J Psycholinguist Res (2012) 41:105–128
DOI 10.1007/s10936-011-9179-x

ORIGINAL PAPER

Comparisons of Online Reading Paradigms: Eye Tracking, Moving-Window, and Maze

Naoko Witzel · Jeffrey Witzel · Kenneth Forster

Published online: 15 October 2011

Abstract This study compares four methodologies used to examine online sentence processing during reading. Specifically, self-paced, non-cumulative, moving-window reading (Just et al. in J Exp Psychol Gen 111:228–238, 1982), eye tracking (see e.g., Rayner in Q J Exp Psychol 62:1457–1506, 2009), and two versions of the maze task (Forster et al. in Behav Res Methods 41:163–171, 2009)—the lexicality maze and the grammaticality maze—were used to investigate the processing of sentences containing temporary structural ambiguities. Of particular interest were (i) whether each task was capable of revealing processing differences on these sentences and (ii) whether these effects were indicated precisely at the predicted word/region. Although there was considerable overlap in the general pattern of results from the four tasks, there were also clear differences among them in terms of the strength and timing of the observed effects. In particular, excepting sentences that tap into clause-closure commitments, both maze task versions provided robust, "localized" indications of incremental sentence processing difficulty relative to self-paced reading and eye tracking.

Keywords Eye tracking · Moving-window reading · Maze task · Sentence processing

N. Witzel · J. Witzel · K. Forster
Cognitive Science Program, University of Arizona, Tucson, AZ, USA

N. Witzel (✉)
Department of Psychology, University of Texas at Arlington, Arlington, TX 76019, USA
e-mail: naoko@uta.edu

J. Witzel
Department of Linguistics & TESOL, University of Texas at Arlington, Arlington, TX, USA

Introduction

Over the years, a number of experimental tasks have been used to investigate how sentences are comprehended. Representative methods in auditory sentence processing research include click detection (see e.g., Garrett et al. 1965), phoneme monitoring (see e.g., Foss 1969), cross-modal priming (see e.g., Nicol and Swinney 1989), self-paced listening (see e.g., Waters and Caplan 2004), and eye tracking under the visual world paradigm (see e.g., Kamide et al. 2003). Sentence processing is most commonly investigated, however, with tasks that involve the visual presentation of sentential stimuli and in which response time or reading time (both of which will be referred to as RT in this paper) is a crucial dependent variable. These tasks include speeded grammaticality judgment (see e.g., Clifton et al. 1984; Nicol et al. 1997), same-different matching (see e.g., Freedman and Forster 1985), (word-by-word) sensicality judgments (see e.g., Boland et al. 1990; Tanenhaus et al. 1989), rapid serial visual presentation (see e.g., Forster 1970; not to be confused with the "not-so-rapid" serial visual presentation of words/phrases that is often used in neuroimaging studies; see e.g., Phillips et al. 2005; inter alia), probe recognition (see e.g., Bever and McElree 1988; McElree and Bever 1989), maze task reading (see e.g., Forster 2010; Forster et al. 2009; Nicol et al. 1997), self-paced reading (see e.g., Just et al. 1982), and eye tracking during reading (for review, see Rayner 2009). This is not to say however that these reading tasks are (even roughly) equally represented in the literature. Rather, eye tracking and self-paced reading (henceforth, SPR), and more specifically, non-cumulative, moving-window SPR (Just et al. 1982; for an example of a non-moving window versions of this task, see Gordon et al. 2004), have emerged as the most widely-accepted experimental tasks for the investigation of sentence comprehension during reading. And perhaps for good reason. Because these tasks conceivably allow for indications of processing ease/difficulty as readers make their way through sentences, they have the potential to shed light on the processes that underlie the incremental integration of words and phrases into developing sentence representations. However, just as it is incumbent on sentence processing researchers to scrutinize and develop theoretical claims, it is also necessary to probe these accepted methodologies for weaknesses as well as to experiment with alternative techniques. The present study addresses these methodological concerns by comparing the findings from SPR, eye tracking, and two versions of the maze task (described below) on sentences involving temporary structural ambiguities. Of particular interest are (i) whether these tasks are capable of revealing processing differences among these sentence types and (ii) whether these effects are indicated precisely at the predicted word/region.

Experimental tasks that involve reading can (and have been) evaluated in a number of ways. Chief among these considerations are (i) whether the task allows for "online" indications of processing ease/difficulty and (ii) the extent to which the task is "natural" (Mitchell 2004). The first of these metrics is of obvious importance. Sentence processing researchers are primarily concerned with revealing the characteristics of sentence comprehension operations, and most assume that reading patterns, and more specifically, indications of processing time differences, on the component parts of sentences can provide important information about these characteristics. As to the latter of these criteria, the assumption is that the more natural the task, the better—a belief that is rooted in the idea that a natural task does not distort the normal operations involved in comprehending sentences (Mitchell 2004).

When evaluated against these criteria, eye tracking appears to fare quite well. In a typical eye-tracking experiment, entire sentences are presented one at a time, and participants are asked to read each one (usually silently) at their normal reading speed. While participants are reading, the location and duration of their eye fixations are recorded, which then allows for the calculation of a number of dependent measures for each word or set of words in the

sentences of interest. Comprehension questions follow all, some, or none of the sentences, depending on the lab or researcher conducting the experiment. In this way, eye tracking makes it possible to assess participants' reading patterns as they make their way through sentences (i.e., it allows for measures of "online" processing) and places very few restrictions on how participants accomplish this task (i.e., it allows for "natural" reading). With regard to the naturalness of this task, its crucial feature is that participants are able to stop their progression through the sentence at any point to reinspect previous content. In fact, the frequency with which such *regressive eye movements* are initiated from (or land on) a given word or region of the sentence is one of the dependent measures analyzed in eye-tracking studies. Because eye tracking allows for natural indications of online sentence processing, it has been used to examine a virtually exhaustive range of issues in sentence comprehension (for review, see Rayner 2009; Rayner and Pollatsek 2006; Rayner and Sereno 1994; Staub and Rayner 2007).

It is important to point out, however, that there is in principle a trade-off between the "naturalness" of eye tracking and its ability to indicate characteristics of the online integration of words/phrases into developing sentence representations. Because eye tracking places few restrictions on the way participants approach reading (other than often requiring them to be able to answer simple comprehension questions), the task allows for a number of reading strategies on any given item in the experiment. Participants might read the sentence very carefully, or might skim through the sentence in order to arrive at an approximate interpretation, or might adopt a reading strategy somewhere between these poles. The strategy adopted influences the extent to which fixation durations and locations accurately indicate online processing differences.

Evaluated against the same criteria of naturalness and sensitivity to online integration processes, SPR fares less well. In a typical SPR experiment, each sentence is presented as a series of dummy characters (usually as a series of dashes), with each dash representing a character in the words making up the sentence. When the participant presses a specified button, the first word of the sentence is displayed. When the participant is ready to view the next word of the sentence, s/he presses the same button. The first word then reverts to dummy characters, and the next word is displayed. The participant proceeds in this manner until all of the words in the sentence have been shown. Depending on the lab/researcher(s), some or all of the sentences are followed by comprehension questions. In terms of naturalness, SPR is a bit of a mixed bag. To the extent that it does not require explicit decisions about the properties of the text as it is presented, this task approximates normal reading. However, it is very different from normal reading in that it displays only one word at a time (and, as such, forces explicit fixation on oft-skipped function words) and in that it does not permit regressive eye movements. (There is also a version of this task in which sentences are presented phrase-by-phrase, rather than word-by-word (see e.g., Van Dyke and McElree 2006). This variant is used less often, so this report will focus on SPR with word-by-word presentation.) In theory, these unnatural aspects of the task might provide for clearer indications of the characteristics of online processing differences. First, the incremental presentation of each word in the sentence makes it possible to examine reading patterns with reference to a single dependent measure—specifically, the time it takes to push the button in recognition of each word. Also, and possibly more importantly, this manner of presentation reduces the number of reading strategies available to subjects and, thus, might help to standardize their approach to each item in the experiment.

Although the method of presentation in SPR might lead to a more consistent approach to reading across subjects (and across items for each subject) than in eye tracking, a variety of strategies are nevertheless available for this task, with the strategy selected potentially influencing how well button-pressing time can be taken to reflect characteristics of online sentence processing. Participants might adopt a strategy whereby they press the button to

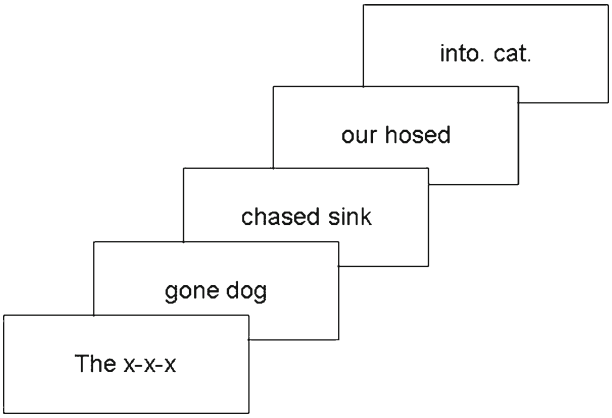

Fig. 1 Sequence of frames in a typical G-maze task. (Adapted from Forster et al. 2009)

move on to the next word as soon as they have recognized the word and integrated it into the developing sentence representation. This strategy would likely yield the most accurate indications of the processing ease/difficulty associated with parsing each word in the sentence. But a given subject (or set of subjects) could easily adopt a more or less conservative button-pushing criterion. A subject might, for instance, delay button-pushing until well after the integration of each word to ensure that the sentence had been encoded into memory for recall when/if a comprehension question occurred. If this strategy were adopted, effects that might otherwise have been obtained could be lost in longer, more variable "reading" times. Worse yet, this strategy could lead to results that overestimate the influence of memory processes during online sentence comprehension. On the other end of the spectrum, a subject might simply press the button as quickly as possible, buffering each word for reconstruction later in the sentence. The use of this buffering strategy might explain why SPR experiments often only yield results on words following the region of interest in the sentence. Such "spillover" effects—or, more appropriately, "holdover" effects—are problematic because they clearly indicate that RTs in SPR experiments often do not reflect the processing associated with the words on which they are recorded.

A more recent word-by-word reading technique—the maze task—places stricter limits on the strategies available to subjects. There are two versions of this task—(i) the grammaticality maze (or G-maze) task and (ii) the lexicality maze (or L-maze). In both versions, a sentence is presented as a sequence of choices between paired alternatives, only one of which continues the sentence. The participant's task is to choose the alternative that continues the sentence as quickly and as accurately as possible. In the G-maze task, both alternatives are words, but only one is a grammatical continuation of the sentence. The L-maze is a somewhat easier version of the task. In this variant, the choice is between a word and a legal nonword, and the participant must choose which of the letter strings is a word. In either case, if the subject makes the correct choice from each pair in the sequence, the selected words form a sentence. These variants of the maze task are illustrated in Figs. 1 and 2.[1]

Evaluated in terms of naturalness, both versions of the maze task fare horribly. There is nothing natural about having to make explicit choices about the grammaticality or lexicality of the component words in sentences. However, what the maze task lacks in naturalness,

[1] Alternatively, a demonstration of G-maze can be found at the following website www.u.arizona.edu/~kforster/MAZE.

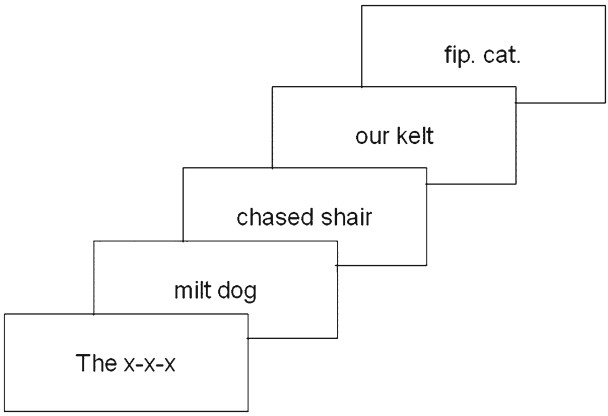

Fig. 2 Sequence of frames in a typical L-maze task

it might make up for in its sensitivity to the processing costs associated with integrating words into sentences. Indeed, this task might be more sensitive precisely because it places unnatural restrictions on the strategies available to subjects. The G-maze task is the more demanding version, both in terms of the difficulty of the decision required and in terms of the constraints it places on the strategies available to participants. This variant essentially forces the participant to adopt an approach to reading whereby a button is pressed for each word in the sentence as soon as it has been recognized and integrated into the developing sentence representation. In this way, the G-maze has the potential to provide highly localized indications of processing time differences during online sentence comprehension. Specifically, it should indicate processing time differences at precisely the words predicted to yield such disparities. And although processing time differences incurred at a given point in a sentence could influence decisions on subsequent words (i.e., lead to "spillover" effects), it is highly unlikely that they would show up only on subsequent words (i.e., show up as "holdover" effects). The L-maze task is slightly more permissive. While it necessitates word recognition at each choice point, it does not strictly require incremental integration into the developing sentence structure. And indeed, this version of the task could be accomplished on purely lexical grounds, without any reference to sentence structure or meaning (but see Forster et al. 2009, for results indicating that this is not what happens in practice). Despite this, the idea is that by necessitating lexical-level processing at each word in the sentence—a minimum level of processing that is not required in either eye tracking or SPR—the L-maze increases the likelihood that higher-level, structurally/semantically-relevant effects will be revealed at each word. If this assumption is correct, the L-maze should also yield highly localized indications of processing time differences during online sentence comprehension.

The present study sought to compare the results from eye tracking, SPR, the G-maze, and the L-maze on a common set of sentence types involving temporary structural ambiguities. The processing of such sentences has been examined extensively in order to shed light on the architecture of the sentence comprehension system. In light of the differences among these tasks, of particular interest were (i) whether each task would be capable of revealing the predicted pattern of processing time differences on each sentence type and (ii), as (if not more) importantly, whether these differences would be indicated precisely at the predicted word/region.

Sentences of Interest

Of interest in this study are three sentence types involving temporarily ambiguous structural configurations—sentences involving relative clause (RC) attachment ambiguity, adverb attachment ambiguity, and noun phrase (NP) versus sentence (S) coordination ambiguity (all of which were drawn from Witzel et al. in press). There were two main reasons for examining these sentence types. First, as mentioned above, sentences containing temporary structural ambiguities in general, and these ambiguities in particular, have been investigated extensively in the sentence processing literature. Therefore, the predictions about which sentences should cause processing difficulty and where that processing difficulty should arise are relatively straightforward and uncontroversial (even if the reasons for these difficulties remain matters of theoretical debate). Secondly, these sentences allow for a novel test of the maze methodology. To date, the maze task has been used to investigate movement violations (Freedman and Forster 1985), subject-verb agreement processing (Nicol et al. 1997), the processing of subject- and object-extracted relative clauses (in English: Forster et al. 2009; in Chinese: Qiao and Forster 2008), the processing of control structures and scrambling in Japanese (Witzel and Witzel 2009), word frequency effects (Forster et al. 2009), and lexical ambiguity resolution (Forster et al. 2009; Witzel and Forster 2009). Although a complete review of these studies is beyond the scope of this article, let it suffice to say that these studies have yielded results in line with those obtained in eye-tracking and SPR examinations of similar phenomena. Structurally ambiguous sentences, however, have not been tested with either version of the maze task, and thus these sentences might offer new insights into the benefits and drawbacks of this technique.

The first sentence type of interest involved RC attachment ambiguity, as in examples (1a) and (1b) below. In these sentences, the RC *who shot herself/himself* is in a structural position where it can modify either of the component NPs in the complex NP *The son of the actress*. Modification of the local NP (*the actress*) is often referred to as low attachment, whereas modification of the nonlocal NP (*the son*) is referred to as high attachment. In this case, the RC attachment site is clearly indicated by antecedent-reflexive gender agreement.

(1a) The son of the actress who shot *herself* on the set was under investigation. (*Low Attachment*)
(1b) The son of the actress who shot *himself* on the set was under investigation. (*High Attachment*)

In English, there appears to be a somewhat inconsistent bias toward low RC attachment, as indicated by longer reading times for disambiguating information necessitating high attachment than for information requiring low attachment. That is, some studies have produced results consistent with this bias (Carreiras and Clifton 1999; Cuetos and Mitchell 1988; Frazier and Clifton 1997), while others have indicated no bias at all (Carreiras and Clifton 1993; Traxler et al. 1998). Of particular importance to the present investigation, these divergent findings have been attributed at least in part to methodological differences. Carreiras and Clifton, for instance, found evidence for a low attachment bias in English with eye tracking (Carreiras and Clifton 1999), but not with SPR (Carreiras and Clifton 1993)—a disparity that they partially explain with reference to differences in the sensitivity of these tasks to online parsing.

The second set of sentences involved adverb attachment ambiguity, as in examples (2a) and (2b) below. This ambiguity is perhaps best illustrated by the globally ambiguous sentence *Jack called the friend he met last week.* In this sentence, it is unclear which of the two events occurred last week, the calling of the friend or the meeting of the friend. More technically, it is ambiguous as to whether the adverb phrase *last week* modifies the nonlocal

verb phrase (VP) *called the friend* (high attachment) or the local VP *met* (low attachment). In the sentences examined in the present study, the adverb attachment site was disambiguated by the respective tenses of the local and nonlocal VPs.

(2a) Susan bought the wine she will drink *next week,* but she didn't buy any cheese. (*Low Attachment*)

(2b) Susan bought the wine she will drink *last week,* but she didn't buy any cheese. (*High Attachment*)

Although this sentence type has not been investigated as extensively as RC attachment sentences, one eye-tracking study in particular showed that English native speakers have a strong preference for low adverb attachment (Altmann et al. 1998)—a bias indicated by much longer reading times for the adverbs that attached to the nonlocal VP (high attachment) than for those that attached to the local VP (low attachment).

The third set of sentences involved NP versus S coordination ambiguity, as in example sentences (3a) and (3b). Example (3b) is temporarily ambiguous in that it is unclear whether *and the salesman* is part of a conjoined direct object (as in *The robber shot the jeweler and the salesman.*) or the subject of a conjoined sentence (as it turns out to be in this example). The structural status of this string is ambiguous until the verb in the conjoined sentence (*reported*) becomes available. In example (3a), this ambiguity is prevented by the comma, which clearly indicates that *and the salesman* starts a new sentence.

(3a) The robber shot the jeweler, and the salesman *reported* the crime to the police. (*Unambiguous*)

(3b) The robber shot the jeweler and the salesman *reported* the crime to the police. (*Ambiguous*)

In an SPR experiment, Frazier and Clifton (1997) showed that English readers preferred NP coordination over S coordination, as evidenced by inflated reading times for verbs that necessitated the latter analysis. Similarly, in Dutch, a language typologically similar to English, both SPR (in its phrase-by-phrase incarnation; Frazier 1987) and eye-tracking (Hoeks et al. 2006) experiments have yielded results consistent with an NP-coordination bias.

The purpose of the present study is to test these sentence types with four different reading paradigms—eye tracking, SPR, the G-maze, and the L-maze—in order to determine whether each task is capable of providing indications of processing difficulty consistent with those obtained in previous studies. Also of interest is the extent to which each task is able to "localize" this processing difficulty to precisely those regions/words that are predicted to cause interpretive problems. As indicated above, the regions/words that should cause processing difficulty are those that conflict with online biases in the interpretation of these structural ambiguities (and that thus require reanalysis of the sentence structure).

Method

Each of the reading tasks was run as a separate experiment. However, in order to make it easier to compare among the tasks (and because there was considerable overlap in their design), we will present the methodological details of each in a single "Method" section.

Participants

Thirty-two undergraduate students enrolled in an introductory psychology course at the University of Arizona participated in each of the four experiments, for which they received course

credit. None of the subjects participated in more than one of the experiments.[2] All of the participants were native speakers of English.

Materials and Design

The sentences tested in all four reading tasks were the same as those used in Witzel et al. (in press). There were 24 items for each of the three experimental sentence types. Each item had two versions, which were counterbalanced over two lists. In addition to these experimental items, 24 fillers were included in each reading task. These fillers included sentences with RCs and time adverbials in structurally unambiguous positions as well as various coordination types, and were included in order to prevent participants from realizing the purpose of the experiments. In total, there were 96 sentences in each reading task.

In both of the maze task experiments, each word in the sentences (except for the first word) was paired with an ungrammatical word (for the G-maze) or a nonword (for the L-maze). For the G-maze, the ungrammatical alternatives were carefully selected (from a random list of words) to ensure that none would make a grammatical continuation at the point in the sentence at which it occurred. For the L-maze, nonword alternatives were drawn from the ARC Nonword Database (Rastle et al. 2002). The pairing of the correct alternative with the incorrect alternative was the same in both conditions for each item in the three experimental sentence types. These correct and incorrect alternatives appeared randomly on the left or the right.

Another difference in the materials among the four reading tasks related to comprehension questions. For the eye-tracking and SPR experiments, half of the 96 sentences were followed by yes-no comprehension questions. These comprehension questions followed the same items in both experiments. Only those subjects who scored at least 80% correct on these questions were included in the analysis for these two experiments. The sentences in the G-maze and L-maze experiments were not followed by comprehension questions, and there was no criterion for exclusion.

Procedure

Eye Tracking

Sentences were presented as single lines of text (with standard capitalization and punctuation) in white letters on a black background on a 21-inch CRT monitor. Participants were asked to read each sentence at their natural reading speed, making sure to comprehend well enough to accurately answer occasional yes-no questions. Participants' eye movements were recorded from the right eye using a Dr. Bouis Monocular Oculometer, at a sampling rate of 200 Hz. The distance from the eye to the monitor was approximately 60 cm, allowing for single character resolution. A bite plate and headrest were used to attenuate head movements. The eye tracker was calibrated at the beginning of the experimental session and then recalibrated after every four trials. Each trial began with a fixation mark (an asterisk) close to the left margin of the computer screen. A sentence would then display, with its first letter located one character space to the left of the fixation point. After reading the sentence, the participant pressed a button under the right hand, at which point the sentence was removed from the screen. If the item was not followed by a comprehension question, a string of dashes appeared on the screen, signaling that the participant could proceed to the next trial when ready by again

[2] Note that part of the eye-tracking data (specifically, the first 30 subjects) are reported as the Native Speaker results in Witzel et al. (in press). In order to equalize the number of participants in all four tasks, 2 more participants were added to this dataset.

pressing the right button. If the item was followed by a comprehension question, participants answered 'yes' with the right button or 'no' with a button under the left hand. The participant then received feedback, and the next trial began automatically. At the beginning of the reading task, participants were given eight practice trials.

SPR

The SPR experiment was controlled by a Pentium PC, using Windows-based DMDX software (Forster and Forster 2003). Items were presented as black letters on a white background, using a color monitor with a refresh cycle of 10 ms. Each sentence initially appeared as a series of dashes, with each dash corresponding to a letter of a word in the sentence. The participant pressed the right button on a button box to see the first word of the sentence. The participant read this word and again pressed the right button to continue on to the next word. When the next word appeared, the first word reverted to dashes, with the interval between each button press recorded by the computer. The participant continued in this manner until the end of the sentence. After the button press for the last word in the sentence, the item was removed from the screen. At that point, the subject either received a comprehension question or the next item (again, as a string of dashes). If the item was followed by a question, the participant answered 'yes' with the right button on the button box or 'no' with the left button. The participant then received feedback, and the next item began automatically. After every 12 items, the subject was encouraged to take a short rest. At the beginning of the reading task, the participant was given eight practice trials.

G-maze

Both the G-maze and the L-maze (described below) were conducted using the same hardware, software, and display settings as in the SPR experiment. In these maze reading tasks, each sentence was presented as a series of frames, the first of which consisted of [The X-X-X]. In the G-maze task, each subsequent frame contained two words side by side, only one of which was a grammatical continuation of the sentence. Participants were instructed to choose the word that best continued the sentence as quickly and as accurately as possible by pushing the corresponding left or right button on a button box. When the correct alternative was selected, the next pair of alternatives was automatically displayed. When the incorrect alternative was chosen, an "⟨error⟩" message was presented, followed by the beginning of the next item. If the participant made the correct selections throughout the frames, the final selection was followed by a "CORRECT" message, and then the beginning of the next item. Unlike in the eye-tracking and SPR experiments, there were no comprehension questions. After every 12 items, the subject was encouraged to take a short rest. At the beginning of the task, the participant was given eight practice trials.

L-maze

The L-maze experiment was conducted in much the same way as its G-maze counterpart. The only difference between these tasks was that in the L-maze, the frames that made up each item consisted of a word and a nonword. Participants were instructed to choose the word in each frame as quickly and as accurately as possible. They were also told that the sequence of words in each set of frames would form a sentence. Note that this procedure is slightly different from that which was used in Forster et al. (2009). Their L-maze task included both

grammatical and ungrammatical sentences (i.e., scrambled sequences of words), and participants were told that sometimes the words in each set would form a sentence, but other times they would not. Ungrammatical sentences were not included in the present experiment, however, in order to allow for a more straightforward comparison of this L-maze task with the other reading paradigms.

Data Analysis

Of particular interest were RTs at the words that provided disambiguating information about the structural properties of the temporarily ambiguous sentences as well as the words/phrases that followed disambiguation. The specific words/regions of interest depended on the sentence structure and are indicated below in the results section for each sentence type. For the eye-tracking data, five measures were calculated—First Fixation Duration, First-Pass RT, Go-Past RT, Right-Bounded RT, and Total RT. First Fixation Duration was defined as the duration of the first fixation in a region, provided that it lasted for at least 50 ms. First-Pass RT was defined as the sum of the fixation durations in a region before leaving that region in any direction (In the eye-tracking literature, this measure is often referred to as Gaze when calculated for single-word regions. However, for the sake of consistency, the term First-Pass RT will be used regardless of the number of words in the region under analysis.). Go-Past RT (also known as Regression Path RT) was defined as the sum of the fixation durations after entering a region before leaving that region to the right. This measure included regressive fixations outside of the region. Right-Bounded RT was defined as the sum of the fixation durations in a region before moving out of that region to the right. This measure did not include regressive fixations outside of the region. Total RT was simply the sum of all of the fixation durations in a region. For regions involving more than one word, both "raw" and "per word" RTs were calculated for the First-Pass RT, Go-Past RT, and Right-Bounded RT measures. The "raw" RTs were not adjusted for the number of words in the region, while the "per word" RTs were calculated by dividing the measure by the number of words in the region. Although "per word" RTs are not commonly reported in more recent eye-tracking studies, they were deemed important for the present study in order to allow for clearer comparisons with the "per word" RTs reported for SPR and for the maze tasks. For these latter tasks, RTs for single words were calculated from the time the word appeared on the screen until the button press. For multiple-word regions, the average button-pressing time was averaged across the relevant words. This averaging process was done prior to the application of the outlier rejection/correction procedures discussed below.

Results

In the eye-tracking and SPR tasks, no participant scored less than 80% correct on the comprehension questions. The mean comprehension accuracy score for participants in the eye-tracking experiment was 90.53% (SD = 4.50); in the SPR task, the mean comprehension accuracy score was 89.25% (SD = 3.96). For both of these tasks, the data from all of the experimental sentences were included in the analyses, subject to the trimming procedures described below. For the two maze tasks, if the wrong alternative was chosen on a given frame, the RT for that frame was discarded. Also, note that the RTs for the remainder of that sentence would not be recorded since the experiment was set up to continue automatically to the next sentence. The data trimming procedures were as follows: For the eye-tracking data, trials with major tracker losses were excluded from the analysis. These trials accounted for 3.34% of the experimental

items. For the SPR, G-maze, and L-maze data, outliers were treated by setting them equal to cutoffs established 2 SD units above and below the mean for each participant. For each comparison, two analyses of variance (ANOVAs) were conducted—one with subjects as the random variable ($F1$), and the other with items as the random variable ($F2$). List/Item Group was included as a non-repeated factor in these analyses in order to remove the variability associated with the counterbalancing procedures.

Relative Clause Attachment

Eye Tracking

Table 1 presents the mean First Fixation Duration, First-Pass RT, Go-Past RT, Right-Bounded RT, and Total RT by condition and disambiguating region. None of these measures yielded differences at the disambiguating reflexive (First Fixation, First-Pass, and Total: all F's < 1; Go-Past: $F1(1, 30) = 2.06, F2(1, 22) = 2.15$; Right-Bounded: $F1 < 1.5, F2(1, 22) = 1.51$). In the region following disambiguation (reflexive + 1), however, several measures revealed trends suggesting that low attachment sentences were read more quickly than high attachment sentences. First Fixation Durations were shorter for low attachment sentences in the by-subjects analysis, $F1(1, 30) = 4.85, p < .05$, but this difference only approached significance by items, $F2(1, 22) = 3.12, p = .09$. In First-Pass RT, this difference was significant by items, $F2(1, 22) = 4.53, p < .05$, and approached significance by subjects, $F2(1, 30) = 3.43, p = .07$—an effect that was not strengthened by the per word

Table 1 Mean reading times for relative clause attachment sentences in eye tracking

		Region		
		Reflexive	Reflexive + 1	Reflexive + 2
	Low attachment	herself	on the set	was under investigation
	High attachment	himself	on the set	was under investigation
First Fixation	Low attachment	243	250	257
	High attachment	245	263	261
First-Pass RT	Low attachment	260	464	653
	High attachment	267	490	641
First-Pass RT (per word)	Low attachment		179	236
	High attachment		188	228
Go-Past RT	Low attachment	314	584	585
	High attachment	339	636	590
Go-Past RT (per word)	Low attachment		227	212
	High attachment		250	211
Right-Bounded RT	Low attachment	277	514	567
	High attachment	289	550	573
Right-Bounded RT (per word)	Low attachment		198	205
	High attachment		212	205
Total RT	Low attachment	399	692	800
	High attachment	414	709	787

calculation of this measure, $F1(1, 30) = 2.99, p = .09, F2(1, 22) = 3.55, p = .07$. In Right-Bounded RT, this difference was significant by subjects, $F1(1, 30) = 5.08, p < .05$, and marginally significant by items, $F2(1, 22) = 4.04, p = .05$. The per word calculation of this measure, however, revealed significant differences under both analyses, $F1(1, 30) = 5.69, p < .05, F2(1, 22) = 4.33, p < .05$. Neither Go-Past RT nor Total RT showed reliable differences in this region (Go-Past, raw: $F1(1, 30) = 2.54, F2(1, 22) = 2.36$; Go-Past, per word: $F1(1, 30) = 3.01, F2(1, 22) = 2.79$; Total: both F's < 1). There were no differences between low and high attachment sentences in the final region (reflexive + 2: all F's < 1). A follow-up analysis was also conducted on the Total RT for the relative clause up to and including the disambiguating reflexive (i.e., *who shot himself/herself*). This was done in order to determine the cumulative time that readers spent inspecting this structurally ambiguous constituent under low and high attachment conditions. Consistent with the pattern of results for the region immediately following disambiguation, the Total RTs in this relative clause region were shorter for low attachment sentences than for high attachment sentences (low attachment = 1091 ms, high attachment = 1175 ms; $F1(1, 30) = 5.28, p < .05, F2(1, 22) = 5.47, p < .05$).

SPR

Table 2 presents the mean SPR RTs for low and high attachment sentences at and after the disambiguating reflexive. The RTs for low and high attachment sentences were not significantly different at the disambiguating reflexive, $F1(1, 30) = 1.13, F2(1, 22) = 1.07$, or at the immediately following word (reflexive + 1: $F1(1, 30) = 1.77, F2(1, 22) = 2.55$). However, the average RTs for the words in the phrase following the disambiguating reflexive were longer in high attachment sentences than in low attachment sentences (combined

Table 2 Mean reading times for relative clause attachment sentences in SPR, G-maze, and L-maze

		Region			
		Reflexive	Reflexive + 1	Reflexive + 2	Reflexive + 3
	Low attachment	herself	on	the set	was under investigation
	High attachment	himself	on	the set	was under investigation
SPR	Low attachment	487	451		426
	High attachment	513	478		415
	Low attachment			412	
	High attachment			443	
G-maze	Low attachment	860	1026		1001
	High attachment	986	1074		966
	Low attachment			975	
	High attachment			1005	
L-maze	Low attachment	663	650		715
	High attachment	702	674		720
	Low attachment			687	
	High attachment			718	

reflexive + 1/reflexive + 2: $F1(1, 30) = 5.75, p < .05, F2(1, 22) = 5.76, p < .05$). This effect was not held-over into the final words of the sentences (reflexive + 3: both F's < 1).

G-maze

Table 2 presents the mean G-maze RTs for low and high attachment sentences at and after the disambiguating reflexive. The results indicated that this reflexive was responded to more quickly in low attachment sentences than in high attachment sentences, $F1(1, 30) = 18.49, p < .001, F2(1, 22) = 19.61, p < .001$. Although the RT patterns at the word following disambiguation and at the phrase following disambiguation were consistent with the spillover of this effect, the differences in these regions were not statistically reliable (reflexive + 1: $F1(1, 30) = 2.08, F2(1, 22) = 1.35$; combined reflexive + 1/reflexive + 2: $F1(1, 30) = 2.28, F2(1, 22) = 2.44$). There was no indication that this effect spilled over into the final words of the sentences (reflexive + 3: $F1(1, 30) = 2.58$, F2 < 1).

L-maze

Table 2 presents the mean L-maze RTs for low and high attachment sentences at and after the disambiguating reflexive. As in the G-maze task, this reflexive was responded to more quickly in low attachment sentences than in high attachment sentences, $F1(1, 30) = 7.46, p < .05, F2(1, 22) = 5.58, p < .05$. There was also a trend suggesting that this effect spilled over onto the processing of subsequent words. Specifically, although the word following disambiguation did not reveal a significant RT difference between the sentence types (reflexive + 1: $F1(1, 30) = 2.55, F2(1, 22) = 1.40$), there was a trend suggesting that the phrase following disambiguation was read more quickly in low attachment sentences (combined reflexive + 1/reflexive + 2: $F1(1, 30) = 8.63, p < .01, F2(1, 22) = 3.51, p = .08$). There was no difference in the RTs over the final words in these sentences (reflexive + 3: both F's < 1).

Taken together, the results of all four experiments were consistent with a low attachment bias in the processing of English relative clause structures. It is important to note, however, that there were differences among the tasks in terms of the timing and strength of the garden-path effect when this bias was violated. The eye-tracking experiment revealed a somewhat weak garden-path effect (in that it was not statistically-reliable across measures) that was delayed until the region immediately following the disambiguating reflexive. This effect also appeared in an analysis of the cumulative reading time on the beginning portion of the temporarily ambiguous relative clause. Comparably, the SPR experiment did not reveal a garden-path effect on the disambiguating reflexive or on the word immediately following disambiguation. Rather, this effect was obtained only when averaging across the RTs to the words in the phrase following disambiguation. In contrast, both the L-Maze and the G-Maze revealed a garden-path effect precisely at the critical disambiguating word.

Adverb Attachment

Eye Tracking

Table 3 presents the mean First Fixation Duration, First-Pass RT, Go-Past RT, Right-Bounded RT, and Total RT by condition and disambiguating region. Low attachment sentences were read more quickly than high attachment sentences at and immediately after the adverb. In these regions, low attachment sentences had significantly faster First-Pass RTs (adverb, raw: $F1(1, 30) = 8.31, p < .01, F2(1, 22) = 11.09, p < .005$;

Table 3 Mean reading times for adverb attachment sentences in eye tracking

		Region		
		Adverb	Adverb + 1	Adverb + 2
	Low attachment	next week,	but she didn't buy	any cheese
	High attachment	last week,	but she didn't buy	any cheese
First Fixation	Low attachment	238	240	252
	High attachment	247	242	264
First-Pass RT	Low attachment	301	526	485
	High attachment	337	577	442
First-Pass RT (per word)	Low attachment	191	149	193
	High attachment	214	165	179
Go-Past RT	Low attachment	358	569	469
	High attachment	525	751	441
Go-Past RT (per word)	Low attachment	235	162	187
	High attachment	335	214	180
Right-Bounded RT	Low attachment	317	542	455
	High attachment	403	648	424
Right-Bounded RT (per word)	Low attachment	202	154	182
	High attachment	254	185	173
Total RT	Low attachment	356	680	556
	High attachment	487	797	513

adverb, per word: $F1(1, 30) = 6.50, p < .05, F2(1, 22) = 8.57, p < .01$; adverb + 1, raw: $F1(1, 30) = 3.65, p = .06, F2(1, 22) = 13.73, p < .005$; adverb + 1, per word: $F1(1, 30) = 4.15, p < .05, F2(1, 22) = 13.17, p < .005$), Go-Past RTs (adverb, raw: $F1(1, 30) = 28.16, p < .001, F2(1, 22) = 28.60, p < .001$; adverb, per word: $F1(1, 30) = 24.66, p < .001, F2(1, 22) = 24.74, p < .001$; adverb + 1, raw: $F1(1, 30) = 25.01, p < .001, F2(1, 22) = 51.42, p < .001$; adverb + 1, per word: $F1(1, 30) = 24.50, p < .001, F2(1, 22) = 50.66, p < .001$), Right-Bounded RTs (adverb, raw: $F1(1, 30) = 27.91, p < .001, F2(1, 22) = 31.68, p < .001$; adverb, per word: $F1(1, 30) = 23.00, p < .001, F2(1, 22) = 25.41, p < .001$; adverb + 1, raw: $F1(1, 30) = 19.63, p < .001, F2(1, 22) = 62.12, p < .001$; adverb + 1, per word: $F1(1, 30) = 20.48, p < .001, F2(1, 22) = 58.11, p < .001$), and Total RTs (adverb: $F1(1, 30) = 38.26, p < .001, F2(1, 22) = 45.56, p < .001$; adverb + 1: $F1(1, 30) = 10.63, p < .005, F2(1, 22) = 23.31, p < .001$). First Fixation Durations in these regions, however, did not differ between the low and high attachment sentences (adverb: $F1 = 1.64, F2 = 2.43$, adverb + 1: both F's < 1). In the final region of the sentence (adverb + 2), several measures revealed trends suggesting that high attachment sentences were read more quickly than low attachment sentences (Go-Past, raw: $F1(1, 30) = 1.85, p = .18, F2(1, 22) = 4.85, p < .05$; Go-Past, per word: $F1 < 1, F2 = 2.46$; Right-Bounded, raw: $F1(1, 30) = 2.27, p = .14, F2(1, 22) = 6.61, p < .05$; Right-Bounded, per word: $F1(1, 30) = 1.11, p = .30, F2(1, 22) = 3.86, p = .06$; Total: $F1(1, 30) = 3.43, p = .07, F2(1, 22) = 5.70, p < .05$), but this difference was only significant in First-Pass RT (first-pass, raw: $F1(1, 30) = 4.29, p < .05, F2(1, 22) = 5.21, p < .05$; First-Pass, per word: $F1(1, 30) = 2.63, p = .11, F2(1, 22) = 4.60, p < .05$).

Table 4 Mean reading times for adverb attachment sentences in SPR, G-maze, and L-maze

		Region			
		Adverb	Adverb + 1	Adverb + 2	Adverb + 3
	Low attachment	next week,	but	she didn't buy	any cheese
	High attachment	last week,	but	she didn't buy	any cheese
SPR	Low attachment	409	349		370
	High attachment	541	387		379
	Low attachment			324	
	High attachment			349	
G-maze	Low attachment	868	840		838
	High attachment	1097	923		883
	Low attachment			877	
	High attachment			928	
L-maze	Low attachment	638	602		668
	High attachment	688	615		674
	Low attachment			631	
	High attachment			638	

SPR

Table 4 presents the mean SPR RTs for low and high attachment sentences at and after the adverb. As in the eye-tracking experiment, RTs were faster at the adverb in low attachment sentences than in high attachment sentences, $F1(1, 30) = 8.15, p < .01, F2(1, 22) = 15.14, p < .001$. This effect spilled over onto the immediately following word (adverb + 1: $F1(1, 30) = 9.91, p < .005, F2(1, 22) = 8.02, p < .01$) and phrase (combined adverb + 1/adverb + 2: $F1(1, 30) = 9.36, p < .005, F2(1, 22) = 8.05, p < .01$), but did not influence response times at the end of the sentence (adverb + 3: $F1 = 1.46, F2 = 2.77$).

G-maze

Table 4 presents the mean G-maze RTs for low and high attachment sentences at and after the adverb. Again, RTs were faster at the adverb in low attachment sentences than in high attachment sentences, $F1(1, 30) = 82.66, p < .001, F2(1, 22) = 46.57, p < .001$. This effect spilled over onto the immediately following word (adverb + 1: $F1(1, 30) = 18.14, p < .001, F2(1, 22) = 10.86, p < .005$) and phrase (combined adverb + 1/adverb + 2: $F1(1, 30) = 8.36, p < .01, F2(1, 22) = 9.52, p < .01$). Low attachment sentences also yielded faster response times in final region of the sentence (adverb + 3: $F1(1, 30) = 10.21, p < .005, F2(1, 22) = 5.56, p < .05$).

L-maze

Table 4 presents the mean L-maze RTs for low and high attachment sentences at and after the adverb. As in all of the other tasks, RTs were faster at the adverb in low attachment sentences than in high attachment sentences, $F1(1, 30) = 11.82, p < .005, F2(1, 22) = 15.84, p < .001$.

Although there was a weak trend suggesting that this effect spilled over onto the immediately following word (adverb + 1: $F1(1, 30) = 2.91$, $p = .10$, $F2 < 1.5$), there was no indication of this spillover in the following phrase (combined adverb + 1/adverb + 2: both F's < 1) or in the final region of the sentences (adverb + 3: both F's < 1).

As in the RC attachment sentences, all four tasks revealed a low attachment bias for adverb attachment. However, in contrast to the results for the RC attachment sentences, there were no major differences in the strength and timing of this effect across the tasks.

NP versus S Coordination

Eye tracking

Table 5 presents the mean First Fixation Duration, First-Pass RT, Go-Past RT, Right-Bounded RT, and Total RT by condition and disambiguating region. Sentences in which the verb in the second clause indicated an S-coordination structure were read more slowly at and immediately after this verb than sentences in which this structure was established unambiguously by a comma. In these regions, temporarily ambiguous sentences had slower Go-Past RTs (verb: $F1(1, 30) = 14.87$, $p < .001$, $F2(1, 22) = 8.09$, $p < .01$; verb + 1, raw: $F1(1, 30) = 10.79$, $p < .005$, $F2(1, 22) = 13.73$, $p < .005$; verb + 1, per word: $F1(1, 30) = 9.12$, $p < .01$, $F2(1, 22) = 10.05$, $p < .005$; combined verb/verb + 1, raw: $F1(1, 30) = 17.01$; $p < .001$, $F2(1, 22) = 27.03$, $p < .001$; combined verb/verb + 1, per word: $F1(1, 30) = 16.59$, $p < .001$, $F2(1, 22) = 23.37$, $p < .001$), Right-Bounded RTs (verb: $F1(1, 30) = 10.42$, $p < .005$, $F2(1, 22) = 9.59$, $p < .01$; verb + 1, raw: $F1(1, 30) = 4.37$, $p < .05$, $F2(1, 22) = 5.29$, $p < .05$; verb + 1, per word: $F1(1, 30) = 2.81$, $p = .10$, $F2(1, 22) = 3.17$, $p = .09$; combined verb/verb + 1, raw: $F1(1, 30) = 16.71$, $p < .001$, $F2(1, 22) = 21.50$, $p < .001$; combined verb/verb + 1, per word: $F1(1, 30) = 15.58$, $p < .001$, $F2(1, 22) = 17.17$, $p < .001$), and Total RTs (verb: $F1(1, 30) = 25.91$, $p < .001$, $F2(1, 22) = 20.69$, $p < .001$; verb + 1: $F1(1, 30) = 12.02$, $p < .005$, $F2(1, 22) = 13.14$, $p < .005$; combined verb/verb + 1: $F1(1, 30) = 25.81$, $p < .001$, $F2(1, 22) = 25.87$, $p < .001$). In these temporarily sentences, there was also a trend toward slower First-Pass RTs at the disambiguating verb, $F1(1, 30) = 4.89$, $p < .05$, $F2(1, 22) = 3.76$, $p = .06$, but not at the region immediately following verb (verb + 1, raw: $F1 < 1$, $F2 < 1.5$; verb + 1, per word: both F's < 1). There were also indications that the processing difficulty for temporarily ambiguous sentences spilled over into the final region of the sentence (verb + 2). In this region, temporarily ambiguous sentences had longer First Fixation Durations, $F1(1, 30) = 4.26$, $p < .05$, $F2(1, 22) = 4.56$, $p < .05$, and Total RTs, $F1(1, 30) = 4.93$, $p < .05$, $F2(1, 22) = 6.41$, $p < .05$.

SPR

Table 6 presents the mean SPR RTs for temporarily ambiguous and unambiguous sentences at and after the second-clause verb. At this verb, there was a trend suggesting that temporarily ambiguous sentences were read more slowly than unambiguous sentences. This difference was significant in the by-items analysis, $F2(1, 22) = 6.26$, $p < .05$, but not in the by-subjects analysis, $F1(1, 30) = 1.60$. There was also an trend suggesting that the words immediately following the second clause verb were read more slowly in temporarily ambiguous sentences (verb + 1: $F1(1, 30) = 4.28$, $p < .05$, $F2(1, 22) = 3.62$, $p = .07$; combined verb/verb + 1: $F1(1, 30) = 3.68$, $p = .06$, $F2(1, 22) = 7.70$, $p < .05$).

Table 5 Mean reading times for NP vs S coordination sentences in eye tracking

		Region		
		Verb	Verb + 1	Verb + 2
	Unambiguous	reported	the crime	to the police
	Ambiguous	reported	the crime	to the police
First Fixation	Unambiguous	251	245	254
	Ambiguous	255	254	270
First-Pass RT	Unambiguous	223	280	483
	Ambiguous	246	293	493
	Unambiguous	522		
	Ambiguous	560		
First-Pass RT (per word)	Unambiguous		175	179
	Ambiguous		181	182
	Unambiguous	199		
	Ambiguous	212		
Go-Past RT	Unambiguous	247	340	461
	Ambiguous	297	464	486
	Unambiguous	587		
	Ambiguous	761		
Go-Past RT (per word)	Unambiguous		215	166
	Ambiguous		288	171
	Unambiguous	224		
	Ambiguous	287		
Right-Bounded RT	Unambiguous	230	304	444
	Ambiguous	263	337	463
	Unambiguous	557		
	Ambiguous	651		
Right-Bounded RT (per word)	Unambiguous		190	162
	Ambiguous		208	166
	Unambiguous	212		
	Ambiguous	246		
Total RT	Unambiguous	296	391	561
	Ambiguous	375	462	605
	Unambiguous	687		
	Ambiguous	837		

G-maze and L-maze

Table 6 also presents the mean G-maze and L-maze RTs for temporarily ambiguous and unambiguous sentences at and after the second-clause verb. There were no RT differences at any of the relevant words/regions in either task (all F's < 1.5).

Taken together, these results show that only the eye-tracking task was able to provide clear indications of processing difficulty for sentences involving NP versus S coordination

Table 6 Mean reading times for NP vs S coordination sentences in SPR, G-maze, and L-maze

		Region		
		Verb	Verb + 1	Verb + 2
	Unambiguous	reported	the crime	to the police
	Ambiguous	reported	the crime	to the police
SPR	Unambiguous	363	332	372
	Ambiguous	415	350	371
	Unambiguous	345		
	Ambiguous	377		
G-maze	Unambiguous	1435	926	903
	Ambiguous	1441	942	895
	Unambiguous	1127		
	Ambiguous	1146		
L-maze	Unambiguous	815	699	680
	Ambiguous	817	691	690
	Unambiguous	747		
	Ambiguous	743		

ambiguity. This task revealed a garden-path effect at and immediately after the verb that indicated the violation of an apparent conjoined-NP interpretive bias. The results from the SPR experiment were consistent with this bias; however, none of the trends that were suggestive of a garden-path effect in this task were statistically reliable. Neither of the maze tasks revealed results consistent with this bias.

Discussion

This study examined the processing of three sentence types containing temporarily ambiguous structural configurations—RC attachment sentences, adverb attachment sentences, and (NP vs. S) coordination sentences—using eye tracking, SPR, the G-maze task, and the L-maze task. We set out to determine (i) whether these tasks would be capable of revealing the predicted pattern of processing time differences on each sentence type and (ii) whether these differences would be indicated precisely at the predicted word/region. With regard to the first question, only eye tracking yielded results that were clearly consistent with the expected effects on each sentence type. The results of the SPR task were consistent with the predicted effects for the RC and adverb attachment sentences, but revealed only a (delayed) statistically-unreliable trend suggestive of the expected processing time difference for the coordination sentences. For their part, both the G-maze and L-maze tasks revealed robust indications of the predicted effects for the RC and adverb attachment sentences, but oddly did not show any effects (not even delayed, "holdover" effects) for the coordination sentences.

In terms of the second question, although eye tracking indicated RT differences at the critical word/region for the adverb attachment and coordination sentences, it revealed only delayed effects for the RC attachment sentences—in the form of (i) a "holdover" effect at the region immediately following the disambiguating word under one of the "initial pass"

RT measures (specifically, per-word Right-Bounded RT) and (ii) a Total RT difference at the relative clause up to and including this critical word. The SPR task, on the other hand, showed an RT difference at the critical region for the adverb attachment sentences, but only delayed effects for both the RC attachment sentences and coordination sentences—specifically, "holdover" effects on both sentence types that were obtained only when averaging over multiple words (and that were not statistically reliable in the case of the latter structure). In contrast to eye tracking and SPR, both the G-maze task and L-maze task showed RT differences at the critical region for both the RC attachment and adverb attachment sentences, but nothing for the coordination sentences. That is, where these maze tasks revealed effects, they did so at precisely the predicted word/region.

In light of the large RT disparities among these tasks on each of the sentence types, comparisons of the relative effects at the predicted word/region are perhaps best expressed in terms of effect size, or as the proportion of the variance accounted for by the predictor variable (i.e., sentence condition). Figure 3 shows the effect size—specifically, partial Eta squared (ηp2)—at the critical word/region for each sentence type under each task. For the sake of simplicity, the effect sizes reported here are for the by-subjects (F1) analyses only. For the eye-tracking analyses, effect sizes are reported only for (raw) First-Pass RT, Go-Past RT, and Total RT—three of the most commonly reported RT measures for this task. For the RC attachment sentences, both the G-maze task (ηp2 = .38) and L-maze task (ηp2 = .20) showed large effect sizes at the disambiguating reflexive, with sentence condition (high attachment, low attachment) accounting for more than 1/3 of the total variance in the former task and for 1/5 of the variance in the latter. The effect sizes at this word in the SPR task (ηp2 = .04) as well as under the measures in the eye-tracking task (First-Pass RT: ηp2 = .02; Go-Past RT: ηp2 = .06; Total RT: ηp2 = .02) were rather modest, with sentence condition accounting for less than 10% of the variance under all of these measures, and in most cases less than 5%. For the adverb attachment sentences, all of the word-by-word reading tasks (SPR: ηp2 = .21; G-maze: ηp2 = .73; L-maze: ηp2 = .28) as well as the eye-tracking task (First-Pass RT: ηp2 = .22; Go-Past RT: ηp2 = .48; Total RT: ηp2 = .56) revealed robust effects at the time adverbial. The G-maze task and eye tracking yielded especially large effect sizes at this critical region, with the predictor variable accounting for close to 3/4 of the variance in the G-maze task, and for around half of the variance in eye tracking (at least under the Go-Past RT and Total RT measures). For the coordination sentences, on the other hand, although the eye-tracking task revealed large effect sizes at the disambiguating second-clause verb (First-Pass RT: ηp2 = .14; Go-Past RT: ηp2 = .33; Total RT: ηp2 = .46), the predictor variable did not account for much of the variance at this region in any of the word-by-word reading tasks (SPR: ηp2 = .05; L-maze: ηp2 = .00; G-maze: ηp2 = .00).

In light of these findings, it is important to reconsider the pros and cons of each task tested in this study. As suggested in the introduction, one way to think of these advantages and disadvantages is in terms of the strategies available to participants during the task and the extent to which the possibility of adopting multiple strategies might detract from the task's ability to indicate characteristics of online sentence processing. With respect to eye tracking, it was suggested that the rather expansive strategy space available during this task might be considered a double-edged sword. On the one hand, this task allows participants to adopt any of the strategies that they might normally use when reading, and thus allows for investigations of relatively "natural" sentence processing. On the other hand, this feature might lead to situations in which effects are obscured if certain reading strategies (e.g., a skimming strategy) are adopted over others (or if a number of different reading strategies are averaged over in the analysis). It is important to point out that although eye tracking was the only task to reveal the expected patterns of processing difficulty on all of the sentence types, concerns

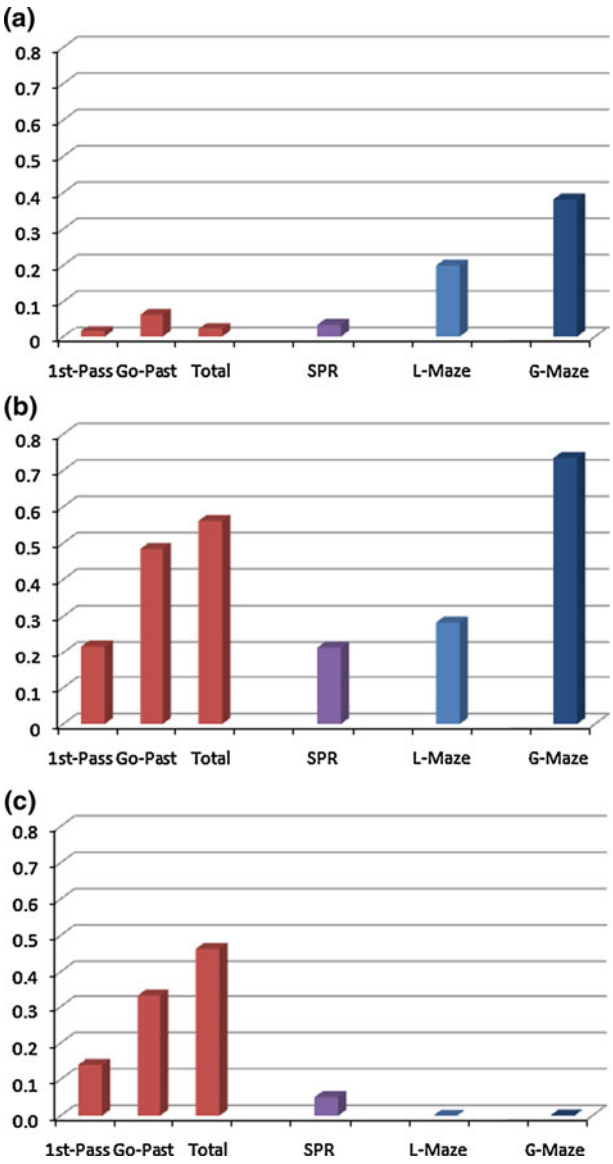

Fig. 3 Effect size comparisons at critical region/word for **a** RC attachment; **b** adverb attachment; and **c** NP versus S coordination

about task insensitivity are relevant to one of these sentence types in particular—namely, RC attachment sentences. As discussed above, on this sentence type, a processing time difference indicating difficulty for high RC attachment was revealed immediately after the critical (disambiguating) word in only one of the "initial pass" RT measures. One could argue that this presents an interpretive dilemma in that this difference was not obtained exactly at the predicted location. However, it is important to reiterate that there was also a Total RT difference in the same direction at the early part of the structurally ambiguous relative clause, up

to and including the disambiguating word. Therefore, it can be inferred from the *complete* pattern of results that the predicted region was indeed (at least part of) the primary locus of processing difficulty in these sentences. This would appear to be a clear illustration that in eye tracking, the richness of the data can largely compensate for the occasional insensitivity of some measures to processing differences.

This saving grace of multiple measures does not apply to SPR. In this task, a single RT measure (button-pushing time) is taken to reflect processing time. And as indicated above, it appears that this measure can regularly yield delayed, distributed effects—or effects that are revealed only after the word/region on which they are predicted and that are "smeared" over multiple words. These characteristics obviously present certain interpretive problems for investigations into how sentence representations are developed during real-time comprehension, so it is important to consider why they might occur. There are a number of possibilities. One is simply that it is difficult to synchronize button-pushing-time and processing time, particularly in the absence of some explicit criterion for initiating this response to a given word. A related explanation is that subjects might develop a regular button-pushing "rhythm" that is sometimes subtly adjusted in (delayed) response to comprehension disruptions. Indeed, this "tapper" strategy is a perfectly reasonable way to approach the SPR task, and one that might account for the delayed, distributed effects found in this and many other studies. To investigate this possibility, several follow-up analyses were conducted in which we eliminated the data from potential "tappers". In order to do this, the subjects in each list were ranked in terms of the overall variability in their button-pushing responses. Those with the least variability were eliminated—first eight subjects from each list, then six, and finally four—and the statistical analyses were rerun. Interestingly, these follow-up analyses did not reveal any change in the patterns of results for the critical words in the sentences of interest. That is, eliminating potential "tappers" did not appear to influence the extent to which processing time differences were indicated at the critical words in these sentences. It could be the case that these procedures for identifying "tappers" were not appropriate, so further investigation into this issue is necessary.

The maze task involves a strategy space that is different from that of either eye tracking or SPR. This is largely due to the fact that, in contrast to eye tracking and SPR, the maze task requires a minimum level of processing to be completed at each word of the sentence. In the case of the G-maze, each word must be integrated into the sentence; whereas in the case of the L-maze, each word must at least be recognized. As demonstrated above, the restrictions that these task features place on the strategies available to participants appear to allow for robust, highly "localized" indications of the processing costs associated with integrating each successive word into developing sentence representations. It should be noted, however, that the complete lack of results for coordination sentences under the maze tasks poses some problems for the wide application of these methodologies. Thus, it is important to consider why these results (or lack thereof) were obtained. One possibility relates to the fact that unlike in eye tracking or SPR, the maze task presents each sentence one word at a time in a way that does not allow the reader to know the length of the sentence. This may encourage readers to close the current clause or constituent whenever possible. That is, maze tasks might encourage hyper-incrementality in the integration of words into sentences. Consistent with this explanation, pilot studies run in our lab have shown that maze task readers have difficulty responding to a word like *office* when it is preceded by *The post*, suggesting that participants tend to close the constituent at *post* (i.e., [NP The post]) and that they then have difficulty reanalyzing the structure of this phrase to allow *office* to be integrated as part of a compound noun. Comparably, with respect to the coordination sentences of interest in the present study, it may be the case that maze task readers closed the clause at the end of the

first (object) NP (*the jeweler*) in both ambiguous and unambiguous sentences. When they then read the second NP (*and the salesman*), they may have been starting a new clause under both conditions. This comparable structural analysis for both ambiguous and unambiguous coordination sentences might account for the similar responses to the "disambiguating" verb in both sentence types. If this explanation is correct, it would suggest that maze tasks (at least as they are instantiated in the present study) are not appropriate for investigating clause/constituent closure commitments during online sentence processing. The lack of results for the coordination sentences also points to the need for further research using the maze task in conjunction with other methods and for careful consideration of the cases in which the results from these tasks fail to converge.

There is of course no task that serves perfectly for all sentence processing questions. Each task necessarily focuses on processing under one presentation modality (auditory/visual) over the other, and each has its (sometimes questionable) linking assumptions as well as its advantages and disadvantages/limitations. In light of this, it is important to develop new methods. We have presented one such method in this paper—a word-by-word reading methodology that requires (in the case of the G-maze) or encourages (in the case of the L-maze) incremental integration during sentence processing and that thus has the potential to yield robust, highly localized indications of the processing costs associated with this integration (again, at least for some sentence types). It should be emphasized that the maze task is not meant to replace any methodology. It is simply meant to add to the methods currently available and to open new avenues of inquiry, particularly in the domains of lexical and sentence processing. It is important to note that the maze task will allow researchers to investigate questions in ways that are impossible with other online reading methodologies. To take one example, in the investigation of agreement phenomena, the G-maze task could be used to present both grammatical and ungrammatical alternatives (e.g., [*agree agrees*]) in order to understand the extent to which the ungrammatical alternative is considered a viable "competitor" for integration (for a methodology that approximates this proposal, see e.g., Staub 2009). Similarly, one could also use the G-maze to present two grammatical alternatives, each of which would force the reader to adopt very different structural analyses for the sentence. For example, with reference to the RC attachment sentences examined in the present study, reflexives that indicate high and low RC attachment could be displayed simultaneously—e.g., after the selecting the words *The son of the actress who shot*, the participant could be presented with the pair [*herself himself*]. This type of experiment might be able to provide an indication of the extent to which competing analyses are entertained during online sentence processing. In this way, it is hoped that the present study will help to motivate novel approaches to the investigation of language processing.

Acknowledgments We would like to thank Stacey Claspill, Leslie Darnell, Katherine Plattner, and Devin St. John for assisting with data collection. This research was supported in part by the Language Learning Dissertation Grant to N. Witzel. Earlier version of this paper was presented at the CUNY Conference on Human Sentence Processing (New York, NY, March 2010).

References

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

Bever, T. G., & McElree, B. (1988). Empty categories access their antecedents during comprehension. *Linguistic Inquiry, 19*, 35–43.

Boland, J. E., Tanenhaus, M. K., & Garnsey, S. M. (1990). Evidence for the immediate use of verb control information in sentence processing. *Journal of Memory and Language, 29*, 413–432.

Carreiras, M., & Clifton, C., Jr. (1993). Relative clause interpretation preferences in Spanish and English. *Language and Speech, 36*, 353–372.

Carreiras, M., & Clifton, C., Jr. (1999). Another word on parsing relative clauses: Eyetracking evidence from Spanish and English. *Memory & Cognition, 27*, 826–833.

Clifton, C., Jr., Frazier, L., & Connine, C. (1984). Lexical expectations in sentence comprehension. *Journal of Verbal Learning and Verbal Behavior, 23*, 696–708.

Cuetos, F., & Mitchell, D. C. (1988). Cross-linguistic differences in parsing: Restrictions on the use of the late closure Strategy in Spanish. *Cognition, 30*, 73–105.

Forster, K. I. (1970). Visual perception of rapidly presented word sequences of varying complexity. *Perception & Psychophysics, 8*, 215–221.

Forster, K. I. (2010). Using a maze task to track lexical and sentence processing. *The Mental Lexicon, 5*, 347–357.

Forster, K. I., & Forster, J. C. (2003). DMDX: A windows display program with millisecond accuracy. *Behavior Research Methods, Instruments, & Computers, 35*, 116–124.

Forster, K. I., Guerrera, C., & Elliot, L. (2009). The maze task: Measuring forced incremental sentence processing time. *Behavior Research Methods, 41*, 163–171.

Foss, D. J. (1969). Decision processes during sentence comprehension: Effects of lexical item difficulty and position upon decision times. *Journal of Verbal Learning and Verbal Behavior, 8*, 457–462.

Frazier, L. (1987). Syntactic processing: Evidence from Dutch. *Natural Language and Linguistic Theory, 5*, 519–559.

Frazier, L., & Clifton, C., Jr. (1997). Construal: Overview, motivation, and some new evidence. *Journal of Psycholinguistic Research, 26*, 277–295.

Freedman, S. E., & Forster, K. I. (1985). The psychological status of overgenerated sentences. *Cognition, 19*, 101–131.

Garrett, M., Bever, T. G., & Fodor, J. A. (1965). The active use of grammar in speech perception. *Perception & Psychophysics, 1*, 30–32.

Gordon, P. C., Hendrick, R., & Johnson, M. (2004). Effects of noun phrase type on sentence complexity. *Journal of Memory and Language, 51*, 97–114.

Hoeks, J. C. J., Hendriks, P., Vonk, W., Brown, C. M., & Hagoort, P. (2006). Processing the noun phrase versus sentence coordination ambiguity: Thematic information does not completely eliminate processing difficulty. *Quarterly Journal of Experimental Psychology, 59*, 1581–1599.

Just, M. A., Carpenter, P. A., & Woolley, J. D. (1982). Paradigms and processes in reading comprehension. *Journal of Experimental Psychology: General, 111*, 228–238.

Kamide, Y., Altmann, G. T. M., & Haywood, S. L. (2003). The time-course of prediction in incremental sentence processing: Evidence from anticipatory eye movements. *Journal of Memory and Language, 49*, 133–156.

McElree, B., & Bever, T. G. (1989). The psychological reality of linguistically defined gaps. *Journal of Psycholinguistic Research, 18*, 21–35.

Mitchell, D. C. (2004). On-line methods in language processing: Introduction and historical review. In M. Carreiras & C. Clifton (Eds.), *The on-line study of sentence comprehension: Eye-tracking, ERP and beyond* (pp. 15–32). Hillsdale, NJ: Erlbaum.

Nicol, J. L., Forster, K. I., & Veres, C. (1997). Subject-verb agreement processes in comprehension. *Journal of Memory and Language, 36*, 569–587.

Nicol, J., & Swinney, D. (1989). The role of structure in coreference assignment during sentence comprehension. *Journal of Psycholinguistic Research, 18*, 5–19.

Phillips, C., Kazanina, N., & Abada, S. H. (2005). ERP effects of the processing of syntactic long-distance dependencies. *Cognitive Brain Research, 22*, 407–428.

Qiao, X. & Forster, K. I. (2008). *Object relatives ARE easier than subject relatives in Chinese*. Poster presented in the 14th annual conference on architectures and mechanisms for language processing. Cambridge, UK.

Rastle, K., Harrington, J., & Coltheart, M. (2002). 358,534 nonwords: The ARC nonword database. *Quarterly Journal of Experimental Psychology, 55A*, 1339–1362.

Rayner, K. (2009). Eye movements and attention in reading, scene perception, and visual search. *Quarterly Journal of Experimental Psychology, 62*, 1457–1506.

Rayner, K., & Pollatsek, A. (2006). Eye movement control in reading. In M. Traxler & M. Gernsbacher (Eds.), *Handbook of psycholinguistics* (2nd ed., pp. 609–653). Cambridge, MA: Elsevier.

Rayner, K., & Sereno, S. C. (1994). Eye movements in reading: Psycholinguistic studies. In M. A. Gernsbacher (Ed.), *Handbook of psycholinguistics* (1st ed., pp. 57–81). San Diego, CA: Academic Press.

Staub, A. (2009). On the interpretation of the number attraction effect: Response time evidence. *Journal of Memory and Language, 60*, 308–327.

Staub, A., & Rayner, K. (2007). Eye movements and on-line comprehension processes. In G. Gaskell (Ed.), *The Oxford handbook of psycholinguistics* (pp. 327–342). Oxford, UK: Oxford University Press.

Tanenhaus, M. K., Carlson, G., & Trueswell, J. C. (1989). The role of thematic structures in interpretation and parsing. *Language and Cognitive Processes, 4*, 211–234.

Traxler, M. J., Pickering, M. J., & Clifton, C., Jr. (1998). Adjunct attachment is not a form of lexical ambiguity resolution. *Journal of Memory and Language, 39*, 558–592.

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

Waters, G. S., & Caplan, D. (2004). Verbal working memory and on-line syntactic processing: Evidence from self-paced listening. *The Quarterly Journal of Experimental Psychology Section A: Human Experimental Psychology, 57*, 129–163.

Witzel, J., & Forster, K. (2009). *Lexical co-occurrence and ambiguity resolution*. Manuscript submitted for publication, University of Arizona.

Witzel, J., & Witzel, N. (2009). *Pre-head gap-filling in Japanese sentence processing*. Poster presented at the 22nd annual meeting of the CUNY conference on human sentence processing, University of California, Davis, CA.

Witzel, J., Witzel, N., & Nicol, J. (in press). Deeper than shallow: Structure-based parsing biases in L2 sentence comprehension. *Applied Psycholinguistics.*