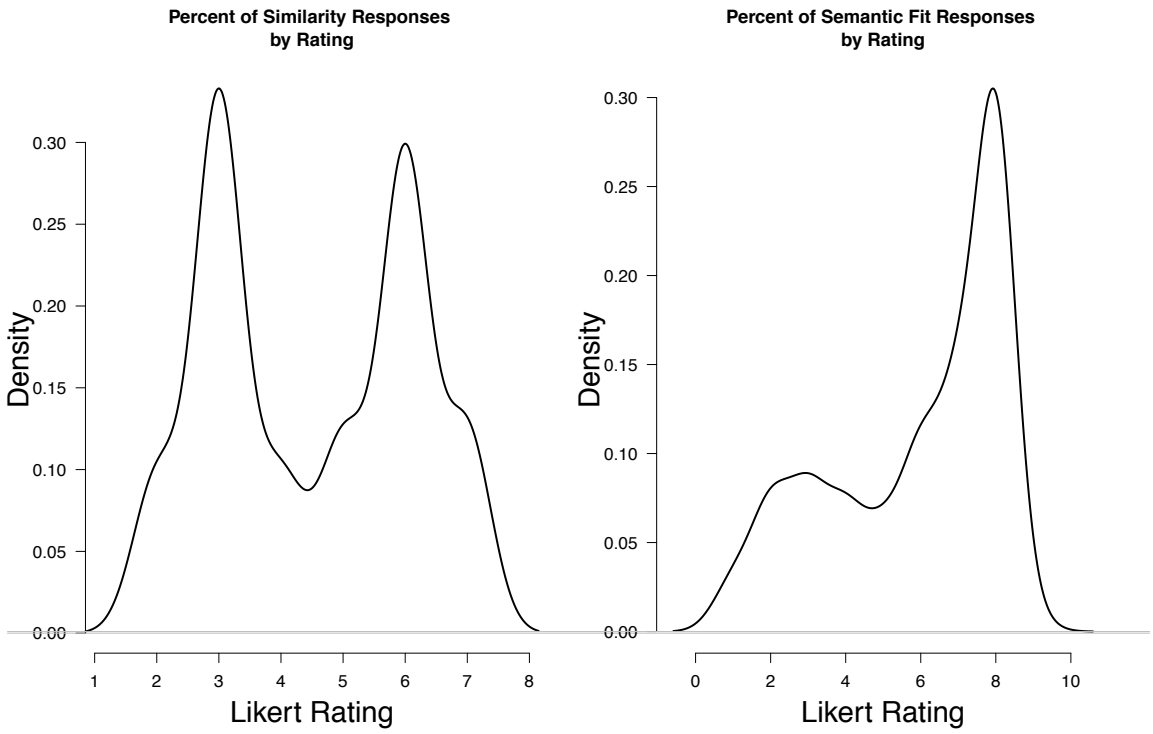


*Figure 4.15.* : Density plots of responses to the similarity judgment task (left) and the semantic fit judgment task (right). Higher ratings indicate higher similarity or stronger semantic fit

of comprehension difficulty.

Discriminating between the effects of similarity and retrieval-cue overlap, if there is a distinction to be made, will be an important step in empirically testing a central architectural claim of the LV05 model, and in shaping memory retrieval models as a whole. But the possibilities for future research should not be allowed to overshadow the insights that have been gained in this eyetracking experiment. The experiment found evidence for separable and non-interactive effects of semantic fit, lexical expectation and similarity at the site of memory retrievals.

# CHAPTER V

# Conclusion

Summary of results and theoretical conclusions

In Chapter III, I presented evidence that the distance intervening between a dependent and its head is a primary determinant of comprehension difficulty. Demonstrating that locality has an effect in simple sentences plays a key role in discriminating between memory-based predictions concerning locality effects others that implicate independent sources of complexity. Specifically, the four experiments presented in Chapter III are an important counterpoint to expectation-based accounts asserting that difficulty integrating long-distance dependencies can be explained without appealing to memory limitations. The high frequency, short length words of the latter two experiments, and the canonical Subject-Verb-Object structures make the locality effects found there difficult to explain under expectation-based accounts; and the lack of argument extraction makes them inexplicable by some other syntactic-complexity accounts.

Having argued for the necessity of some account for memory processes in parsing theory, Chapter IV shifted focus from trying to discriminate between the predictions of memory theory and expectation-based theories like surprisal, on one hand, to exploring the conjunctive effects of memory limitations and expectation-based process-

ing, on the other. An eyetracking study tested the as-yet-empirically-unsupported assumption that expectation-based processing, putatively involving pre-activation of predicted lexical input, occurs independently from interference. Separate effects of memory difficulty and expectation were both found at the resolution of a subject-verb dependency. However, there was no evidence of an interaction between semantic expectation and similarity-based interference.

Taken together, Experiments 1–5 support the conclusion that short-term memory processes must be accounted for (because they correctly predict locality phenomena where other theories, including extant implementations of surprisal, do not). Additionally, they establish that expectation and retrieval interference both have independent (and ostensibly non-interactive) effects on sentence comprehension. Both conclusions constrain the development of theories about parsing. First, they suggest that any complete parsing model must account for both memory processes and expectation effects. Second, they point towards architectures that do not produce interactive effects of expectation and interference.

## Future directions

Having identified the theoretical import of the evidence from five experiments, I now turn my attention to where the preceding research might lead. Many open questions concerning the memory processes and expectation still remain, and answering each of them will refine our understanding of the structure of memory and both the enabling and limiting consequences of that architecture for language processing.

First: While the locality experiments produced novel evidence that locality effects are not restricted to a small set of complex structures in English, the exact mechanism that produces those effects remains elusive. The simple sentences examined

in the latter experiments (3 and 4) are not easily explained by non-memory-based parsing theories. Although memory theories do predict locality effects, this paper does not reach any new conclusions about whether they arise from activation decay, interference, or both. Locality effects are predicted by decay-based theories, but as I pointed out in that chapter, recent research aimed at finding any significant effect of decay have been unable to verify that it plays any substantial role in memory performance. This is not a trivial point. With no (obvious) sources of retroactive interference to explain difficulty integrating subject-verb dependency, memory-based theories may make the right prediction, but no one has articulated in specific terms precisely how the prediction could plausibly be supported by what we know about the memory system. Perhaps the degradation of memory traces that has been attributed to decay is actually a result of retroactive interference, but our understanding of the sources of retroactive interference is underdeveloped. This is one possibility, but admittedly only a speculative hypothesis. Without disambiguating empirical evidence (from any experiments, not specifically those presented above), the exact cognitive underpinnings of locality effects remain unspecified— but the experiments presented here offer compelling evidence that any complete theory of sentence comprehension must be able to account for them.

Chapter IV posed important questions surrounding the relative merits of existing methods for estimating interference, and their utility as operationalized predictors of interference. Depending upon one's theoretical predispositions, the differential influences of similarity and semantic fit might inspire different questions. One might ask, "Are similarity effects just a more sensitive measure of retrieval cue overlap in experiments similar to Experiment 5?", and "What role do the tasks themselves play in either detecting or obscuring the effects of retrieval interference?" Alternatively,

one might well ask, "What are the similarity ratings measuring that is not accounted for by semantic fit," and "Have interference effects been mischaracterized as byproducts of retrieval cue overlap, when they are really similarity effects?" Interpreting the results of Experiment 5 as a serious challenge to the claim that retrieval-cue overlap causes interference would be a stretch too far. The similarity effect found at the critical verb may have been a novel observation of the effect of lexical similarity quite apart from the effect of retrieval cue interference; but the similarity judgments collected just as well may have captured variation in retrieval cue overlap more effectively than the semantic fit judgment paradigm, for reasons explained at the conclusion of Chapter IV. Too little is understood about the tools used to estimate retrieval cue overlap and lexical similarity to draw strong conclusions on this question. Ergo, we cannot ascertain whether raw lexical similarity and semantic fit have completely separate effects on comprehension, or whether, to the contrary, they are completely confounded and they only manifest differently because of measurement error in the experiments that estimate them. Every one of the questions posed above warrants earnest experimental investigation. With additional research it may be possible to precisely quantify the dimensions of similarity captured by semantic fit ratings and similarity ratings, respectively. The answers are well beyond the scope of this thesis; but they point out what I believe to be one of the strengths of this research agenda: it generates numerous important questions and motivates research that will advance models of sentence-processing— conceivably also advancing domain-general theories of memory, if the basis of retrieval interference can be resolved.

In the near term, Experiment 5 points to several actionable research ideas that might contribute to solving the questions posed above. Some of these ideas are

actively being pursued currently. For instance, I am in the first stages of using scan-path analyses (von der Malsburg & Vasishth, 2011) to analyze characteristic patterns of fixation across the sentences in Experiment 5. This technique may help identify signature patterns in regression triggers and landing sites that dissociate the effects of lexical similarity and semantic fit. Interference effects at the verb, for instance, might tend to be followed by regressions to the target subject to re-encode it, or even to the distracter subject, as readers back-track to the source of difficulty. Scan-path analyses do not always yield clear insights, in part because readers do not always target regressions selectively to a region that supplies information they failed to retrieve or need to re-encode for other reasons. At the same time, there is the potential for some of the less easily interpreted eyetracking results to become clearer as part of a larger pattern. In particular, scan-path analyses could determine whether the similarity effect in total fixation times before the verb is a reflection of re-reading after fixating the verb, or whether the slowdown associated with high distracter–target NP similarity is effectively smeared across multiple first-pass measures.

Several other experiments might follow from Experiment 5. As a sample of the possibilities, consider the following.

1. A parallel experiment in self-paced reading could confirm the effects found in this experiment, and also contribute to the development of the eye-movement control model alluded to in Chapter III and presented in (Bartek et al., 2011). This experiment is, in fact, already underway.

2. New materials, incorporating concretely identifiable sources of retrieval interference like noun-phrases specificity (Gordon et al., 2001) may clarify the differences between the predictive validity of semantic fit measurement and similarity judgments,

and suggest new interpretations of their relationship. Controlling the features that contribute to retrieval-cue overlap would improve the chances of manipulating orthogonal (or nearly orthogonal) dimensions of similarity, and consequently disentangling any effects that might arise from either similarity or retrieval interference.

3. Assuming the result of Experiment 5 are replicated, the same paradigm can be extended to include a locality manipulation. This step would be a logical extension of some expectation-based research in which anti-locality effects emerge from the consolidation of syntactic expectation over the course of a long-distance dependency (Konieczny, 2000), and connect Chapter III to the objectives of Chapter IV.

Each of these experiments could substantially advance theories of both linguistic expectation and memory function in sentence-processing. In the meantime, the experiments that generated the ideas listed above make their own substantial contribution by providing strong evidence for the necessity of short-term memory effects in parsing models and indicating that a non-interactive architecture integrating memory effects and expectation effects could well be an accurate reflection of the language-processing architecture of the brain.

# APPENDICES

| item | version | % target | N |
|---|---|---|---|
| 1 | a | 42 | 12 |
| 1 | b | 23 | 13 |
| 1 | c | 0 | 10 |
| 1 | d | 0 | 13 |
| 2 | a | 50 | 8 |
| 2 | b | 42 | 12 |
| 2 | c | 9 | 11 |
| 2 | d | 18 | 11 |
| 3 | a | 0 | 11 |
| 3 | b | 0 | 12 |
| 3 | c | 0 | 11 |
| 3 | d | 0 | 9 |
| 4 | a | 55 | 11 |
| 4 | b | 44 | 9 |
| 4 | c | 0 | 10 |
| 4 | d | 30 | 10 |
| 5 | a | 22 | 9 |
| 5 | b | 0 | 10 |
| 5 | c | 12 | 8 |
| 5 | d | 0 | 7 |
| 6 | a | 44 | 9 |
| 6 | b | 30 | 10 |
| 6 | c | 25 | 12 |
| 6 | d | 10 | 10 |
| 7 | a | 100 | 9 |
| 7 | b | 70 | 10 |
| 7 | c | 18 | 11 |
| 7 | d | 45 | 11 |

Table .11:: Cloze completion results from Experiment 5.1. The six versions of each sentence are arbitrarily labeled A – D. No single version of a sentence (for instance, version "A") always had the same characteristics (for instance, strong expectation and high semantic fit).

| item | version | % target | N |
|---|---|---|---|
| 8 | a | 73 | 11 |
| 8 | b | 45 | 11 |
| 8 | c | 38 | 8 |
| 8 | d | 10 | 10 |
| 9 | a | 67 | 12 |
| 9 | b | 64 | 11 |
| 9 | c | 0 | 14 |
| 9 | d | 8 | 13 |
| 10 | a | 67 | 9 |
| 10 | b | 55 | 11 |
| 10 | c | 12 | 8 |
| 10 | d | 0 | 10 |
| 11 | a | 33 | 9 |
| 11 | b | 64 | 11 |
| 11 | c | 0 | 13 |
| 11 | d | 25 | 8 |
| 12 | a | 0 | 13 |
| 12 | b | 45 | 11 |
| 12 | c | 9 | 11 |
| 12 | d | 27 | 11 |
| 13 | a | 20 | 10 |
| 13 | b | 23 | 13 |
| 13 | c | 0 | 9 |
| 13 | d | 0 | 11 |
| 14 | a | 55 | 11 |
| 14 | b | 64 | 11 |
| 14 | c | 0 | 12 |
| 14 | d | 0 | 12 |

Table .12:: Cloze completion results from Experiment 5.1 *continued*. The six versions of each sentence are arbitrarily labeled A – D. No single version of a sentence (for instance, version "A") always had the same characteristics (for instance, strong expectation and high semantic fit).

| item | version | % target | N |
|---|---|---|---|
| 15 | a | 46 | 13 |
| 15 | b | 10 | 10 |
| 15 | c | 0 | 9 |
| 15 | d | 0 | 11 |
| 16 | a | 25 | 12 |
| 16 | b | 12 | 8 |
| 16 | c | 0 | 11 |
| 16 | d | 0 | 11 |
| 17 | a | 9 | 11 |
| 17 | b | 50 | 12 |
| 17 | c | 0 | 13 |
| 17 | d | 0 | 12 |
| 18 | a | 50 | 14 |
| 18 | b | 36 | 11 |
| 18 | c | 8 | 12 |
| 18 | d | 9 | 11 |
| 19 | a | 25 | 12 |
| 19 | b | 36 | 11 |
| 19 | c | 0 | 14 |
| 19 | d | 8 | 12 |
| 20 | a | 38 | 13 |
| 20 | b | 42 | 12 |
| 20 | c | 20 | 10 |
| 20 | d | 21 | 14 |
| 21 | a | 30 | 10 |
| 21 | b | 33 | 9 |
| 21 | c | 9 | 11 |
| 21 | d | 11 | 9 |

Table .13:: Cloze completion results from Experiment 5.1 *continued*. The six versions of each sentence are arbitrarily labeled A – D. No single version of a sentence (for instance, version "A") always had the same characteristics (for instance, strong expectation and high semantic fit).

| item | version | % target | N |
|---|---|---|---|
| 22 | a | 40 | 10 |
| 22 | b | 0 | 5 |
| 22 | c | 23 | 13 |
| 22 | d | 23 | 13 |
| 23 | a | 62 | 8 |
| 23 | b | 27 | 11 |
| 23 | c | 12 | 8 |
| 23 | d | 50 | 6 |
| 24 | a | 55 | 11 |
| 24 | b | 50 | 10 |
| 24 | c | 30 | 10 |
| 24 | d | 18 | 11 |
| 25 | a | 22 | 9 |
| 25 | b | 45 | 11 |
| 25 | c | 0 | 7 |
| 25 | d | 0 | 11 |
| 26 | a | 57 | 14 |
| 26 | b | 62 | 8 |
| 26 | c | 46 | 13 |
| 26 | d | 17 | 12 |
| 27 | a | 9 | 11 |
| 27 | b | 45 | 11 |
| 27 | c | 22 | 9 |
| 27 | d | 36 | 11 |
| 28 | a | 10 | 10 |
| 28 | b | 9 | 11 |
| 28 | c | 14 | 14 |
| 28 | d | 10 | 10 |

Table .14:: Cloze completion results from Experiment 5.1 *continued*. The six versions of each sentence are arbitrarily labeled A – D. No single version of a sentence (for instance, version "A") always had the same characteristics (for instance, strong expectation and high semantic fit).

| item | version | % target | N |
|---|---|---|---|
| 29 | a | 11 | 9 |
| 29 | b | 0 | 9 |
| 29 | c | 0 | 9 |
| 29 | d | 0 | 14 |
| 30 | a | 42 | 12 |
| 30 | b | 17 | 12 |
| 30 | c | 0 | 11 |
| 30 | d | 8 | 12 |
| 31 | a | 20 | 10 |
| 31 | b | 50 | 12 |
| 31 | c | 0 | 13 |
| 31 | d | 9 | 11 |
| 32 | a | 38 | 13 |
| 32 | b | 33 | 6 |
| 32 | c | 0 | 8 |
| 32 | d | 27 | 11 |
| 33 | a | 42 | 12 |
| 33 | b | 36 | 14 |
| 33 | c | 9 | 11 |
| 33 | d | 0 | 13 |
| 34 | a | 75 | 12 |
| 34 | b | 30 | 10 |
| 34 | c | 25 | 12 |
| 34 | d | 29 | 7 |

Table .15:: Cloze completion results from Experiment 5.1 *continued*. The six versions of each sentence are arbitrarily labeled A – D. No single version of a sentence (for instance, version "A") always had the same characteristics (for instance, strong expectation and high semantic fit).

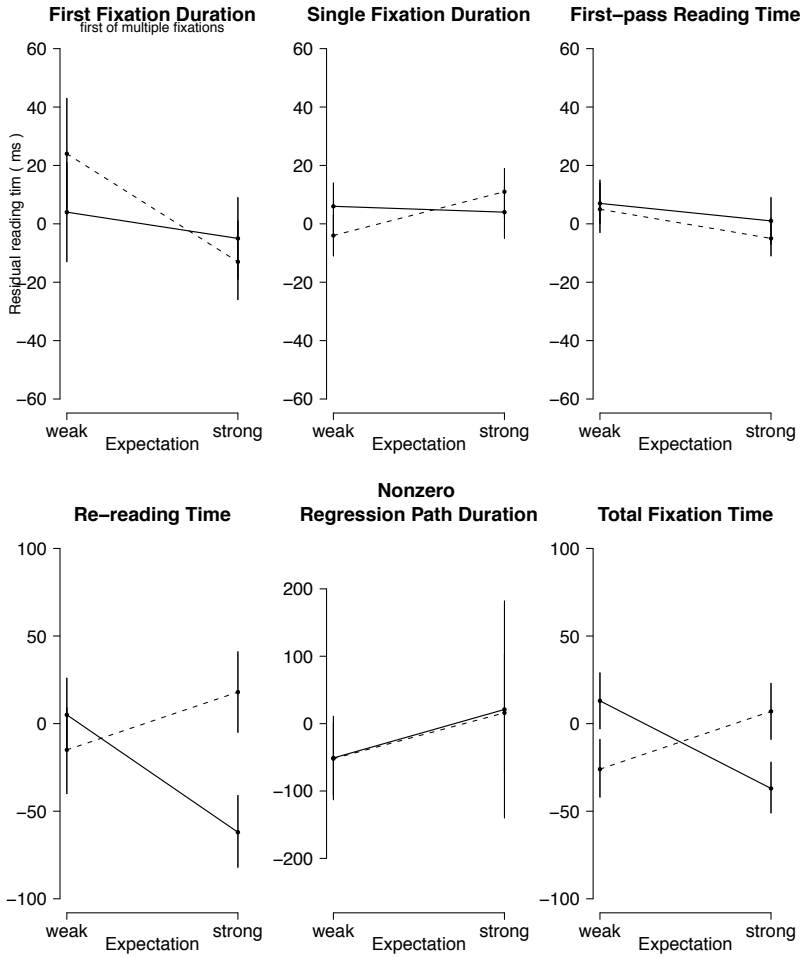


*Figure .16.* : PRE-CRITICAL REGION residual reading times: expectation plotted against semantic-fit ratio. Variance attributed to spillover, word length, word frequency, and plausibility has been factored out to show the relationship of interference and expectation. The four means plotted are taken from a median split performed on continuous predictors. High-fit-ratio sentences are shown with dotted lines; low-fit-ratio sentences are shown with solid lines.

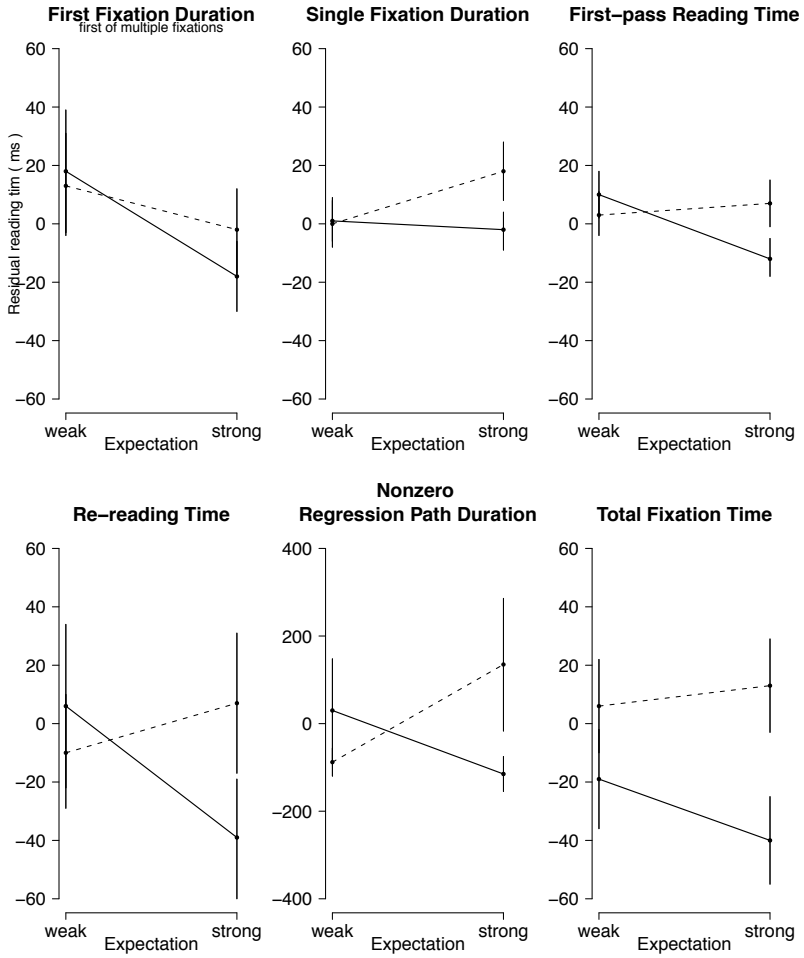


*Figure .17.* : PRE-CRITICAL REGION residual reading times: expectation plotted against similarity. Variance attributed to spillover, word length, word frequency, and plausibility has been factored out to show the relationship of semantic fit and expectation. The four means plotted are taken from a median split performed on continuous predictors. High-similarity sentences are shown with dotted lines; low-similarity sentences are shown with solid lines.

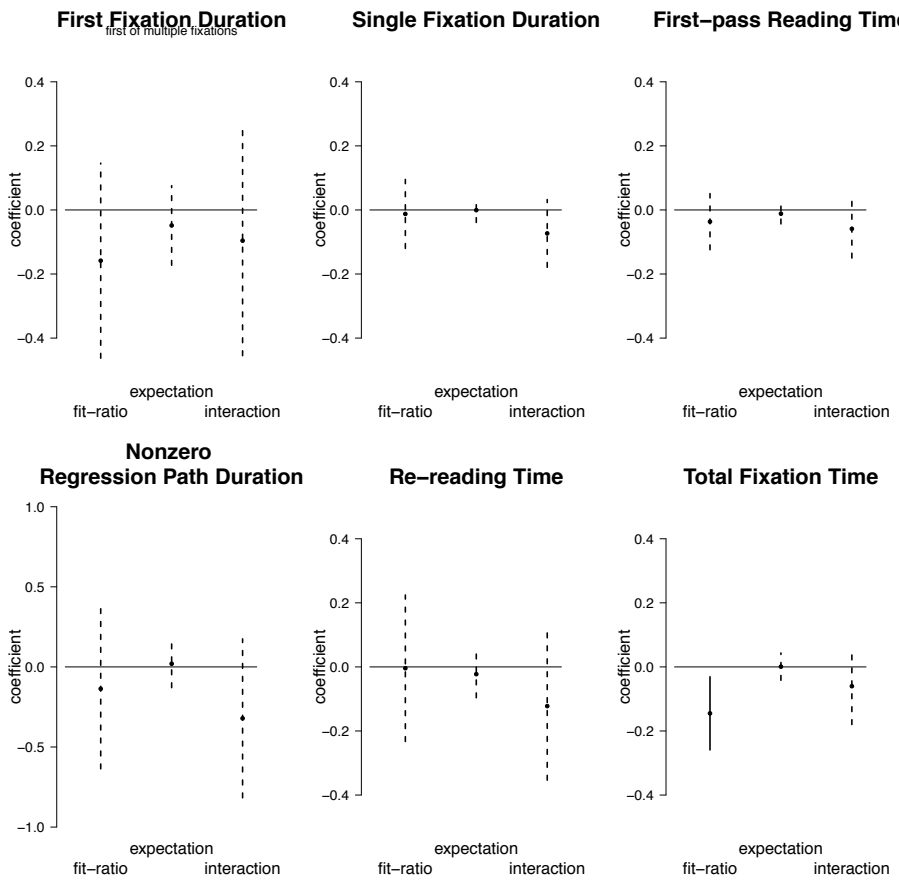


*Figure .18.* : PRE-CRITICAL REGION HPD showing the estimated regression coefficients for the expectation effect, interference effect, and their interaction. These intervals serve as significance tests at an alpha level of .05. Intervals that include zero are non-significant; those that do not include zero are significant.

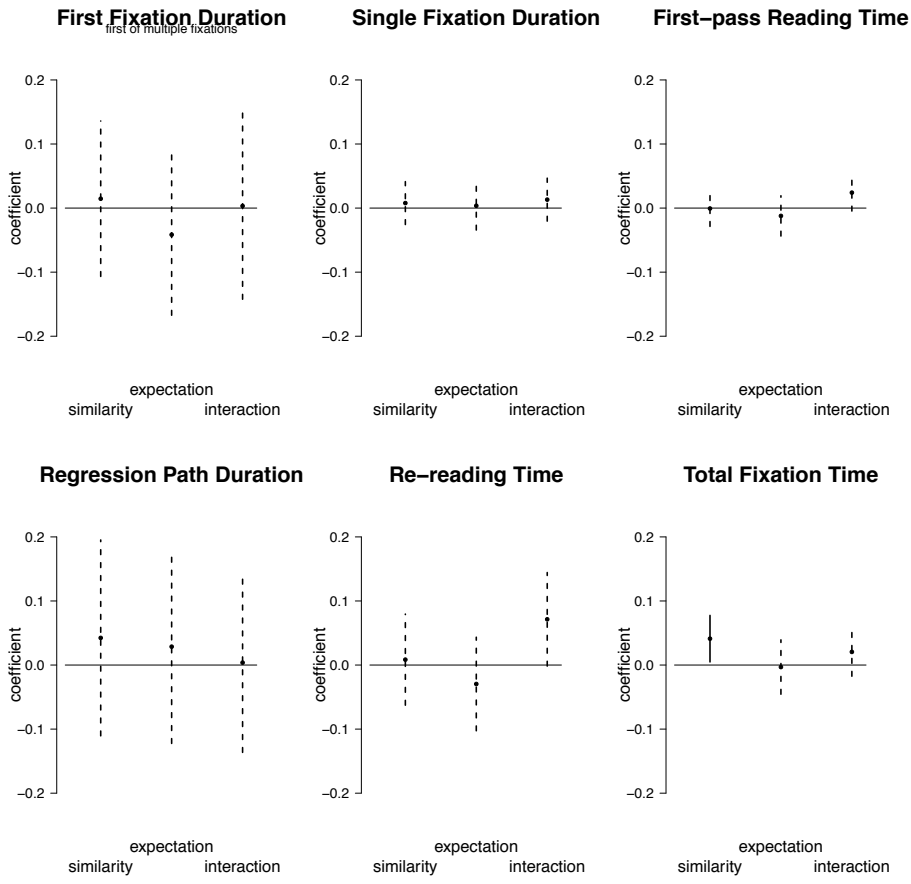


*Figure .19.* : PRE-CRITICAL REGION HPD intervals showing the estimated regression coefficients for the expectation effect,similarity, and their interaction.

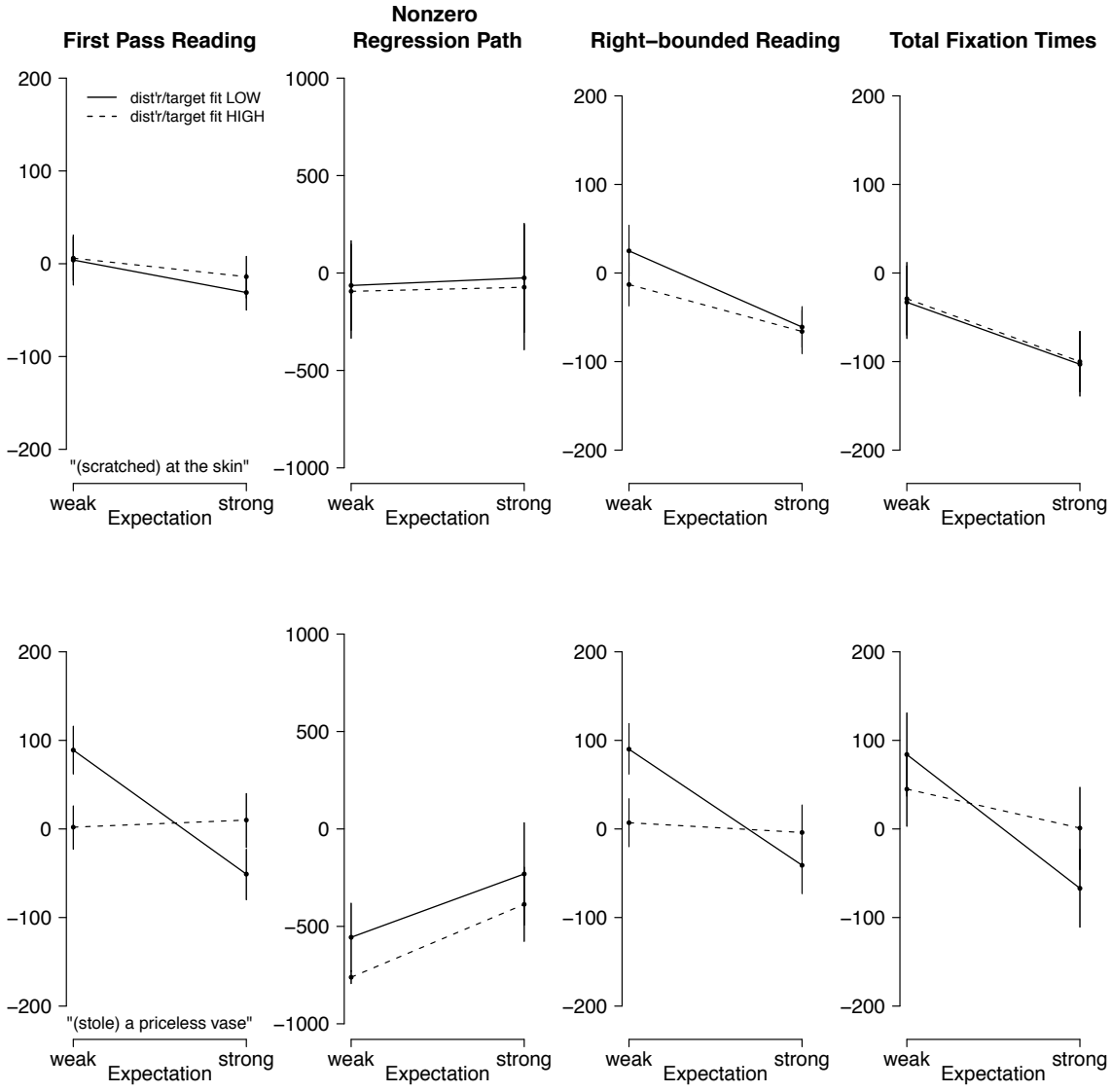


*Figure .20.* : SPILLOVER REGION residual reading times. Top row: " (scratched) at the skin" sentences; Bottom row: "(stole) a priceless vase" Both rows plot expectation against semantic fit. A subset of fixation measures are shown because first-fixation, single-fixation and re-reading time are not interpretable when aggregating over several regions.

# BIBLIOGRAPHY

# BIBLIOGRAPHY

Altmann, G., Nice, K. van, Garnham, A., & Henstra, J. (1998). Late closure in context. *Journal of Memory and Language*, *38*, 459–484.

Anderson, J. R., Bothell, D., Byrne, M. D., Douglass, S., Lebiere, C., & Qin, Y. (2004). An integrated theory of the mind. *Psychological Review*, *111*(4), 1036-1060. Available from `http://act-r.psy.cmu.edu/publications/pubinfo.php?id=526`

Ashby, J., & Rayner, K. (2005). Eye movements of highly-skilled and average readers: differential effects of frequency and predictability. *Quarterly Journal of Experimental Psychology*, *58A*, 1065–1086.

Atkinson, R. C., & Shiffrin, R. M. (1968). Human memory: A proposed system and its control processes. In K. W. Spense & J. T. Spense (Eds.), *The psychology of learning and motivation.* Academic Press.

Baddeley, A. (1986). *Working memory* (A. D. Baddeley, Ed.). Oxford: Oxford University Press.

Baddeley, A., Thomson, N., & Buchanan, M. (1975). Word length and structure of short-term memory. *Journal of verbal learning and verbal behavior*, *14*, 575–589.

Baddeley, A., & Warrington, E. (1970). Amnesia and the distinction between long- and short-term memory. *Journal of verbal learning and verbal behavior*, *9*, 176–189.

Bartek, B., Lewis, R. L., Vasishth, S., & Smith, M. R. (2011). In search of on-line locality effects in sentence comprehension. *Journal of Experimental Psychology: Learning, Memory and Cognition*. (In-press)

Bates, D., & Sarkar, D. (2007). lme4: Linear mixed-effects models using s4 classes [Computer software manual]. (R package version 0.9975-11)

Berman, M. G., Jonides, J., & Lewis, R. L. (2009). In search of decay in verbal short-term memory. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *35*, 317–333.

Boston, M., Hale, J., Kliegl, R., Patil, U., & Vasishth, S. (2008). Parsing costs as predictors of reading difficulty: an evaluation using the potsdam sentence corpus. *Journal of Eye Movement Research*, *2*, 1–12.

Brysbaert, M., & Vitu, F. (1998). Word skipping: Implications for theories of eye movement and control in reading. In G. Underwood (Ed.), *Eye guidance in reading and scene perception* (pp. 125–148). Oxford, England: Elsevier.

Chomsky, N. (1965). *Aspects of the theory of syntax.* Cambridge, MA: MIT Press.

Clifton, C., Staub, A., & Rayner, K. (2007). Eye movements in reading words and sentences. In R. van Gompel (Ed.), *Eye movements: A window on mind and brain* (pp. 341–372). Amsterdam: Elsevier.

Daneman, M., & Carpenter, P. A. (1980). Individual differences in working memory and reading. *Journal of Verbal Learning and Verbal Behavior*, *19*, 450–466.

Ehrlich, S., & Rayner, K. (1981). Contextual effects on word perception and eye movements during reading. *Journal of Verbal Learning and Verbal Behavior*, *20*, 641–655.

Ferreira, F., & Clifton, J., C. (1986). The independence of syntactic processing. *Journal of Memory and Language*, *25*, 348–368.

Ferreira, F., & Henderson, J. (1990). Use of verb information in syntactic parsing: Evidence from eye movements and word-by-word self-paced reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *16(4)*, 555–568.

Ferreira, F., & Henderson, J. M. (1991). Recovery from mis-analyses of garden-path sentences. *Journal of Memory and Language*, *30*, 725–745.

Fodor, J. D., & Ferreira, F. (Eds.). (1998). Hillsdale, NJ: Erlbaum.

Foss, D. (1982). A discourse on semantic priming. *Cognitive Psychology*, *14*, 590–607.

Foss, D., & Ross, I. (1983). Great expectations: Context effects during sentence processing. In G. F. d'Arcais & R. Jarvella (Eds.), *The process of language understanding.* New York: John Wiley.

Frazier, L. (1987). Syntactic processing: Evidence from Dutch. *Natural Language and Linguistic Theory*, *5*, 519–560.

Frazier, L., & Clifton, C. (1989). Successive cyclicity in the grammar and the parser. *Language and Cognitive Processes*, *4*, 93–126.

Frazier, L., & Fodor, J. D. (1978). The sausage machine: A two-stage parsing model. *Cognition*, *6*, 291-325.

Frazier, L., & Rayner, K. (1982). Making and correcting errors during sentence comprehension: Eye movements in the analysis of structurally ambiguous sentences. *Cognitive Psychology*, *14*, 178–210.

Frisson, S., Rayner, K., & Pickering, M. (2005). Effects of contextual predictability and transitional probability on eye movements during reading. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *25*, 1366–1383.

Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models.* Cambridge, UK: Cambridge University Press.

Gennari, S., & MacDonald, M. (2008). Semantic indeterminacy in object relative clauses. *Journal of Memory and Language*, *58*, 161–187.

Gennari, S., & MacDonald, M. (2009). Linking production and comprehension processes: the case of relative clauses. *Cognition*, *111*, 1–23.

Gibson, E. (1991). *A computational theory of human linguistic processing: Memory limitations and processing breakdown.* Unpublished doctoral dissertation, Carnegie Mellon University, Pittsburgh, PA.

Gibson, E. (1998). Linguistic complexity: Locality of syntactic dependencies. *Cognition*, *68*, 1–76.

Gibson, E. (2000). Dependency locality theory: A distance-based theory of linguistic complexity. In A. Marantz, Y. Miyashita, & W. O'Neil (Eds.), *Image, language, brain: Papers from the first mind articulation project symposium.* Cambridge, MA: MIT Press.

Gibson, E. (2007). *Locality and anti-locality effects in sentence comprehension.* Leipzig, Germany. (Presented at Workshop on processing head-final languages)

Gibson, E., Pearlmutter, N., Canseco-Gonzales, E., & Hickock, G. (1996). Recency preference in the human sentence processing mechanism. *Cognition*, *59*, 23–59.

Gibson, E., Pearlmutter, N., & Torrens, V. (1999). Recency and lexical preference in spanish. *Memory and Cognition*, *27*, 603–611.

Gibson, E., & Warren, T. C. (2004). Reading time evidence for intermediate linguistic structure in long-distance dependencies. *Syntax*, *7*, 55–78.

Gillund, G., & Shiffrin, R. M. (1984). A retrieval model of both recognition and recall. *Psychological Review*, *91*, 1-65.

Gordon, P. C., Hendrick, R., & Johnson, M. (2001). Memory interference during language processing. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *27(6)*, 1411–1423.

Gordon, P. C., Hendrick, R., & Johnson, M. (2004). Effects of noun phrase type on sentence complexity. *Journal of Memory and Language*, *51*, 97–104.

Gordon, P. C., Hendrick, R., Johnson, M., & Lee, Y. (2006). Similarity-Based Interference During Language Comprehension: Evidence from Eye Tracking During Reading. *Journal of Experimental Psychology: Learning Memory and Coginition*, *32*(6), 1304–1321.

Gordon, P. C., Hendrick, R., & Levine, W. H. (2002). Memory-load interference in syntactic processing. *Psychological Science*, 425–430.

Gough, P., Alford, J., & Holly-Wilcox, P. (1981). Words and contexts. In *Perception of print: Reading research in experimental psychology.* Hillsdale, NJ: Erlbaum.

Grodner, D., & Gibson, E. (2005). Consequences of the serial nature of linguistic input. *Cognitive Science*, *29*, 261–290.

Grodner, D., Gibson, E., & Tunstall, S. (2002). Syntactic complexity in ambiguity resolution. *Journal of Memory and Language*, *46*, 267–295.

Grodzinsky, Y. (2000). The neurology of syntax: Language use without broca's area. *Behavioral and Brain Sciences*, *23*, 1–21.

Hale, J. (2001). A probabilistic Earley parser as a psycholinguistic model. In *Proceedings of the Second Meeting of the North American Chapter of the Association for Computational Linguistics.* Pittsburgh, PA.

Hebb, D. O. (1961). Distinctive features of learning in the higher animal. In J. F. Delafresnaye (Ed.), *Brain mechanisms and learning: A symposium.* London: Oxford: Blackwell Scientific Publications.

Hyona, J. (1993). Effects of thematic and lexical priming on readers' eye movements. *Scandinavian Journal of Psychology*, *34*, 293–304.

Jaeger, F., Fedorenko, E., & Gibson, E. (2005). Dissociation between production and comprehension difficulty. In *Proceedings of the cuny sentence processing conference.* Arizona.

Jonides, J., Lewis, R. L., Nee, D. E., Lustig, C. A., Berman, M. G., & Moore, K. S. (2008). The mind and brain of short-term memory. *Annual Review of Psychology*, *59*, 15.1–15.32.

Jurafsky, D. (2002). Probabilistic modeling in psycholinguistics: Linguistic comprehension and production. In R. Bod, J. Hay, & S. Jannedy (Eds.), *Probabilistic linguistics* (pp. 39–96). MIT Press.

Just, M. A., & Carpenter, P. A. (1992). A capacity theory of comprehension: Individual differences in working memory. *Psychological Review*, *99(1)*, 122–149.

Kennison, S. (2002). Comprehending noun phrase arguments and adjuncts. *Journal of Psycholinguistic Research*, *31(1)*, 65–81.

Kimball, J. (1973). Seven principles of surface structure parsing in natural language. *Cognition*, *2*, 15–47.

Kintsch, W. (1988). *Comprehension: A paradigm for cognition.* New York: Cambridge University Press.

Konieczny, L. (2000). Locality and parsing complexity. *Journal of Psycholinguistic Research*, *29*, 627-645.

Kruschke, J. K. (2010). What to believe: Bayesian methods for data analysis. *Trends in Cognitive Sciences*, *14*, 293–300.

Lau, E. F. (2009). *The predictive nature of language comprehension.* Unpublished doctoral dissertation, University of Maryland, MD.

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, *106*(3), 1126-1177.

Lewis, R. L. (1996). Interference in short-term memory: The magical number two (or three) in sentence processing. *Journal of Psycholinguistic Research*, *25(1)*, 93–115.

Lewis, R. L., & Vasishth, S. (2005, May). An activation-based model of sentence processing as skilled memory retrieval. *Cognitive Science*, *29*, 1-45.

Lewis, R. L., Vasishth, S., & Van Dyke, J. (2006, October). Computational principles of working memory in sentence comprehension. *Trends in Cognitive Science*, *10*(10), 447-454.

Logie, R. H., Della Sala, S., Laiacona, M., Chalmers, P., & Wynn, V. (1996). Group aggregates and individual reliability: the case of verbal short-term memory. *Memory and Cognition*, *24*, 305–321.

MacDonald, M. C., Perlmutter, N. J., & Seidenberg, M. S. (1994). Lexical nature of syntactic ambiguity resolution. *Psychological Review*, *101(4)*, 676–703.

MacDonald, S., & Shillcock, R. (2003a). Eye movements reveal the on-line computation of lexical probabilities during reading. *Psychological Review*, *112*, 814–840.

MacDonald, S., & Shillcock, R. (2003b). Low-level predictive inference in reading: the influence of transitional probabilities during reading. *Psyhological Science*, *14*, 648–652.

Mahajan, A. K. (1990). *The A/A-bar Distinction and Movement Theory*. Unpublished doctoral dissertation, MIT, Cambridge, MA.

McElree, B. (2000). Sentence comprehension is mediated by content-addressable memory structures. *Journal of Psycholinguistic Research*, *29(2)*, 111–123.

McElree, B., Foraker, S., & Dyer, L. (2003). Memory structures that subserve sentence comprehension. *Journal of Memory and Language*, *48*, 67–91.

Meyer, D., Schvaneveldt, R., & Ruddy, M. (1975). Loci of contextual effects on visual word recognition. In P. Rabbitt & S. Dornic (Eds.), *Attention and performance* (Vol. 5). New York: Academic Press.

Morris, R. (1994). Lexical and message-level sentence context effects on fixation times in reading. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *20*, 92–103.

Nakatani, K., & Gibson, E. (2008). Distinguishing theories of syntactic expectation cost in sentence comprehension: evidence from japanese. *Linguistics*, *46*, 63–86.

Pearlmutter, N. J., & Gibson, E. (2001). Recency in verb phrase attachment. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *27*, 574–590.

Posner, M., & Snyder, C. (1975a). Attention and cognitive control. In P. Rabbitt & S. Dornic (Eds.), *Attention and performance v.* New York.

Posner, M., & Snyder, C. (1975b). Facilitation and inhibition in the processing of signals. In P. Rabbitt & S. Dornic (Eds.), *Attention and performance v.* New York.

Pritchett, B. L. (1992). *Grammatical competence and parsing performance*. Chicago: University of Chicago Press.

R Development Core Team. (2006). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Available from http://www.R-project.org (ISBN 3-900051-07-0)

Rayner, K. (1979). Eye guidance in reading: Fixation locations within words. *Perception*, *8*, 21–30.

Rayner, K., & Well, A. D. (1996). Efects of contextual constraint on eye movements in reading: a further examination. *Psychonomic Bulletin and Review*, *3*, 504–509.

Schustack, M., Ehrlich, S., & Rayner, K. (1987). Local and glocal sources of contextual facilitation in reading. *Journal of Memory and Language*, *26*, 322–340.

Shallice, T., & Warrington, E. (1970). Independent functioning of verbal memory stores: a neuropsychological study. *Quarterly Journal of Experimental Psychology*, *22*, 261–273.

Sharkey, A. J., & Sharkey, N. E. (1992). Weak contextual constraints in text and word priming. *Journal of memory and language*, *31*, 543–572.

Traxler, M., Morris, R., & Seely, R. (2002). Processing subject and object relative clauses: evidence from eye movements. *Journal of Memory and Language*, *35*, 69–90.

Trueswell, J., Tanenhaus, M., & Kello, C. (1993). Verb-specific contraints in sentence processing: Separating effects of lexical preference from garden paths. *Journal of Experimental Psychology*, *19*, 528–553.

Vallar, G., & Papagno, C. (2002). Neuropsychological impairments of verbal short-term memory. In A. Baddeley, M. Kopelman, & B. Wilson (Eds.), *The handbook of memory disorder* (pp. 249–270). Wiley.

Van Dyke, J. (2007). Interference effects from grammatically unavailable constituents during sentence processing. *Journal of Expermiental Psychology: Learning, Memory and Cognition*, *33*, 407–430.

Van Dyke, J., & Lewis, R. L. (2003). Distinguishing effects of structure and decay on attachment and repair: A cue-based parsing account of recovery from misanalyzed ambiguities. *Journal of Memory and Language*, *49*, 285–316.

Van Dyke, J., & McElree, B. (2006). Retrieval interference in sentence comprehension. *Journal of Memory and Language*, *55*, 157–166.

Vasishth, S. (2003). *Working memory in sentence comprehension: Processing Hindi center embeddings*. New York: Garland Press. (Published in the Garland series Outstanding Dissertations in Linguistics, edited by Laurence Horn)

Vasishth, S. (2004). Decay and similarity in sentence processing. In *Technical report, ieice.* Tokyo, Japan.

Vasishth, S., Brüssow, S., Lewis, R. L., & Drenhaus, H. (2008). Processing polarity: How the ungrammatical intrudes on the grammatical. *Cognitive Science*, *4*, 685–712. Available from http://www.ling.uni-potsdam.de/ vasishth/Papers/npi1final.pdf

Vasishth, S., & Drenhaus, H. (2011). *Locality in German.* (Submitted to Dialogue and Discourse)

Vasishth, S., & Lewis, R. L. (2006). Argument-head distance and processing complexity: Explaining both locality and antilocality effects. *Language*, *82*, 767–794.

von der Malsburg, T., & Vasishth, S. (2011). The scanpath signature of syntactic reanalysis. *Journal of Memory and Language*. (Accepted pending minor revisions)

Vosse, T., & Kempen, G. (2000). Syntactic structure assembly in human parsing: a computational model based on competitive inhibition and lexicalist grammar. *Cognition*, *75*, 105–143.

Wanner, E., & Maratsos, M. (1978). An ATN approach to comprehension. In M. Halle, J. Bresnan, & G. A. Miller (Eds.), *Linguistic theory and psychological reality.* Cambridge, MA: MIT Press.

Wu, H. I., & Gibson, E. (2008). Processing Chinese relative clauses in context. In *Proceedings of the 21st cuny conference on sentence processing.* University of North Carolina.