Language and Cognitive Processes

Scanpaths reveal syntactic underspecification and reanalysis strategies

Titus von der Malsburg [a] & Shravan Vasishth [a]

[a] Department of Linguistics, University of Potsdam, Potsdam,
Germany

Version of record first published: 01 Nov 2012.

ⓘFirst

PLEASE SCROLL DOWN FOR ARTICLE

Scanpaths reveal syntactic underspecification and reanalysis strategies

Titus von der Malsburg and Shravan Vasishth

Department of Linguistics, University of Potsdam, Potsdam, Germany

What theories best characterise the parsing processes triggered upon encountering ambiguity, and what effects do these processes have on eye movement patterns in reading? The present eye-tracking study, which investigated processing of attachment ambiguities of an adjunct in Spanish, suggests that readers sometimes underspecify attachment to save memory resources, consistent with the good-enough account of parsing. Our results confirm a surprising prediction of the good-enough account: high-capacity readers commit to an attachment decision more often than low-capacity participants, leading to more errors and a greater need to reanalyse in garden-path sentences. These results emerged only when we separated functionally different types of regressive eye movements using a scanpath analysis; conventional eye-tracking measures alone would have led to different conclusions. The scanpath analysis also showed that rereading was the dominant strategy for recovering from garden-pathing. Our results may also have broader implications for models of reading processes: reanalysis effects in eye movements occurred late, which suggests that the coupling of oculo-motor control and the parser may not be as tight as assumed in current computational models of eye movement control in reading.

Consider the sentence in (1), taken from Frazier and Rayner (1982). After the human sentence comprehension mechanism (hereafter, the parser) encounters the verb *forgot*, it must decide whether the subsequent phrase, the noun phrase *her husband*, is an object of the verb *forgot*, or the subject of a subsequent clause (e.g., ...*forgot(,) her husband*...). If the parser commits to the object reading, it encounters a problem at *didn't*, because this word cannot be attached to the

previously built structure. The parser is *garden-pathed* and must somehow revise its earlier decision at the verb.

(1) Although Mary forgot her husband didn't seem very upset yesterday.

Garden-pathing is, of course, not limited to English; for example, in Spanish, temporary attachment ambiguities can also lead to garden-pathing (Carreiras & Clifton, 1993; Cuetos & Mitchell, 1988; Igoa, Carreiras, & Meseguer, 1998; Meseguer, Carreiras, & Clifton, 2002). For example, in (2a) and (2b), the adverbial clause (AdvC) beginning with *cuando* ("when"/"if") preferentially attaches to the more local verb *se levantaran* ("stand up") rather than the nonlocal verb *dijo* ("say"), but the mood of the verb inside the adverbial clause determines whether the attachment is in fact nonlocal instead of local: *entraron* ("came") in (2a) is in indicative mood as the main verb and results in nonlocal high attachment to *dijo* (i.e., a garden-path), whereas *entraran* ("come") in (2b) is in subjunctive mood and results in local attachment to *se levantaran*.

(2)	a.	*El*	*profesor*	**dijo**	*que*	*los*	*alumnos*	*se levantaran*	*del asiento*
		The	teacher	said	that	the	students	had to stand up	from their seats
		[*AdvC*	*cuando*	*los*	*directores*	**entraron**	*en la clase*].		
		[*AdvC*	when	the	directors	came	into the class room].		
	b.	*El*	*profesor*	*dijo*	*que*	*los*	*alumnos*	**se levantaran**	*del* asiento
		The	teacher	said	that	the	students	had to stand up	from their seats
		[*AdvC*	*cuando*	*los*	*directores*	**entraran**	*en la clase*].		
		[*AdvC*	when	the	directors	come	into the class room].		

What processes unfold when the parser is garden-pathed? Garden-path theory (Frazier, 1979), a classical account of sentence comprehension difficulty, assumes that a reanalysis step is triggered at the disambiguation point that restructures the parse to the correct one. The details underlying this reanalysis step are of interest not only to sentence comprehension researchers, but are also of great relevance to models of oculo-motor control in reading, for example, E-Z Reader 10 (Reichle, Warren, & McConnell, 2009), that assume a role of higher-level cognitive processes in driving eye movements in reading. A theory of reading that includes effects of parsing difficulty would need to explain how reanalysis processes and eye movements interact.

As a first step towards uncovering the strategies that the parser deploys when faced with temporary ambiguities, we can begin by posing a simple question: are the eyes and parser tightly coupled during syntactic reanalysis, or are they only loosely related? Several interesting proposals exist regarding the eye–parser connection; these have been proposed specifically in the context of garden-path constructions. For example, Frazier and Rayner (1982) provided evidence for an eye–parser relationship that they termed selective reanalysis: when the parser is garden-pathed, the eye moves selectively to the previously ambiguous region, suggesting that the parser is triggering a targeted repair operation. Selective reanalysis implies a tight coupling of the eye with parser actions, and in fact, using items like (2) above, Meseguer et al. (2002) provided evidence consistent with selective reanalysis. They found more frequent eye movements from the postdisambiguation region to the head of the adverbial clause in the garden-path condition.

Selective reanalysis has not gone uncontested. Mitchell, Shen, Green, and Hodgson (2008) proposed an alternative, called the time-out hypothesis, according to which the parser is decoupled from oculo-motor control during reanalysis. In this account, regressions are executed not to seek out information but simply to buy time for the

parser for finishing processing of earlier material. These regressive eye movements are assumed to be influenced by low-level visual details of the sentence, e.g., the position of line breaks, but not by its linguistic structure. Although Mitchell et al. found evidence showing that the layout does indeed influence regression trajectories, they also found some evidence for at least a loose coupling between the eye and parser (p. 289). Regressions were found to be largely driven by layout, but there was also a tendency of the eyes to "gravitate" to informative material.

An important problem that emerges when we investigate the eye-parser connection is that conventional eye movement measures may occasionally fall short of giving us interpretable answers. Specifically, fixation durations at particular words cannot tell us much about *patterns* of eye movements, which vary in time and space. For this reason, Frazier and Rayner (1982) informally examined scanpaths (trajectories of eye movements), and both Meseguer et al. (2002) and Mitchell et al. (2008) developed more scanpath-oriented measures, which Mitchell et al. call "regression signatures", to characterise eye movement patterns. However, one limitation of regression signatures is that they only inform us about the distribution of landing sites of regressive saccades from a specific region, but they cannot provide information about more complex spatio-temporal patterns in sequences of fixations. To address this methodological problem, von der Malsburg and Vasishth (2011) developed a method for analysing whole scanpaths instead of just single saccades, and show that it can uncover novel information about eye movement patterns. Using Meseguer and colleagues' data, von der Malsburg and Vasishth showed that one common fixation pattern in their data consisted of rereading the sentence, suggesting that, at least in some cases, the parser simply discarded the incorrect parse and started over. Rereading seems inefficient compared to targeted repair, but it may be a good strategy for readers with low working memory capacity. To our knowledge, Lewis (1998) was the first to notice that rereading allows one to trade off processing time for savings in memory. As Lewis puts it:

> This strategy [rereading] requires no memory for input or prior parsing states—it needs just enough memory for the parser to remember not to continue going down the same path. The drawback of this method is time, but it is a reliable strategy when all else fails.

Indeed, Frazier and Rayner (1982, p. 196) also found evidence for rereading in their data (they did not examine the interaction with working memory capacity—we take this up in our experiment, described below). However, they concluded that rereading was probably not an index of reanalysis processes (p. 203) because rereading began from the end of the sentence; the sentence-final region is widely believed to trigger processes that go beyond initial parsing events. In the Meseguer et al. data, regressive rereading was also launched from the final region of the sentence—although not necessarily from the final word—because the disambiguating material was close to the sentence end. Hence, the possibility that readers in the Meseguer et al.'s study just read the high-attachment sentences a second time to make sure they got it right cannot be ruled out; the preliminary conclusion by von der Malsburg and Vasishth that rereading may indeed be signalling the parser's attempt to escape from garden-paths requires more empirical support. There was also a second property of our earlier results that did not allow us to draw definitive conclusions about the effect of parsing events on eye movements. If we assume a deterministic account of parsing along the

lines of garden-path theory,[1] reanalysis is expected only in the high-attachment condition; however, rereading was also found in the low-attachment condition (although this occurred less often than in the high-attachment case). This finding allows two conclusions: either rereading does not (only) reflect reanalysis, or reanalysis can occur in both the high-attachment and low-attachment conditions. The latter is predicted by the unrestricted race model of parsing, in which nonsyntactic cues such as discourse- and world-knowledge can influence parsing decisions and sometimes override a syntactic preference for a structure (Traxler, Pickering, & Clifton, 1998).

This state of affairs demonstrates that the interpretation of scanpath patterns is contingent on what kind of parser we assume. If we assume a deterministic parser, not all cases of rereading can be attributed to reanalysis. If, however, we assume a race parser, this is quite possible. This shows that the interpretation of effects observed upon disambiguation crucially depends on assumptions about parsing events that occur during the first pass over the ambiguous material. Various earlier studies, including our own, simply assumed that the parser would operate in a fashion predicted by garden-path theory. Therefore, the present study was designed to thoroughly investigate the contingencies between first pass parsing events and effects of disambiguation. Our approach consists of two components: first, we investigate parsing decisions on the ambiguous region by contrasting ambiguous sentences such as those in (2) with unambiguous sentences of an analogous structure. (Note that when we refer to ambiguous sentences, we mean *temporarily* ambiguous sentences. Globally ambiguous sentences were not studied in the reported experiment.) This allows us to evaluate the predictions of various theories of parsing for first pass parsing mechanisms. Second, following Lewis, we examine the processes leading up to garden-pathing from an individual differences angle. Specifically, we examine how individual participants' performance on the operation span test (typically assumed to reflect working memory capacity) affects their parsing decisions. Working memory effects, if they turn up, provide us with additional constraints for pinning down first pass parsing decisions but they may also allow us a functional interpretation of the various eye movement patterns that are triggered by disambiguation. Two other such patterns were found in our earlier scanpath analysis of the Meseguer et al. data, but their function remained unclear.

In this study, these goals will be achieved by extending the Meseguer et al. design with an unambiguous baseline condition, as shown in (3c) and by assessing the working memory capacity of the participants with a working memory span task.

(3)	El	profesor	dijo	que	los	alumnos	se levantaran	del asiento ...
	The	teacher	said	that	the	students	had to stand up	from their seats ...
a.	[AdvC	cuando	los	directores	**entraron**	en la clase de música].
	[AdvC	when	the	directors	came	into the music class].
b.	[AdvC	cuando	los	directores	**entraran**	en la clase de música].
	[AdvC	when	the	directors	come	into the music class].
c.	[AdvC	**si**	los	directores	entraban	en la clase de música].
	[AdvC	if	the	directors	come	into the music class].

---

[1] Frazier and Fodor (1978) consider the possibility that the mechanism underlying minimal attachment and late closure consists of a race between structural alternatives and that the parser adopts whatever structure is finished first. This idea of a competition amounts to the proposal that the parser may in fact be nondeterministic. However, Frazier and Fodor (1978) argue that the structures preferred by minimal attachment and late closure are easier to construct because they require fewer applications of grammar rules. Unless further assumptions are made, this renders the garden-path model effectively deterministic.

In (3c), the word *si* ("if") replaces *cuando* ("when"/"if"). The word *si* rules out high attachment right from the beginning and the conditional clause can only attach low. High-attachment is ruled out because in Spanish a verb with episodic tense *pretérito perfecto simple* cannot be modified with a *si*-clause. Below we will briefly discuss the predictions of various accounts of parsing and we will see that these three conditions allow us to distinguish between them.

In one of the scanpath patterns that our earlier scanpath analysis identified, the eyes regressed from the spillover region (*en la clase*) back to the disambiguating material. This frequent pattern occurred equally often in the high- and low-attachment conditions and we speculated that it reflects a kind of diagnosis procedure. If true, the pattern should occur less often in condition (3c), because there the target region of these regressions (*entraban*, "come") does not contain relevant information because the sentence has already been disambiguated much earlier (at *si*).

Another difference compared with the original material by Meseguer et al. is that we extend the end of the sentence with extra material (*de música*), separating disambiguation effects more clearly from sentence wrap-up. If we still find rereading scanpaths starting from the spillover region (not including the final region), we can more safely attribute them to syntactic reanalysis.

How assumptions about the parser architecture determine the predictions for reanalysis

In the above discussion, we framed the reanalysis question mainly in terms of Frazier's garden-path theory, but the assumptions made in that model do not go uncontested in the literature and it is worth briefly considering the alternative proposals. Frazier's garden-path theory assumes that (1) syntactically complete structural representations are built as the parse progresses; (2) the parser relies, initially, on syntactic (as opposed to, e.g., semantic) information to make decisions; (3) the parser is serial (only one structure is built at any choice point); and (4) the parser is deterministic in the sense that it will always execute the same sequence of operations for a type of sentence even if there are multiple possibilities at some point in the sentence. Relaxing these assumptions has an interesting impact on what happens at the verb of the adverbial clause (*entraron/entraran/entraban*, "came/come/come") in (3), and earlier at the region that is ambiguous in the two first conditions (the phrase *cuando/si los directores*, "when/if the directors"). The resulting theories make intriguing predictions some of which are quite different from those made by garden-path theory. Let us unpack these predictions using example (3) above. (Refer to Table 1 for a list of the theoretically interesting regions of these sentences.) Most of these theories are designed to be testable using reading times and therefore do not make explicit predictions about scanpaths. Hence, we will focus on reading times as an indicator of first pass parsing events. It is important to note that the goal of this exercise to formulate and evaluate predictions of the various theories of sentence processing is not to draw definite conclusions about them but rather to get a sense of the spectrum of proposals that have been made for the processing of temporarily ambiguous sentences. Also we will see that some of these ideas are not mutually exclusive.

Construal

Consider first a refinement of garden-path theory, construal (Carreiras & Clifton, 1993; Frazier & Clifton, 1996). This theory makes a distinction between primary and

TABLE 1
The critical regions of the material in (3)

Condition	Head	Pre-verbal	Verb	Spillover
(a)	*cuando*	*los directores*	*entraron*	*en la clase de ...*
	when	the directors	came	into the class ...
			(disambiguates, high)	
(b)	*cuando*	*los directores*	*entraran*	*en la clase de ...*
	when	the directors	come	into the class ...
			(disambiguates, low)	
(c)	*si*	*los directores*	*entraban*	*en la clase de ...*
	if	the directors	come	into the class ...
	(unambiguous, low)			

nonprimary relations; primary relations are obligatory elements of a clause (e.g., arguments of a verb), while nonprimary relations include adjuncts such as the adverbial clause in (3). According to construal, primary relations are resolved immediately by the parser, but nonprimary relations are initially left underspecified and are only loosely "associated" with the last theta domain, i.e., the extended maximal projection of the last theta assigner (Carreiras & Clifton, 1993; Frazier & Clifton, 1996, 1997). Later processes can refine this association and bind the phrase to a target site inside that theta domain without any noticeable processing effort even if there are several candidate sites within this domain. If, however, the attachment site ultimately turns out to be outside this theta domain, structural revision is necessary, and this will cause an observable slowdown. One prediction of construal is thus the absence of reanalysis effects for certain types of ambiguities. For (3), the construal hypothesis predicts that the initial processing cost for the first part of the adverbial clause (*cuando/si los directores*, "when/if the directors") should be the same across both the temporarily ambiguous and the unambiguous conditions because the clause is in all cases only associated with the last theta domain, i.e., the projection of the embedded verb *se levantaran* ("stand up").

At the disambiguating region (the verb *entrar*, "come"), of course, the reanalysis cost should be apparent just as in garden-path theory because the correct attachment site is outside the last theta domain in the high-attachment condition. In other words, underspecification at the ambiguous region results in a similar set of predictions as that produced by garden-path theory. However, construal assumes different processes for dealing with adjunct attachment that operate at a "more leisurely pace", presumably because the attachment of adjuncts has little consequences for other parsing decisions. Hence, reanalysis effects may show up later in garden-path sentences involving the attachment of adjuncts than, e.g., in sentences with a subject/object ambiguity as in (1).

Good-enough parsing

One can relax even further the assumption that the parser is eager in its attachment decisions—it may sometimes refrain from committing to one or another parse altogether, maintaining only a partial representation of the sentence structure. This is an assumption in the good-enough account of parsing (Ferreira, Bailey, & Ferraro, 2002; Sanford & Sturt, 2002). Good-enough parsing focuses on the notion that the parser is not necessarily aiming for a complete and fully specified analysis of the sentence but that it might accept a partial or even inconsistent analysis if there is a

chance of succeeding at the task at hand without more in-depth processing. The good-enough account is mainly supported by studies demonstrating that readers often fail to provide accurate answers to comprehension questions relating to garden-path sentences (e.g., Christianson, Hollingworth, Halliwell, & Ferreira, 2001; see Ferreira et al., 2002 and Ferreira & Patson, 2007 for an overview). The influence of the task on online sentence processing was shown by Swets, Desmet, Clifton, and Ferreira (2008) in a self-paced reading study that manipulated the difficulty of comprehension questions. In this study, the well-known ambiguity advantage—ambiguous regions are easier to process than unambiguous ones—was stronger when the comprehension questions were easy. This was interpreted as showing that the speed-up on ambiguous material is due to underspecification and that this underspecification is strategic.

To our knowledge, there are—unfortunately—no descriptions of good-enough processing that allow precise predictions of reading behaviour. The predictions adopted in this study are based on the interpretation of good-enough processing presented in Swets et al. (2008): the parser may or may not decide to attach a phrase when the attachment site is ambiguous. When it does attach, it operates in the fashion predicted by garden-path theory, which means that it attaches low initially and revises later if necessary. When the parser decides not to attach the phrase, it stays with this decision until the end of the sentence. This means that the parser does not attach even when disambiguation information becomes available later in the sentence. Compared to cases where an attachment has to be made, the decision not to attach results in a speed-up in the ambiguous region because not doing any attachment presumably requires less time. Another prediction of good-enough processing is that the answers to comprehension questions should be (i) less accurate in the high-attachment condition than in the ambiguous low-attachment condition and (ii) less accurate in the ambiguous low-attachment condition than in the unambiguous low-attachment condition. Prediction (i) follows from the assumption that the language system, in absence of an interpretation supported by syntax, applies heuristics that favour the more common low-attachment interpretation. Prediction (ii) follows from the assumption that a fully determined syntactic interpretation is more reliable than the heuristics used in cases of underspecification. However, the answers to comprehension questions will certainly also be influenced by other mechanisms occurring after the sentence was read, and therefore the precise pattern of accuracies is hard to predict.[2]

The predictions for sentences as in (3) are (i) a speed-up in the pre-verbal part of the ambiguous conditions (*cuando los directores*, "when the directors") because in some trials the parser does not attach at all, which saves time; (ii) slight (relative to garden-path theory's predictions) reanalysis-driven difficulty in the disambiguation region because the parser commits to an attachment site only in some trials; in trials where it does not make a commitment, there is nothing to revise; (iii) underspecification preserves resources and may, therefore, occur more often in readers with low working memory capacity; this means that low-capacity readers should have a greater speed-up in the ambiguous regions and smaller reanalysis effects; and (iv) good-enough parsing

---

[2]Another version of good-enough processing was presented by Christianson et al. (2001), who assume that the low performance when answering questions about garden-path sentences results from failed reanalysis rather than from underspecification. According to this account, a good-enough parser uses the same parsing mechanisms as the garden-path model. The difference is that it maintains imprecise representations. The speed-up in ambiguous sentences observed by Swets et al. (2008) does not favour this interpretation because the Christianson et al. account does not predict such a speed-up.

predicts a worse performance in answering comprehension questions in the high-attachment condition than in the low-attachment conditions. This prediction rests on the assumption that readers have a preference to answer questions according to a low-attachment analysis. In earlier studies, this preference has been demonstrated for English sentences with an attachment ambiguity of a relative clause (Swets et al., 2008; Traxler et al., 1998).

Parallel models

Relaxing the seriality assumption in the garden-path theory also has an interesting impact on the predictions. Two classes of parallel parsers that have been proposed in the literature are constraint-based parsing (e.g., McRae, Spivey-Knowlton, & Tanenhaus, 1998) and ranked parallel parsing (e.g., Gibson, 1991). Both accounts assume that several interpretations of the sentence so far can be maintained in parallel. In constraint-based models, the various alternatives are competing and suppressing each other in a dynamical process. A criterion decides that an analysis is the winner when it has suppressed its alternative parses sufficiently. In ranked parallel parsers, no such competition exists; alternative interpretations are ranked according to how many resources they consume but co-exist peacefully otherwise. When new material renders the highest ranked analysis inconsistent, the alternatives are re-ranked, which requires processing effort. Both types of model are usually taken to predict more processing effort when faced with multiple structural alternatives (in the pre-verbal region of the ambiguous conditions). In constraint-based models, the additional effort is attributed to competition, and in ranked parallel parsers to the complexity of building several structures.

Garden-pathing triggers a re-ranking of structural alternatives and takes time in both types of models, but since all possible structures have already been built, re-inspection of earlier material should not be required. Thus, we would predict that a parallel parser does not trigger any regressive eye movements at all in these situations but only longer gaze durations. However, the oculo-motor system may have a strong drive to keep up the pace of saccades and may, therefore, engage in random backwards-directed eye movements as long as the way forward is blocked (Time-out hypothesis, Mitchell et al., 2008).

It is important to note that the predictions of the constraint-based models have been shown to crucially depend on implementation details, e.g., the criterion ending the competition and assumptions about pre-activation (cf. Green & Mitchell, 2006). Depending on these details, a constraint-based model can predict either a slow-down or a speed-up. Since there is apparently no agreement about these details in the literature, evaluating the predictions of this class of models is difficult. Another difficulty with constrained-based models is that parameter settings may differ across languages. Finding the appropriate parameter settings for the Spanish material studied here would require simulations using corpus data—something that is beyond the scope of this study. In the remainder of this paper, we will assume (following McRae et al., 1998) that constraint-based models predict a slow-down when faced with ambiguity, but the reader should keep this caveat in mind.

Another prominent class of parallel models is based on the idea that an incoming word is more difficult to process when it was unexpected given the previous part of the sentence. This unexpectedness is called surprisal (Hale, 2001). In order to calculate surprisal, a parser has to compute all viable interpretations at a point in the sentence and calculate their probabilities. If an incoming word rules out many high-probability

interpretations, it is assumed to be difficult to process. Surprisal has been demonstrated to make interesting predictions about the processing of garden-path sentences. For example, Hale (2001) showed that surprisal can account for the processing difficulty observed at the disambiguating word in garden-path sentences and Levy (2008) argued that surprisal can explain the ambiguity advantage observed by Traxler et al. (1998). However, for two reasons we will not consider surprisal in the remainder of this article. The first is a technical reason. In order to make predictions, a surprisal-based model has to be trained on a large annotated corpus that informs the model about the relative frequencies of syntactic constructions. In our case, the model would have to be informed, among other things, about the relative frequencies of *si-* vs. *cuando*-clauses and about the frequencies of low- vs. high-attachment of *cuando*-clauses. Unfortunately, a representative corpus of Spanish sentences is not available to us. The other reason why we will not consider surprisal is a theoretical one. The goal of surprisal is not to investigate the mechanics of the parser but rather to clarify the role of certain information-theoretic principles in parsing. Consequently, different parsing mechanisms can be and have been used to calculate surprisal, among them parsers for probabilistic context-free grammars (Hale, 2001) and for dependency grammars (Boston, Hale, Kliegl, Patil, & Vasishth, 2008; Boston, Hale, Vasishth, & Kliegl, 2011), *n*-gram models (Levy, 2008), and recurrent neural networks (Frank, 2009). While this article is concerned with the mechanisms used to build and revise interpretations, surprisal abstracts away from these mechanisms.

Unrestricted race model

Finally, relaxing the assumption of determinism brings us back to the unrestricted race model, which adds a surprising twist to the predictions. The race model assumes a competition in which all possible structures are evolving in parallel. The structure that is finished first is adopted and other unfinished structures are discarded; only the winning parse is pursued. In a majority of cases, the low-attachment structure is likely to win because it will take (on average) less time to complete than high attachment. Sometimes, however, the high-attachment structure would finish faster. The reason for this nondeterminism is that the race is not only influenced by syntactic cues but also by other information, such as discourse and world-knowledge which may sometimes favour high-attachment. Thus, one prediction of the race model is that reanalysis difficulty can occur not only in temporarily ambiguous high-attachment sentences but also in low-attachment sentences when the parser initially adopted high-attachment. This prediction was evaluated in a series of experiments that used material as in (4) (taken from Traxler, 2007).

(4)   a. The writer of the letter . . .

    b. The letter of the writer . . .

    c. The sister of the writer . . .

       . . . that had blonde hair arrived this morning

In (4a), the relative clause is disambiguated to high attachment by *blonde hair*, which can only modify *writer* but not *letter*. Similarly, (4b) is disambiguated to low attachment. However, in (4c) the relative clause can attach high or low—it is globally ambiguous. The prediction of the race model is that the region containing *blonde hair* takes longer to process in the disambiguated conditions because in both (4a and 4b) reanalysis sometimes takes place. In the unambiguous condition, however, the initial

attachment decision is never disconfirmed and hence no reanalysis takes place. Various studies have confirmed the resulting processing advantage of the ambiguous sentence (Traxler et al., 1998; van Gompel, Pickering, & Traxler, 2001; Traxler, 2007; Swets et al., 2008).

The material in (3) gives us another possibility to test the unrestricted race model. Above, we already mentioned that the race model assumes variance in the times it takes the competing structures to be built. This variance has an interesting and surprising consequence: the more structures enter the competition, the faster the race will finish on average. The reason is that a higher number of competitors increases the chance that one of them will finish really fast.[3] For the material in (3), this means that the ambiguous conditions should be processed faster in the pre-verbal region (*cuando los directores*, "when/if the directors") than in condition (c) where the adverbial clause unambiguously attaches low (*si los directores*, "if the directors"). This prediction cannot be tested with the material by Traxler et al. because there the ambiguous region (*that had*) had two possible interpretations in all three conditions, i.e., the number of competitors was always the same.

A second way we can test the race model is by looking into reanalysis effects. Sentences (3a) and (3b) should both show more signs of reanalysis in the disambiguation region (*entraron/entraran*, "came/come") than the analogous material in (3c) (*entraban*, "come"). The logic for this second prediction is the same as for the material in (4): like a globally ambiguous sentence, an unambiguous sentence does not require reanalysis but the ambiguous conditions both require reanalysis occasionally, because the parser can nondeterministically initially adopt either attachment site. Thus, the race model predicts that (3b) should be more difficult to process than (3c). In addition, the high-attachment condition (3a) should be more difficult than the ambiguous low-attachment condition (3b) because the parser adopts low-attachment more often initially, leading to more cases of reanalysis in the high-attachment condition. The race model is the only one among the candidate accounts that predicts reanalysis effects in both temporarily ambiguous conditions (high and low attachment). The predictions of the various accounts are presented schematically in Table 2.

The aim of this study is to investigate the processes that unfold when the parser is garden-pathed and the effects these processes have on eye movements. Specifically, we plan to pin down the function of eye movement patterns triggered by disambiguation by investigating (i) how their occurrence is influenced by first pass operation of the parser and (ii) how both processing stages (first pass and disambiguation) are affected by individual difference in working memory capacity. This will be achieved in two steps: first, we will examine reading times on the pre-verbal region of the adverbial clause to investigate which of the discussed parsing strategies readers may use for initial structure-building. The second part of the analysis focuses on the scanpath phenomena that occur after the verb in the adverbial clause (i.e., the disambiguating material in the ambiguous conditions) is read. We will investigate the circumstances under which the various types of scanpath phenomena occur and how they are influenced by the syntactic structure, the reading time during the first pass over the pre-verbal region, and the working memory capacity of participants. Seeing how the scanpath categories are modulated by these factors should provide a better understanding of their function.

---

[3]Two key assumptions necessary for this to work are that (i) competing structures do not share processing resources and (ii) that the construction times of the competing structures are largely uncorrelated.

TABLE 2
Summary of predictions of the various models for the critical regions in (3). $X < Y$ means ''X is easier to process than Y'', and $X = Y$ means ''X and Y are predicted to show no difference in processing difficulty''

Region	Garden-path	Construal	Good-enough	Parallel	Race
Pre-verbal *los directores* "the directors"	$a = b = c$	$a = b = c$	$a = b < c$	$a = b > c$	$a = b < c$
Verb *entrar*, "come"	$a > b = c$	$a > b = c$	$a > b = c$	$a > b = c$	$a > b > c$

Note: the predictions given for parallel models do not apply to all parallel models. See the main text for details.

MATERIALS AND METHODS

Participants

The experiment was conducted in Berlin. Native speakers of Spanish were recruited via Facebook but were not acquainted with anybody involved in the study. We posted an ad and asked Spanish native speakers to share it with their friends who shared it with their friends in turn. All 70 participants had normal or corrected-to-normal vision, and reported no reading or other language-related disorders. They received €25 for their participation. Most participants had higher education in fields such as art, music, film, and so on. Their mean age was 30 years ($SD = 5.2$), 44 were female. Their native countries comprised: Argentina, Colombia, Mexico, Spain, and Venezuela. All participants spoke other languages in addition to Spanish.

Assessment of working memory capacity

Although variants of the reading span task by Daneman and Carpenter (1980) have been used in many psycholinguistic studies, it is possible that the measure indexes reading experience rather than working memory capacity (MacDonald & Christiansen, 2002; Wells, Christiansen, Race, Acheson, & MacDonald, 2009). An alternative to the reading span task is the operation span test (Conway et al., 2005) in which the distracter task consists of solving simple equations instead of reading sentences. Hence, an operation span test has a reduced task overlap with our main experiment and correlations between capacity and reading performance can more safely be attributed to working memory. Does the operation span task measure a resource relevant to reading? Kane et al. (2004) argue yes, because they found that a large part of the variance in how people perform in domain-specific span tasks can be explained by a domain-general factor.

In order to measure operation span we developed a software that automatises the test. The procedure follows the recommendations given in Conway et al. (2005). There were two tasks: In the first, the memory task, participants had to memorise words for later recall. The second is a distracter task that is intended to prevent participants from actively rehearsing the memory items. The distracter task consisted of checking the correctness of simple equations, e.g., $(5–3) \times 3 = 8$.

In a practice phase, participants had to judge the correctness of eight equations. The response was given with a keyboard that had two labelled keys. After each response, participants received feedback. During this practice, the reaction time was measured for each equation. The average reaction time plus two standard deviations

were used as a time-out at later stages. The time-out was intended to prevent participants from using extra time, after checking the equations, for rehearsing the memory items. Calculating a time-out for every participant separately allows participants to work at their own pace. People who are slow at checking equations will not be penalised and those who are fast do not have time left that they can use to rehearse. In a second practice phase, participants were exposed to the full task: checking equations and memorising words that were shown between the equations. The words were shown for 1,000 ms and came from a set of two-syllable Spanish content words. English tests typically use mono-syllabic words, but there are too few such words in Spanish. We took words from a norming study that compiled high-frequency words that were selected to be psychometrically equivalent (Bishop, 2009). When there was a typo in a word entered by a participant it was still counted as correctly recalled if the typo was only a one-character difference. Some participants systematically did not use accents where the Spanish orthography prescribes them, so this tolerance to small typos was useful to correct for that. The order in which the recalled words were entered did not matter. There were four practice trials in the second training phase: two that presented two memory items and two that consisted of four items. In the main test, set sizes from two to five were presented and there were four sets for each set size resulting in 16 overall trials. Sets were presented in random order and no feedback was given about the correctness of the judgements of equations or the number of correctly recalled items. In all parts of the test participants had to read the equations and memory items aloud. This was intended to prevent vocal rehearsal strategies. After the test was finished we calculated partial-credit unit scores as recommended by Conway et al. (2005). This score indicates the mean proportion of correctly recalled items within the sets.

Experimental sentences

The 72 experimental items (see Appendix 2 for examples) presented in the eye-tracking part of the experiment followed the grammatical structure of the sentences in (3). (The full set of stimuli is available on request from the first author.) In order to construct natural Spanish sentences, we wrote a script that used Google to find sentences with the approximate sentence structure. These sentences were modified to fit into the desired grammatical scheme. Most sentences came from Spanish newspapers and had political content. We avoided material that was funny or disturbing. The regions of the sentences had varying numbers of words across items but we tried to keep them similarly long. The experimental sentences had a mean length of 18.5 words $(SD = 1.6)$. After the disambiguating word, there were on average 4.9 more words $(SD = 1.1)$ until the end of the sentence. Informants from Costa Rica, Argentina, and Spain helped to formulate the sentences such that they could be understood by speakers of different varieties of Spanish equally well.

Using a similar procedure, 72 filler sentences (available on request) were specifically designed to disguise the key contrasts of the experimental sentences. They featured similar lexical material, grammatical structure, and content.

Comprehension questions

We designed yes/no comprehension questions for all sentences, experimental, and fillers. The main purpose of these questions was to motivate participants to deeply analyse the sentences particularly with respect to the attachment of the adverbial clause in the experimental sentences. This was achieved with questions whose answer

depended on the attachment site of the adverbial clause. When the adverbial clause is re-attached in the course of reanalysis, what changes semantically is the temporal order of events. In (3a), the directors enter the room at the time when the teacher is giving the command, whereas in (3b) and (3c) the students are supposed to stand up when the directors enter. There were four types of questions which differed in two respects: (i) half used *before* and half *after* to probe the order of events, (ii) half related to the event in the main clause and the other half to the event in the embedded clause. Here is an example question in which *before* is used and where the adverbial clause is related to the event in the main clause:

(5) ¿Los directivos interrumpieron la clase antes de que el profesor diera las órdenes a los estudiantes?
"Did the directors interrupt the class before the teacher gave the command to the students?"

If this kind of question is interpreted strictly, the answer is not clear in the low-attachment conditions. The reason is that the reader does not know whether the directors entered the room at all. In order to see how Spanish comprehenders interpret this kind of question, we conducted a questionnaire pretest in which we gave lists with 24 sentences each to 12 participants (who did not participate in the later eye-tracking experiment). Of these sentences 18 were from the low-attachment condition and 6 from the high-attachment condition. This way, every low-attachment sentence was judged by three participants, and every high-attachment sentence was seen by one participant. The results suggested that participants answered the questions assuming that the event described in the adverbial clause did take place: in the low-attachment condition, 81% of the questions were answered as desired, compared to 82% in the high-attachment condition. In the main experiment, participants using a strict interpretation strategy should be easily identifiable by good performance in the high-attachment condition vs. poor performance in the other conditions.

Apparatus

For recording eye movements an EyeLink 1000 system by SR Research Ltd. was used, which was recording at 500 Hz in desktop configuration: the camera was located below the screen and participants placed their heads on a chin rest that was adjusted to allow a comfortable position. All materials were presented on a NEC Multisync 2080UX screen at a resolution of 1600×1200 pixels. The distance between the eyes and the camera was 60 cm and the distance between eyes and screen 70 cm. Viewing was binocular but only the right eye was recorded.

Procedure

First, the participants got an overview of the procedure. The first part of a session consisted of the operation span test described above and took about 15 minutes. In the second part, the eye-tracking experiment was conducted which took from 50 to 80 minutes depending on reading speed. Each participant was randomly assigned one of three stimuli lists which comprised different item-condition combinations according to a Latin square. The item order was randomised for each participant individually. At the start of the experiment, the experimenter performed the standard EyeLink calibration procedure, which involves looking at a grid of thirteen fixation targets in random succession. Calibration was repeated during the session if the experimenter noticed that measurement accuracy was poor. At the beginning of each trial, a dot

appeared at the left edge of the screen. After participants fixated on this dot, the sentence appeared. All experimental sentences were presented without line breaks. Seven out of 72 filler sentences spanned across two lines. The font was Arial and the height of a capital letter was approximately $0.5°$ in the visual field. Text was displayed in black on a white background. Participants had to look at the bottom right corner of the screen to proceed to the comprehension question. The response to the questions was given using a 7-button Microsoft Sidewinder game pad. When a response was given, the next trial started. Before the actual experiment started, there were six practice sentences after which participants could ask clarification questions.

Pre-processing of eye movement data

All data processing and analyses were carried out in GNU-R (R Development Core Team, 2009). Fixations were detected using the algorithm described by Engbert and Kliegl (2003) which is based on the velocity profile of saccades. Trials in which more than 20% of the fixations were outside the sentence were removed. For the scanpath analysis, we removed fixations that were directed at regions outside the sentence. Since the regions of interest did not have the same size in all sentences, it was not possible to compare eye movement patterns recorded for different sentences directly. A fixation at a specific coordinate could be in region 3 in one sentence but in region 4 in another sentence. To address this, we calculated the mean size for all regions of interest and transformed all fixations to the resulting standard sentence layout. If a fixation was in the middle of region 3 in sentence 1, it was placed in the middle of region 3 in the standard sentence, which may be at a slightly different screen location than the original fixation.

Scanpath analysis

The analysis of the eye movement trajectories that ensued after participants read the verb in the adverbial clause relies on the similarity measure (Scasim) for such trajectories that was introduced by von der Malsburg and Vasishth (2011). This measure is an edit-distance especially tailored for eye movements. It evaluates the similarity of a pair of scanpaths as a function of temporal and spatial differences taking into account the high acuity in the fovea and the drop in resolution towards the periphery. A difference to other similarity measures that have been proposed in the literature (Brandt & Stark, 1997; Cristino, Mathôt, Theeuwes, & Gilchrist, 2010; Salvucci & Anderson, 2001) is that this measure does not require that the stimulus be partitioned into discrete regions of interest. Instead it operates on the continuous coordinates of the eye fixations. Similarly, time is also handled in a continuous way. Thus, Scasim is highly sensitive even to small spatial and temporal differences between scanpaths. See von der Malsburg and Vasishth (2011) for the definition of Scasim and an extensive discussion.

A variety of analyses of eye movement patterns can be conducted based on similarity scores (Feusner & Lukoff, 2008; Hacisalihzade, Stark, & Allen, 1992; Josephson & Holmes, 2002). In this study, we use a refined version of the method that already proved useful in von der Malsburg and Vasishth (2011). The goal of this method is to identify categories of scanpaths that are similar with respect to spatial and/or temporal properties. Since the clustering procedure that we use for inferring such categories does not operate on similarity scores but on vector representations of the items to be classified, an intermediate step is necessary that generates such a representation. We call this vector representation a map of scanpath space. On such a

map, every scanpath is represented as a point, and the distances between points on the map reflect the dissimilarities between scanpaths; if two scanpaths are located close to each other on the map, this is because they are similar. Clusters of points thus indicate groups of similar scanpaths.

Once we have a map and a cluster model of the distribution of scanpaths on that map, we can examine how the clusters are affected by the experimental manipulation. If a cluster has more member scanpaths recorded in the garden-path condition, these scanpath patterns may be implicated with reanalysis processes.

RESULTS

The results section consists of three parts. In the first, we analyse the performance of the participants in answering comprehension questions. The second part presents an analysis of reading times for the pre-verbal region of the adverbial clause, and the last part an analysis of the scanpath patterns.

Accuracy in comprehension questions

The accuracy of the answers to comprehension question was analysed because one of the core predictions of good-enough processing is that participants show lower performance in high-attachment sentences than in analogous low-attachment sentences.

Comprehension questions for filler sentences were correctly answered in 85% of the cases but only 65% of answers for questions about experimental sentences were correct. In the high-attachment condition, 58% of the questions were answered correctly. In both low-attachment conditions, 70% of the questions were answered correctly. A linear mixed model was used to test these differences between the conditions (Bates, 2005). The model had accuracy in a trial (1 = correct; 0 = incorrect) as the dependent variable and a binomial link function. Comparisons were carried out using Helmert contrasts: the high-attachment condition (a) was compared with the mean of the two low-attachment conditions (b, c), and the low-attachment conditions to each other. This contrast coding was chosen because we expected that the high-attachment condition (a) would be harder and that the low-attachment conditions (b, c) would elicit similar performance because in both conditions the preferred attachment site was ultimately the correct one. In addition, the model had random intercepts for participants and items. Accuracy was significantly worse in the high-attachment condition than in the low-attachment conditions ($\beta = -0.18$, $SE = 0.021$, $z = -8.3$). There was no significant difference between the two low-attachment conditions ($\beta = -0.035$, $SE = 0.039$, $z = -0.91$).

Analysis of eye-tracking measures

Analysis

We calculated first pass reading times for the pre-verbal region of the adverbial clause. This region covers the larger part of the ambiguous region in the ambiguous conditions (a, b). The first word of the adverbial clause (*cuando/si*) was excluded because this word was shorter in the unambiguous condition than in the ambiguous conditions which affects the reading times. There is still a substantial possibility that *cuando* and *si* have differential spillover effects in the pre-verbal region. Two phenomena may contribute to that. First, short words are skipped more often than longer words and these skips are sometimes followed by regressions back to the

skipped word (Vitu & McConkie, 2000). These regressions may reduce first pass reading times on the region following the skipped word because they terminate the regular first pass prematurely. Applied to our material, this means that *si* is more likely to be skipped than *cuando* and that the first pass reading time for the pre-verbal region may, therefore, be shortened in the unambiguous condition. This speed-up, if it was observed, could not be attributed to syntactic processing alone. The second phenomenon is that fixation durations following skippings have been found to be longer than fixation durations following nonskipping forward saccades (Reichle, Pollatsek, Fisher, & Rayner, 1998, p. 147). Hence, if a skip is not followed by a regression, this may increase the reading time for the pre-verbal region. Whether skipping and regressions in response to skipping increase or decrease reading times overall depends on the relative rates of these events, the respective associated effect sizes, and whether the word *si* or the word *cuando* was skipped. (Skipping of a short word may have different effects than skipping of longer word such as *cuando*.) In order to interpret effects in first pass reading time in terms of syntactic processing we have to exclude the possibility that these effects are due to oculo-motor constraints instead of syntactic differences. We address this by correcting the raw first pass reading times using a regression model that estimates and extracts variance that can be explained by skipping of *si*/*cuando* and by subsequent regressions.

The word *si* was skipped in 73% of the trials and *cuando* was skipped in 10% of the trials. When *si* was skipped there was a subsequent regression in 31% of the cases. Regressions after skipping of *cuando* occurred in 41% of the cases.

The linear regression model used for the correction had the natural logarithm of the first pass reading times for the pre-verbal region as the dependent variable. The log-transform was used because it resulted in residuals that were approximately normally distributed. Predictors were skipping status of the previous region (skipped or not), whether or not a regression occurred on the pre-verbal region, and the interaction of these factors. Only first pass skippings and regressions were considered. Skipping and regressions were coded as treatment contrasts with no skip and no regression as the baseline. The corrected values were calculated by subtracting the estimated effects of skipping, regressions, and their interaction from the original reading times. This way the corrected values reflect the reading times expected for trials without a skip and without a regression. This procedure was applied for trials with *cuando* and *si* separately because the effects of skipping may be different for these two words.

Below, we will only report statistical tests that used the corrected reading times. However, we also calculated all tests with the original reading times and found that the results are qualitatively the same in each test. Some *t*- and *z*-scores for the estimated effects of the various predictors became smaller, some larger.

We used a linear mixed model to analyse the corrected reading times. This mixed model had ambiguity as a fixed effect. In other words, it compared the ambiguous conditions taken together vs. the unambiguous condition. The sentences in the ambiguous conditions are identical up to the following region. Therefore, we did not compare them to each other. In addition, the model had a fixed effect of working memory capacity and random intercepts for participants and items. Random slopes were not added for two reasons: (i) they did not qualitatively change the results and (ii) they led to pathological estimates of the correlation of random intercepts and slopes which indicate that there was insufficient data for estimating random slopes.

RESULTS

The means of the reading times (raw and corrected) can be seen in Table 3. First pass reading times in the pre-verbal region were longer in the unambiguous condition than in the ambiguous conditions ($\beta = 0.034$, $SE = 0.013$, $t = 2.6$). (This difference was larger in the uncorrected reading times showing that effects of low-level oculo-motor constraints in fact did contaminate the raw reading times. However, the remaining difference in the corrected reading times could not be explained by oculo-motor effects of the *cuando/si* difference in the previous region.) There was no main effect of working memory capacity. However, there was a significant interaction of working memory with ambiguity: the difference between ambiguous and unambiguous sentences was larger in low-capacity readers than in high-capacity readers ($\beta = 0.031$, $SE = 0.013$, $t = 2.3$).

Scanpath analysis of regressions starting on the spillover region

There were three goals of the scanpath analysis: the first was to determine which scanpath patterns occurred after the verb in the adverbial clause was read. This was done using a cluster analysis that identified categories of scanpaths that are distinguished by their temporal and spatial properties. The second goal was to investigate under which circumstances these scanpath patterns occur. In order to do this, we analysed the rates of these patterns in the three conditions and how their occurrence was influenced by the first pass reading time for pre-verbal region and by the working memory capacity of the readers.

Identification of scanpath categories

We extracted fixation sequences from trials which had a regression launching either from the verb or the spillover region. The sequences started after the verb region was entered and ended when the eyes returned to the launch site of the first regression. From a total of 4,850 trials, 2,520 contained such a pattern. Scasim scores for the similarity of all pairs of regression scanpaths were computed. Raw Scasim scores were divided by the total duration of the two corresponding scanpaths to avoid trivial effects of scanpath length (see von der Malsburg & Vasishth, 2011, for a discussion of normalisation procedures). These similarity scores were used to fit a map of scanpaths using the function isoMDS from the package MASS (Kruskal, 1964a, 1964b). On this map, similar scanpaths are located close to each other and dissimilar scanpaths are far away from each other. A cluster analysis of this map was carried out by applying

TABLE 3
Mean first pass reading times for the pre-verbal region of the adverbial clause (in milliseconds). Raw scores are given along with corrected scores. The overall pattern is the same in raw and corrected scores: reading times were longer in the unambiguous condition. In addition, the effect of ambiguity was larger in low-capacity readers. However, the differences are smaller in the corrected scores

	Raw		Corrected	
	Ambiguous	*Unambiguous*	*Ambiguous*	*Unambiguous*
High capacity	458.31	501.31	498.12	496.15
Low capacity	430.92	519.52	463.17	498.74
All	444.33	510.65	480.28	497.48

mixture of Gaussian modelling (Fraley & Raftery, 2002, 2007). The purpose of the cluster analysis was to identify categories of scanpaths that exhibit qualitatively different gaze trajectories. All parameters of the clusters (position, variance, rotation) were allowed to vary freely. For a specific number of clusters, these parameters were estimated using expectation maximisation. Cluster models were fit for numbers of clusters ranging from 2 to 30. The best model was selected using a Bayesian information criterion and consisted of 15 clusters.

The number of dimensions for the map of scanpaths was determined by fitting maps for increasingly many dimensions (from 2 to 29). For each of these maps, we calculated a cluster model. The faithfulness of a map can be quantified using a residual sum of squares, called stress, that indicates how much variance in similarity scores is not explained by the distances between scanpaths on the map. Figure 1 shows the stress values and the sizes of cluster models as a function of the number of dimensions. Maps with a low dimensionality have high stress, i.e., they are not very faithful. Maps with many dimensions, however, have small cluster models because the Bayesian information criterion used for model selection penalises the estimation of large numbers of parameters (each extra dimension requires the estimation of additional parameters). For this analysis, we selected 7 dimensions (stress: 6) because at this number the size of the cluster model peaks.

Figure 2 shows the prototypical regression patterns for each cluster, i.e., the regression patterns that maximise the similarities to all other cluster members. There are three broad categories of regression patterns. We see a set of eye movement patterns that involve briefly going back to the disambiguation region before proceeding to the end of the sentence (panels 1, 5, 6, 9, 12). We will call this "checking", following the terminology used by von der Malsburg and Vasishth (2011), where this pattern was also found. Another category of scanpaths that was found in our previous study consists of variations of a rereading strategy where the eyes go back to the beginning of the sentence and reread the material so far (panels 4, 11, 13, 15). In a third category, the eyes rapidly regress from the verb region to the pre-verbal region (panels 2, 3, 7, 8, 10, 14). Since the distinctions within these categories are not relevant for the main theoretical questions being asked here, we combine the 15 clusters into three super clusters. See Figure 3 for representative examples of these three categories. The map of scanpaths and the location of the three super clusters can
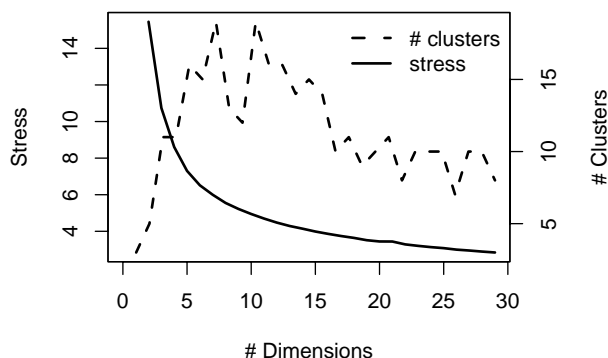
Figure 1. Stress values and numbers of clusters for increasing numbers of map dimensions. As the number of dimensions increases, the stress of maps decreases; more variance is explained by higher-dimensional maps. The number of clusters detected on those maps peaks at seven dimensions.
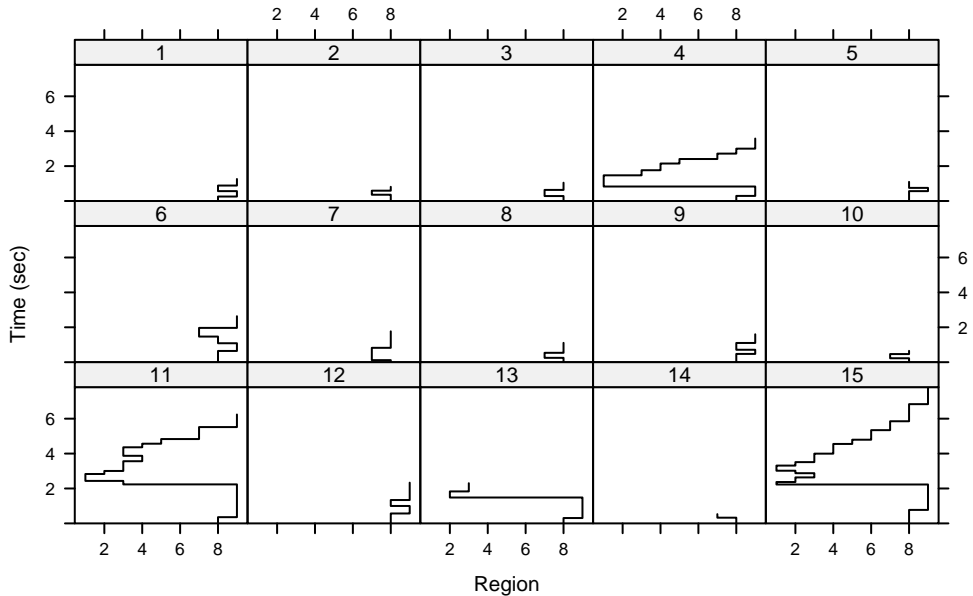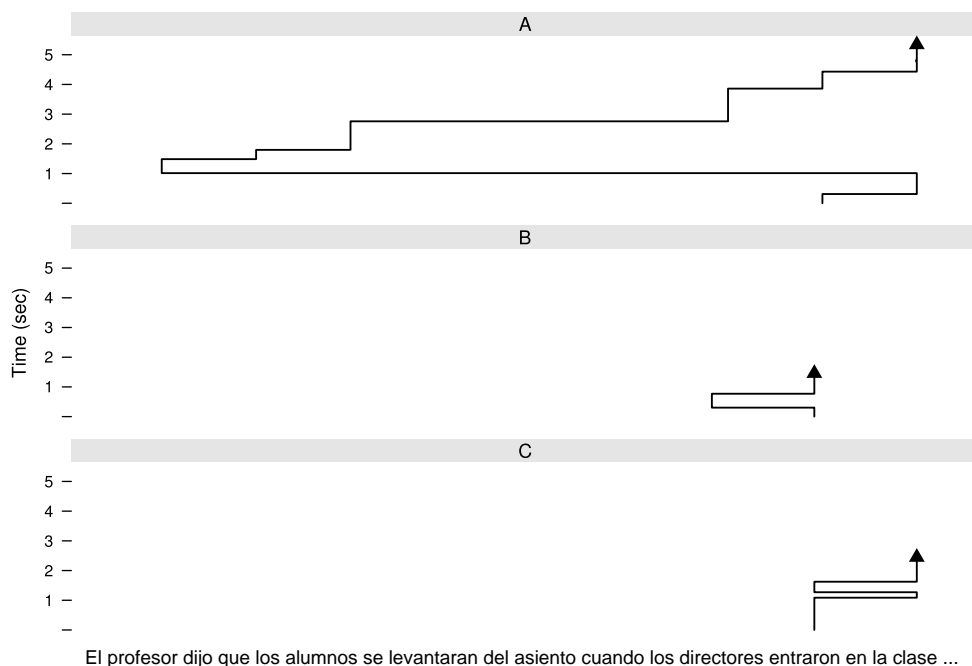
Figure 2. Prototypical scanpaths for the 15 clusters detected on the 7-dimensional map of scanpaths.

be seen in Figure 4. Figure 5 shows the rates at which scanpaths from these clusters occurred in the three conditions.

Analysis of the rates of occurrence of scanpath patterns

Three linear mixed models were used to check under which circumstances these patterns occurred. These models had the presence of a pattern in a trial as a binary dependent variable and the following predictors: condition, the corrected first pass reading time for the pre-verbal region in the same trial, working memory score of the participant, and all interactions of these variables. Interactions were dropped from the model when they were not significant. A main effect term for a variable was dropped when neither the main effect was significant nor any interaction in which this variable was involved. The contrast coding for condition was a Helmert contrast for all models, but the conditions were grouped differently depending on the hypothesis that was tested (see below for details). First pass reading times and working memory capacity were centred and scaled to have a standard deviation of 1. The models had random intercepts for participants and items. For every model, we checked if adding random slopes changed the results qualitatively, which was never the case. All estimates can be found in Appendix 1.

Model 1 tested the occurrence of the rereading pattern. We compared the high-attachment condition with the low-attachment conditions and the low-attachment conditions to each other. This contrast was chosen because our earlier study suggested that rereading occurs more often in the high-attachment condition. A marginally significant effect of attachment (high vs. both low conditions) shows that this was also the case here ($\beta = 0.046$, $SE = 0.026$, $z = 1.7$). There were no main effects of working memory and of first pass reading time on the pre-verbal region. However, both factors interacted with attachment site: if more time was spend reading the pre-verbal region, the difference between high- and low-attachment sentences was larger, meaning that there was considerably more rereading in the high-attachment condition than in the

El profesor dijo que los alumnos se levantaran del asiento cuando los directores entraron en la clase ...

Figure 3. Prototypes of the three super clusters: rereading of the preceding material (A), rapidly triggered regressions from the verb to the pre-verbal region (B), checking regressions from the spillover region to the verb region (C).

low-attachment conditions ($\beta = 0.082$, $SE = 0.027$, $z = 3.1$). See Figure 6 for a graph showing the effect of first pass reading time of the pre-verbal region. A marginally significant interaction of working memory capacity and attachment site shows that participants with a high memory capacity reread more often in the high-attachment condition than those with low capacity ($\beta = 0.045$, $SE = 0.027$, $z = 1.7$). See Figure 7 for a graph of the effect of working memory.

Model 2 tested influences on the rate of the checking pattern (regressions back to the verb of the adverbial clause). This time we used a contrast coding that compared the ambiguous conditions with the unambiguous condition and then the ambiguous conditions to each other. This follows from our hypothesis that this pattern reflects diagnosis procedures that are primarily used in ambiguous sentences. The model shows that checking regressions occurred significantly more often in the temporarily ambiguous conditions ($\beta = 0.13$, $SE = 0.032$, $z = 4$). There was no difference between the two ambiguous conditions.

Model 3 tested rapid regressions from the verb. Since this pattern looks similar to the checking pattern, we used the same contrasts. Rapid regressions were more frequent in the unambiguous condition ($\beta = -0.06$, $SE = 0.031$, $z = -2$). An interaction of working memory and ambiguity shows that this difference was larger in high-capacity participants ($\beta = -0.062$, $SE = 0.032$, $z = -2$). Furthermore, there was a strong main effect of first pass reading time in the pre-verbal region: if the eyes spent a short time on that region, the probability of rapid regression was increased ($\beta = -0.15$, $SE = 0.05$, $z = -3$). However, this effect was driven by only a few trials in which the pre-verbal region was initially skipped.

A final, redundant, mixed model was used to investigate the proportion of trials in which no regression at all occurred after the verb region was read. This model is
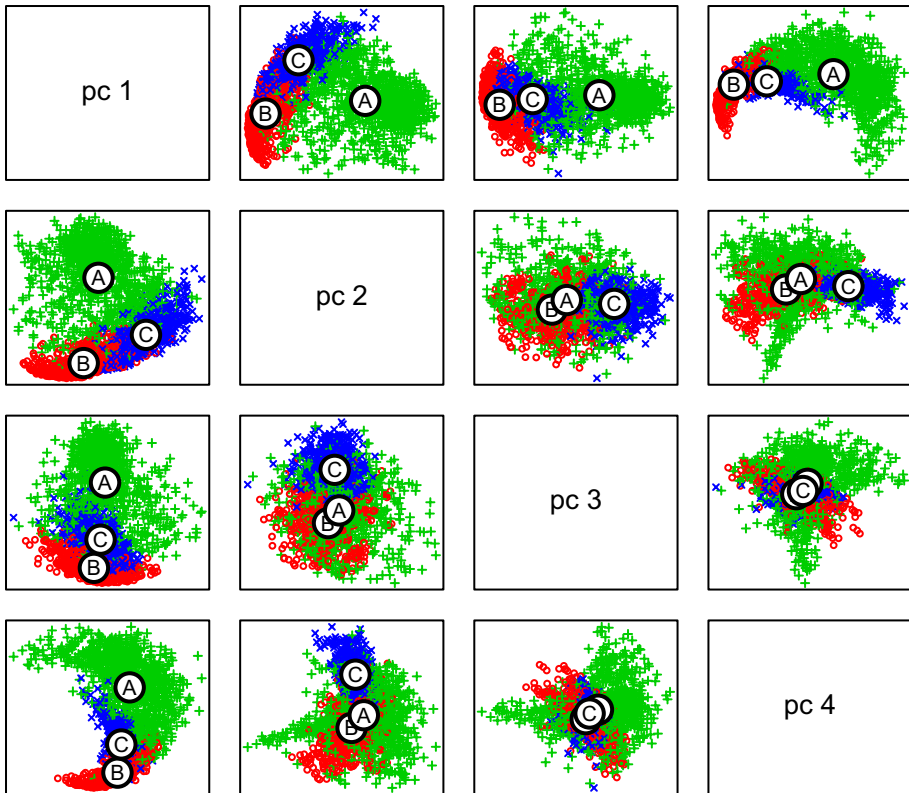
Figure 4. Projections of the 7-dimensional map of scanpaths and the location of the three super clusters (best viewed in colour): cluster A contains rereading scanpaths, cluster B rapid regressions from the verb, and cluster C checking regressions from the spillover region to the verb. Only the first four principal components of the map were used for this plot.

redundant because the outcome is predictable given the above models. We used a contrast coding comparing high vs. low attachment and the low-attachment conditions to each other. In all three conditions, regressions were often absent. However, trials without regressions occurred significantly more often in the low-attachment conditions than in the high-attachment condition ($\beta = -0.059$, $SE = 0.021$, $z = -2.7$). When the first pass reading time on the pre-verbal region was long, it was more likely that no regression occurred ($\beta = 0.077$, $SE = 0.036$, $z = 2.1$).

DISCUSSION

We will first briefly summarise our main results and then discuss how they speak to the questions asked in this study.

The accuracy in the comprehension questions was poor (but above chance) in the high-attachment condition (58%) and fair in the two low-attachment conditions (70%). The pre-verbal region was read faster when it was ambiguous and this effect was larger in readers with a low working memory score. The scanpath analysis found three types of regression patterns that ensued after the verb in the adverbial clause was read: rereading starting from the spillover region, rapid regressions starting from the verb region, regressions back from the spillover region to the verb region (checking).
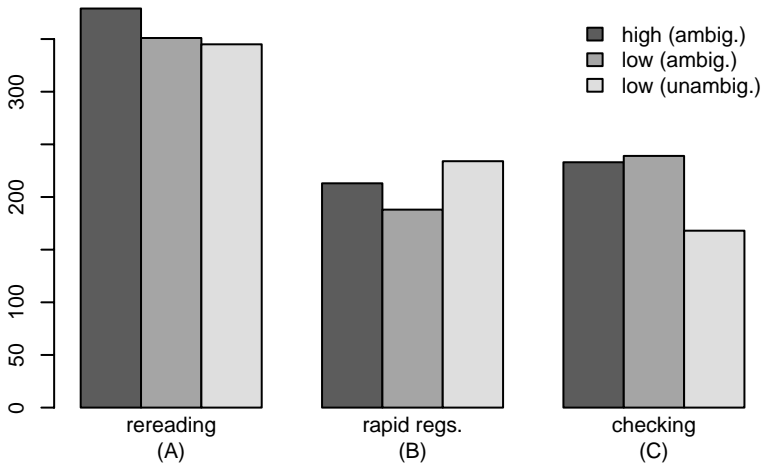
Figure 5. Occurrences of the three scanpath patterns in the three conditions: rereading (A), rapid regressions from the verb (B), checking (regressions from the spillover region to the verb, C).

Rereading occurred more often in the high-attachment condition and this effect was larger in readers with high working memory capacity (marginally significant). When the eyes spent more time on the pre-verbal region, rereading occurred more often in the high-attachment condition than in the low-attachment conditions. The checking pattern was more frequent in the ambiguous conditions and rapid regressions from the verb of the adverbial clause were more frequent in the unambiguous condition. The latter effect was larger in high-capacity readers and when more time was spent on the pre-verbal region.

In our earlier scanpath analysis of the Meseguer et al. data, we also found a regression pattern in which the eyes returned briefly to the first region before the trial stopped. We speculated that this pattern did not reflect language processing but rather
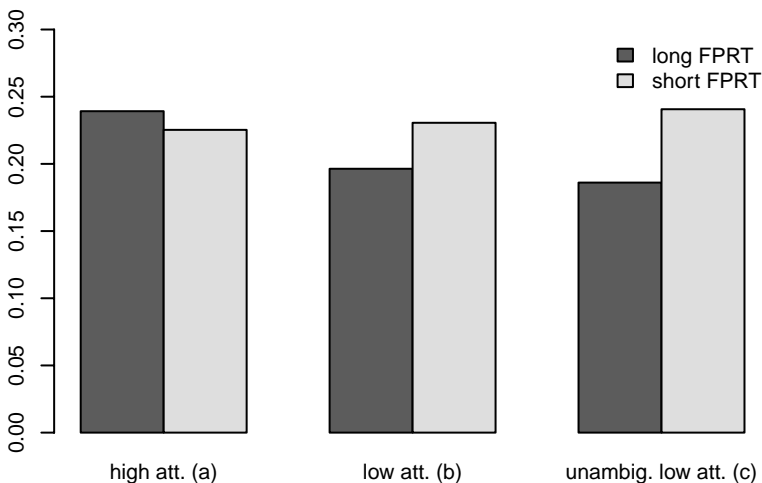
Figure 6. The effect of first pass reading time of the pre-verbal region of the adverbial clause on the rate of rereading in the three conditions. When the pre-verbal region was inspected only for a short time, there was no difference between the conditions suggesting that the parser did not process the attachment. If, however, the pre-verbal region was inspected longer, there was a clear difference between the conditions: rereading occurred more often in the high-attachment condition, which required reanalysis, and less often in conditions that did not require reanalysis.
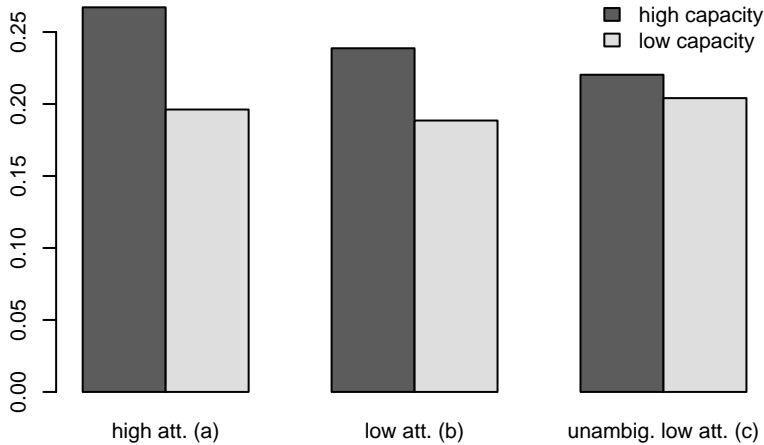
Figure 7. The effect of working memory capacity of participants on the rate of rereading in the three conditions. When capacity was low, there was no difference between the conditions. However, when capacity was high, there was an increased rate of rereading in the high-attachment condition compared to the low-attachment conditions suggesting that high-capacity participants processed the attachment more thoroughly. Note that the differences between the two low-attachment conditions were not significant.

anticipation of the next trial: the procedure used by Meseguer et al. did not require participants to look at a corner of the screen when they were finished reading. In our experiment, participants had to look at a corner of the screen before proceeding to the comprehension question. As expected, this scanpath pattern did not occur anymore. This confirms that it was not related to the processing of the stimulus.

Another change in procedure was that we presented the sentences on one line (Meseguer et al. had a line break in the middle of the embedded clause). This change did not influence the types of patterns that were detected.

The main question in this study was: What strategies do readers use to make provisional parsing decisions in temporally ambiguous sentences and how do they recover when upcoming material indicates that these decisions were wrong? We will first discuss the results that allow us to draw conclusions about initial structure building and then turn to reanalysis processes. Finally, we will discuss consequences for theories about the interaction of parsing and oculo-motor control.

First pass attachment decision

The theories of parsing discussed in the introduction are most clearly distinguished by their predictions for the pre-verbal region. (See Table 2 for a summary of the predictions.) Trivially, there are no differences expected on that region between the two ambiguous conditions which are identical up to the verb of the adverbial clause, but for the comparison of ambiguous vs. the unambiguous sentences we have all possible predictions: garden-path theory and construal predict no difference, parallel models predict that the ambiguous sentences require more effort, and the race model and the good-enough account predict that the unambiguous sentences are harder to process. Our results for this region are relatively easy to interpret. The longer reading times in the unambiguous condition suggest that processing of the pre-verbal region was harder in the unambiguous condition than in the two ambiguous conditions. More support for the idea that the pre-verbal region was harder to process when the sentence was unambiguous comes from the scanpath analysis: in the unambiguous condition,

rapid regressions back to the pre-verbal region occurred more often than in the ambiguous conditions, suggesting that this region was hard to process: Rapid regressions can be interpreted as a consequence of premature forward saccades. In the unambiguous condition, the eyes may often have moved on before the parser had finished processing of the pre-verbal region. In this situation, regressions could give the parser time to finish processing before moving on to new material (cf. Mitchell et al., 2008). This interpretation of rapid regressions is supported by the fact that these regressions occurred more often when the first pass reading time for the pre-verbal region was short.

When interpreting the slowed reading times in the pre-verbal region of the unambiguous sentences it is important to keep in mind a fact about the experimental material: the conditional clause starting with *si* ("if") cannot be attached to the main verb *dijo* ("said") because there is a constraint on the tense of the modified verb: verbs with episodic tense *pretérito perfecto simple* are not eligible for modification with a *si*-clause. A similar constraint does not exist for adverbial clauses starting with *cuando*. Hence, the slowdown in the pre-verbal region of the unambiguous sentences could in part be caused by checking of the tense feature of the main verb. However, the observed difference (the ambiguity advantage) is established in the literature and has been reported by several studies (Traxler et al., 1998; Swets et al., 2008; van Gompel et al., 2001; van Gompel, Pickering, Pearson, & Liversedge, 2005; Traxler, 2007). For this reason, we will assume that checking of the tense feature is not the sole cause of the effects in the pre-verbal region.

The race model predicts the observed speed-up in the ambiguous conditions because it assumes that both structures are built in parallel and whatever structure is finished first is adopted. The interpretation of good-enough processing adopted in this paper predicts the speed-up because it assumes that the parser, when faced with an ambiguity, may simply refrain from making an attachment. By contrast, construal, the other underspecification account considered here, does not make an attachment decision depending on whether or not there is an ambiguity. If the material in question cannot be interpreted as an argument or complement, it is associated with the last theta domain (the projection of *se levantaran*, "stand up"). Consequently, the *si*- and the *cuando*-clauses are both associated with the last theta assigner, and no difference in processing load is expected.[4] Similarly, garden-path theory assumes that in all three conditions the same process takes place (the adverbial clause is attached low), which should again result in equal processing costs. At least some parallel models predict that the ambiguous material should be harder to process. Our results are, therefore, not easily explained by those parallel models, construal, and garden-path theory. These results are also difficult to reconcile with the version of good-enough parsing adopted by Christianson et al. (2001) because that account basically assumes the same processing strategy as garden-path theory. We thus focus on the race model and the Swets et al. (2008) interpretation of good-enough processing as candidates for explaining first pass parsing decisions.

---

[4]Swets et al. (2008) interpret the construal hypothesis differently. Their interpretation, under which the ambiguous conditions should be processed faster than the unambiguous condition, is consistent with our data. However, this interpretation requires additional assumptions which are not part of the original definition of construal.

Reanalysis

Focusing on the race model and good-enough processing allows us to narrow down our expectations for reanalysis effects (see the second row in Table 2). The race model does not commit to a particular reanalysis mechanism but makes the interesting prediction that reanalysis should occur in the ambiguous high- *and* low-attachment conditions. This predicts a pattern of results where the high-attachment condition is harder to process than the ambiguous low-attachment condition which should in turn be harder than the unambiguous condition. Good-enough parsing, however, predicts reanalysis effects only for the high-attachment condition; the two low-attachment conditions (ambiguous and unambiguous) should be equally difficult. The predicted reanalysis effects, however, are relatively small because, according to this account, the parser often just does not bother with determining the attachment site. If the parser does not commit to an attachment site, there is nothing to revise later in the sentence and hence no reanalysis difficulty. However, if the parser does make a commitment, it is assumed to operate in a garden-path theory-like fashion, which means that high attachment should be consistently harder than low-attachment.

Reanalysis effects are usually measured and discussed in terms of the conventional eye-tracking measures. These measures are reliable indicators for processing effort but they do not provide us easily interpretable answers to the question: what did this effort consist of? In our earlier study, we found preliminary evidence showing that the various types of scanpaths identified by the Scasim analysis reflect mechanisms that serve different functions. If true, this means that a measure like the probability of regressing from a word conflates functionally different types of regressions. Any observed pattern in regression probabilities when comparing experimental conditions may, therefore, be difficult to interpret. Frazier and Rayner (1982) were the first to consult scanpaths in reading in order to determine the nature of reanalysis processes, and our previous study (von der Malsburg & Vasishth, 2011) demonstrated how scanpaths can be formally evaluated to arrive at new answers for questions about reanalysis. In our earlier scanpath analysis of the Meseguer et al. (2002) data, we found evidence that rereading constitutes a recovery strategy for garden-path sentences of the type studied here. In the present work, we found further support for this idea because rereading again was a frequent strategy that occurred more often in the high-attachment condition. Because the eye movements examined here started from words preceding the final word of the sentence, this replication of our previous result rules out the alternative explanation that rereading just reflects a second pass over the sentence (Frazier & Rayner, 1982), or that it is triggered by sentence wrap-up effects (Mitchell et al., 2008).

Another eye movement pattern that we found in von der Malsburg and Vasishth (2011) is revisiting the verb region of the adverbial clause (checking). This region contains the disambiguation material in the ambiguous conditions and the difference between the conditions consisted of just one letter in that region (*entraron/entraran*). Therefore, we speculated that this pattern reflects a diagnosis procedure or a kind of sanity check that is executed when the verb of the adverbial clause can potentially resolve an ambiguity. In other words, this type of regression would serve an additional intake of visual information and to increase the confidence about what has been seen earlier (cf. Bicknell & Levy, 2010, who propose a model of eye movement control that explains regressions in terms of uncertainty about the identity of words). One prediction of this idea is that the checking pattern should occur less often in the unambiguous condition because there the verb of the adverbial clause is not informative about attachment. This difference is exactly what we found here.

Moreover, the checking pattern occurred equally often in the two ambiguous conditions, something that we would not expect if this pattern was reflective of reanalysis mechanisms.

The presence of these two functionally different regression patterns, which both started on the same region (the spillover region), shows that the concern about the interpretability of aggregated regressions measures is in fact warranted. Regression probabilities would show a difference between the ambiguous high- and low-attachment conditions (due to the rereading pattern) *and* a difference between the two low-attachment conditions (due to the checking pattern). Such a pattern of results could very well be interpreted as supporting the race model of parsing which is the only model that predicts such a difference between all three conditions. However, looking at the scanpath results, which suggest that only rereading is indicative of garden-pathing, leads us to different conclusions.

So which model *is* supported by the scanpath results? A whole array of effects favours the idea that the parser sometimes leaves the attachment site unspecified, as assumed by the good-enough account. One of these effects is that if the pre-verbal region was read slowly there was a higher rate of rereading in the high-attachment conditions than in the low-attachment conditions. This is expected when we assume that attaching takes additional time: if the parser attaches, which takes time, it has to revise this decision later in the sentence if the verb of the adverbial clause indicates the dispreferred high-attachment; this leads to increased rereading. If, however, the parser does not attach the adverbial clause, there should be no difference between the conditions with respect to rereading because no revision is required then at the disambiguating region.

In reading research, shifts in reading strategies have been observed where older readers, for instance, produce longer fixation durations and more regressions (e.g., Kliegl, Grabner, Rolfs, & Engbert, 2004). Can such correlations of reading times and regressions explain our effect of first pass reading time on the rate of rereading? Probably not, because these differences in general reading strategies would show up as a main effect of reading time, which was not found. Instead, we found an interaction of reading time with attachment site, which suggests that the reading time on the pre-verbal region is related to garden-path processing specifically.

Effects of working memory capacity point in the same direction: high-capacity readers reread more often than low-capacity readers in the high-attachment condition (marginally significant). This suggests that they made attachment decisions more often, which leads to garden-pathing in cases where the decision was wrong, i.e., in the high-attachment condition. A low-capacity reader, however, when he does not attach, should not stumble on the disambiguating material because it cannot possibly conflict with earlier parsing decisions. Consistent with that, there were only small differences between the conditions with respect to rereading in the low-capacity readers. In the analysis of first pass reading times on the pre-verbal region, we also found that low-capacity readers had a larger effect of ambiguity: they had a greater speed-up in the ambiguous condition than the high-capacity readers, which supports the idea that they underspecified the attachment site more often.

Surprisingly, high-capacity readers appear to have more processing difficulty than low-capacity readers. The above results suggest that the explanation for this counter-intuitive effect is that high-capacity readers build dependencies in situations where low-capacity readers do not. At this point, we could ask: isn't it a rather trivial fact that people sometimes do not care to identify the meaning of a sentence? Perhaps the low-capacity readers in our study just did not sleep well in the night before the

experiment? We argue that the situation is not that simple. If the low-capacity readers just were less willing to perform well, we would expect strong main effects of working memory capacity. For example, we may expect generally slowed reading and worse performance in the comprehension questions. However, no such main effects were found. We only found interactions of working memory capacity with attachment site and ambiguity. This suggests that low-capacity readers used a rather sophisticated, grammar-sensitive underspecification strategy in order to preserve memory resources. They selectively adapted their processing to their more limited resources.[5]

A final piece of evidence for underspecification is delivered by the accuracy in the comprehension questions, which confirms the signature prediction of good-enough processing: a low performance in comprehension questions about sentences in which the dispreferred attachment site was ultimately the correct one (Christianson et al., 2001; Swets et al., 2008).

Do these results generalise to other types of ambiguities? Probably not, because underspecification is not a viable strategy for all kinds of structural ambiguities. The attachment of an adverbial clause can be left unspecified without major consequences for the processing of the remaining sentence. However, if the ambiguity affects a primary relation such as in subject/object ambiguities (e.g., Example 1), a decision must be made because structure building cannot proceed otherwise. Likewise, reanalysis is urgently needed when incoming material indicates that a primary relation has been misinterpreted whereas the reanalysis of an attachment may be deferred. It appears that attachment decisions are less constraining and therefore provide more strategic wiggle room that the parser can use to adapt to task demands and available resources. This idea is also supported by a study by Kemper, Crow, and Kemtes (2004) who compared high- and low-capacity readers' performance on garden-path sentences with a main clause/relative clause ambiguity such as in example (6).

(6)   The experienced soldiers warned about the dangers conducted the midnight raid.

In this sentence, it is initially unclear if the word *warned* is the main verb or the verb of a reduced relative clause. Only at *conducted* is the sentence disambiguated towards the dispreferred reduced relative clause interpretation, leading to garden-path effects. Kemper et al. (2004) compared eye-tracking measures for this sentence type with those of three other types which differed in whether or not the critical verb belonged to a relative clause and in whether or not there was an ambiguity. No ambiguity advantage

---

[5]As an aside, recall that the suggestion by Lewis (1998) that rereading is a memory-preserving strategy implies that we would expect low working memory capacity participants to show more rereading. We found the opposite pattern. Note, however, that Lewis was assuming a parser that always fully attempts to build structure; once the assumptions of good-enough parsing are taken into account, our findings can be reconciled with Lewis' memory-preserving claim for rereading: only the high working memory capacity participants try to engage in structure building, and therefore make more mistakes, triggering more rereading, compared to the low-capacity participants, who underspecify and, therefore, make fewer mistakes, leading to less rereading.

was found, suggesting that readers always built structure no matter whether the sentence was ambiguous or not. However, at the disambiguating region, low-capacity readers were slower and performed more regressions than high-capacity readers. These results show that if the sentence material requires a commitment to a structural interpretation, low-capacity readers exhibit the expected greater processing difficulty compared to high-capacity readers. Depending on the type of ambiguity, the effects of working memory capacity can apparently be in one direction or the exact opposite direction.[6]

Taken together these results suggest that good-enough processing is not just a theoretically uninteresting general sloppiness of language comprehenders. It rather appears that it may consist of a dynamic adaptation to continually fluctuating processing constraints. Situations where the language system cannot afford an exhaustive analysis of the sentence and has to cut corners may even be the norm and not just an occasional deviation from "normal" processing. If true, this would call for a theory of sentence processing in which dynamic adaptation to processing constraints is a central theme and not just an add-on. Over 30 years ago, resource limitations and the resulting need of the parser to cut corners were one important motivation for garden path theory. Perhaps these ideas should be developed further.

Alternative reanalysis strategies

Having established that rereading served as a reanalysis strategy for material as that studied here, one obvious questions is: why does the parser employ rereading for garden-path recovery instead of targeted regressions as predicted by selective reanalysis? After all, rereading is relatively time-consuming and might, therefore, seem inefficient. For the present material, repair of the interpretation consists of finding an alternative attachment site for the adverbial clause. If the parser, due to memory limitations, does not keep track of all nonlocal alternatives during the first pass, it has to resort to searching. Rereading from left to right seems like an efficient way to perform that search. Targeted regressions, however, are not viable when the location of the target is unknown. Compare the material studied here with the material used by Frazier and Rayner (1982):

(7)
 a. While Mary was mending the clock it started to chime
 b. While Mary was mending the clock started to chime.

In (7b), *the clock* can initially be interpreted as the object of *mending* but the verb *started* invalidates this analysis. Revision changes the status of *the clock* to being the subject of the main clause. The earliest material affected by the revision is the verb *mending*. Hence, the changes are relatively local and affect recent material. This lets targeted regressions to relevant material appear more viable and efficient than rereading. Thus we can formulate the hypothesis that rereading is used if the reanalysis involves

---

[6]In a related study, MacDonald, Just, and Carpenter (1992) tested sentences as in (6) and found more processing difficulty in high-capacity readers. They explained this result in terms of a model that assumes that high-capacity readers maintain multiple representations of ambiguous sentences, which causes additional effort. Their experiments, however, used the self-paced reading paradigm which constrains the range of possible reading strategies considerably. For example, rereading and re-checking of potentially decisive material are not possible in self-paced reading. Also, there was a considerable difference in the performance on comprehension questions between high- and low-capacity readers: low-capacity participants performed at chance in some conditions, while high-capacity readers performed better than chance. For these reasons the implications for the MacDonald et al. results for our proposal are difficult to determine.

searching for an attachment site for a clause and/or affects nonlocal material. However, we do not claim that rereading is only used for reanalysis—it might also reflect other processes—or that rereading is the only reanalysis strategy for the material used in this study. For instance, we cannot exclude the possibility that readers also used covert reanalysis strategies (Lewis, 1998), which involve no regressive eye movements.

In sum, it seems that the kind of reanalysis strategy that is used (rereading or selective reanalysis), and whether readers deal with the ambiguity at all, depends on the linguistic details of the ambiguity and on available processing resources. Interestingly, some of the results that lead to this conclusion emerged only when we separated functionally different types of regressions using a scanpath analysis. The various regression types would have been conflated in aggregated eye-tracking measures.

Interaction of oculo-motor control and parsing

Finally, we would like to discuss an observation about the timing of regression effects which may have interesting consequences for theories about the coupling of oculo-motor control and parsing. Visual inspection of the scanpaths in Figure 2 suggests that the first regressive saccade was triggered considerably later in the rereading pattern than in the checking pattern. In fact, regressions were launched after 2.9 fixations on average and from 2.4 words after the verb of the adverbial clause in the rereading pattern. Regressions in the checking pattern occurred a bit earlier but still surprisingly late: after 2.5 fixations and 1.8 words (see Figure 8). (We checked these differences using linear mixed models which showed that both were significant.) This
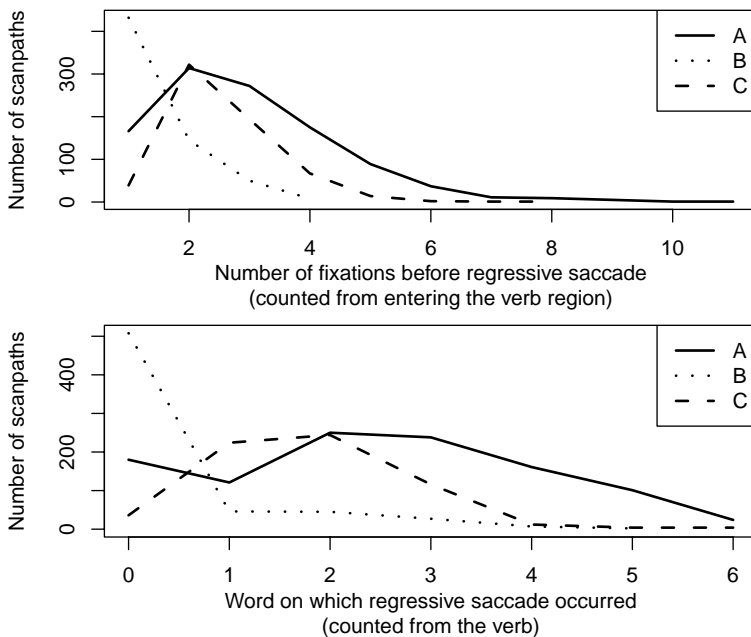
Figure 8. Plots showing when and where regressive saccades were executed after the verb region was entered. The first plot shows after how many fixations the first regression occurred. The second plot shows on which word the first regression was launched (0 is the verb in the adverbial clause). In the pattern that we called "rapid regressions" (B), regressions occurred, by definition, early. In the "checking" pattern (revisiting the verb of the adverbial clause, C), regressions occurred later. Regressions in the rereading pattern (A) had the longest delay.

long delay before effects of the verb appear is surprising given that current theories about the interaction of oculo-motor control and parsing assume relatively rapid effects of sentence processing on eye movements. E-Z Reader 10 (Reichle et al., 2009), for instance, allows the eyes to peek at the next word, but not further ahead, while the current word is being integrated into the sentence context, giving rise to some spillover of processing difficulty. However, if the integration of a word is not completed when lexical access of the next word is finished, the normal forward progression of the eyes is suspended. Similarly, the Bayesian model proposed by Bicknell and Levy (2010), which is specifically targeted at modelling regressions, assumes that regressions in response to a word can happen on the subsequent word at the latest. Thus, these models assume that sentence processing and oculo-motor control are highly synchronised and proceed in lockstep through the sentence.

Why did regressions occur so late in our experiment? The distinction between primary and nonprimary grammatical relations made in the construal theory may again help to explain this effect. According to Frazier and Clifton (1997), primary relations are dealt with rapidly whereas decisions about nonprimary relations can be "made at a more leisurely pace" (p. 46). The rationale for this distinction is that the disambiguation of a primary relation can have consequences for the lexical status of words which potentially ripple through larger parts of the sentence structure. Decisions about nonprimary relations do not have such consequences and can be decided at a later stage, e.g., on a semantic level. This might explain why, in our experiment, the parser did not block the progression of the eyes immediately when a potential structural problem became apparent (at the verb of the adverbial clause). It is conceivable that a process that is only loosely coupled to oculo-motor control determines the attachment site of the adverbial clause while the next words are already being integrated into that clause. More generally, structure building processes dealing with primary relations may interact rapidly with oculo-motor control, whereas decisions about nonprimary relations may be deferred and therefore have late eye movement consequences.[7] Of course, experimental work is needed to test the hypothesis that there are varying degrees of coupling between parser operations and the eyes. Whatever the mechanism behind late regression effects may be, the long delays cannot easily be explained by models of oculo-motor control that assume that the coupling of parsing and eye movements is always tight.

REFERENCES

Bates, D. (2005). Fitting linear mixed models in R. *R News*, *5*(1), 27–31.

Bicknell, K., & Levy, R. (2010). A rational model of eye movement control in reading. In J. Hajič (Ed.), *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics* (pp. 1168–1178). Uppsala, Sweden: Association for Computational Linguistics.

---

[7]This proposal should not be confused with the Time-out hypothesis (Mitchell et al., 2008) which was also described with the term "loose coupling". The Time-out hypothesis assumes that, during regressions triggered by garden-pathing, the parser is largely disconnected from the eyes. Our proposal is that during normal forward operation, processing of information relating to the attachment of the adjunct may lag behind.

Bishop, A. M. (2009). *Psychometrically equivalent bisyllabic word-lists for word recognition testing in Spanish* (Unpublished master's thesis). Brigham Young University, Utah.

Boston, M. F., Hale, J. T., 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), 1–12.

Boston, M. F., Hale, J. T., Vasishth, S., & Kliegl, R. (2011). Parallel processing and sentence comprehension difficulty. *Language and Cognitive Processes*, 26(3), 301–349.

Brandt, S. A., & Stark, L. W. (1997). Spontaneous eye movements during visual imagery reflect the content of the visual scene. *Journal of Cognitive Neuroscience*, 9(1), 27–38.

Carreiras, M., & Clifton, C. (1993). Relative clause interpretation preferences in Spanish and English. *Language and Speech*, 36(4), 353.

Christianson, K., Hollingworth, A., Halliwell, J. F., & Ferreira, F. (2001). Thematic roles assigned along the garden path linger. *Cognitive Psychology*, 42(4), 368–407.

Conway, A. R., Kane, M. J., Bunting, M. F., Hambrick, D. Z., Wilhelm, O., & Engle, R. W. (2005). Working memory span tasks: A methodological review and user's guide. *Psychonomic Bulletin & Review*, 12(5), 769–786.

Cristino, F., Mathôt, S., Theeuwes, J., & Gilchrist, I. D. (2010). ScanMatch: A novel method for comparing fixation sequences. *Behavior Research Methods*, 42(3), 692.

Cuetos, F., & Mitchell, D. C. (1988). Cross-linguistic differences in parsing: Restrictions on the use of the late closure strategy in Spanish. *Cognition*, 30(1), 73–105.

Daneman, M., & Carpenter, P. A. (1980). Individual differences in working memory and reading. *Journal of Verbal Learning & Verbal Behavior*, 19(4), 450–466.

Engbert, R., & Kliegl, R. (2003). Microsaccades uncover the orientation of covert attention. *Vision Research*, 43(9), 1035–1045.

Ferreira, F., Bailey, K. G., & Ferraro, V. (2002). Good-enough representations in language comprehension. *Current Directions in Psychological Science*, 11(1), 11–15.

Ferreira, F., & Patson, N. D. (2007). The 'good enough' approach to language comprehension. *Language and Linguistics Compass*, 1(1–2), 71–83.

Feusner, M., & Lukoff, B. (2008). Testing for statistically significant differences between groups of scan patterns. In K.-J. Räihä & A. T. Duchowski (Eds.), *Proceedings of the 2008 Symposium on Eye-tracking Research & Applications* (pp. 43–46). Savannah, GA: Association for Computing Machinery.

Fraley, C., & Raftery, A. E. (2002). Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association*, 97(458), 611–632.

Fraley, C., & Raftery, A. E. (2007). *MCLUST version 3 for R: Normal mixture modeling and model-based clustering* (Tech. Rep. No. 504). Seattle, WA: Department of Statistics University of Washington.

Frank, S. (2009). Surprisal-based comparison between a symbolic and a connectionist model of sentence processing. In N. Taatgen, & H. van Rijn (Eds.), *Proceedings of the 31st Annual Conference of the Cognitive Science Society* (pp. 1139–1144). Amsterdam, the Netherlands: Cognitive Science Society.

Frazier, L. (1979). On comprehending sentences: Syntactic parsing strategies. *ETD Collection for University of Connecticut*.

Frazier, L., & Clifton, C. (1996). *Construal*. Cambridge, MA: The MIT Press.

Frazier, L., & Clifton, C. (1997). Construal: Overview, motivation, and some new evidence. *Journal of Psycholinguistic Research*, 26(3), 277–295.

Frazier, L., & Fodor, J. D. (1978). The sausage machine: A new two-stage parsing model. *Cognition*, 6(4), 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(2), 178–210.

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

Green, M. J., & Mitchell, D. C. (2006). Absence of real evidence against competition during syntactic ambiguity resolution. *Journal of Memory and Language*, 55(1), 1–17.

Hacisalihzade, S. S., Stark, L. W., & Allen, J. S. (1992). Visual perception and sequences of eye movement fixations: A stochastic modeling approach. *IEEE Transactions on Systems, Man and Cybernetics*, 22(3), 474–481.

Hale, J. (2001). A probabilistic early parser as a psycholinguistic model. In A. Kehler, L. Levin, & D. Marcu (Eds.), *Proceedings of NAACL 2001* (pp. 1–8). Morristown, NJ: Association for Computational Linguistics.

Igoa, J. M., Carreiras, M., & Meseguer, E. (1998). A study on late closure in Spanish: Principle-grounded vs. frequency-based accounts of attachment preferences. *The Quarterly Journal of Experimental Psychology Section A*, 51(3), 561–592.

Josephson, S., & Holmes, M. E. (2002). Attention to repeated images on the world-wide web: Another look at Scanpath theory. *Behavior Research Methods*, *34*(4), 539–548.

Kane, M. J., Hambrick, D. Z., Tuholski, S. W., Wilhelm, O., Payne, T. W., & Engle, R. W. (2004). The generality of working memory capacity: A latent-variable approach to verbal and visuospatial memory span and reasoning. *Journal of Experimental Psychology: General*, *133*(2), 189–217.

Kemper, S., Crow, A., & Kemtes, K. (2004). Eye-fixation patterns of high-and low-span young and older adults: Down the garden path and back again. *Psychology and Aging*, *19*(1), 157–170.

Kliegl, R., Grabner, E., Rolfs, M., & Engbert, R. (2004). Length, frequency, and predictability effects of words on eye movements in reading. *European Journal of Cognitive Psychology*, *16*(1/2), 262–284.

Kruskal, J. (1964a). Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, *29*(1), 1–27.

Kruskal, J. (1964b). Nonmetric multidimensional scaling: A numerical method. *Psychometrika*, *29*(2), 115–129.

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

Lewis, R. L. (1998). Reanalysis and limited repair parsing: Leaping off the garden path. In J. Fodor & F. Ferreira (Eds.), *Reanalysis in sentence processing* (pp. 247–284). Boston, MA: Kluwer Academic Publisher.

MacDonald, M. C., & Christiansen, M. H. (2002). Reassessing working memory: Comment on Just and Carpenter (1992) and Waters and Caplan (1996). *Psychological Review*, *109*(1), 35–54.

MacDonald, M. C., Just, M. A., & Carpenter, P. A. (1992). Working memory constraints on the processing of syntactic ambiguity. *Cognitive Psychology*, *24*(1), 56–98.

McRae, K., Spivey-Knowlton, M. J., & Tanenhaus, M. K. (1998). Modeling the influence of thematic fit (and other constraints) in on-line sentence comprehension. *Journal of Memory and Language*, *38*(3), 283–312.

Meseguer, E., Carreiras, M., & Clifton, C. (2002). Overt reanalysis strategies and eye movements during the reading of mild garden path sentences. *Memory & Cognition*, *30*(4), 551–561.

Mitchell, D. C., Shen, X., Green, M. J., & Hodgson, T. L. (2008). Accounting for regressive eye-movements in models of sentence processing: A reappraisal of the selective reanalysis hypothesis. *Journal of Memory and Language*, *59*(3), 266–293.

R Development Core Team. (2009). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from http://www.R-project.org (ISBN 3-900051-07-0)

Reichle, E. D., Pollatsek, A., Fisher, D., & Rayner, K. (1998). Toward a model of eye movement control in reading. *Psychological Review*, *105*, 125–157.

Reichle, E. D., Warren, T., & McConnell, K. (2009). Using E-Z reader to model the effects of higher-level language processing on eye movements during reading. *Psychonomic Bulletin & Review*, *16*(1), 1–21.

Salvucci, D. D., & Anderson, J. R. (2001). Automated eye-movement protocol analysis. *Human-Computer Interaction*, *16*, 39–86.

Sanford, A. J., & Sturt, P. (2002). Depth of processing in language comprehension: Not noticing the evidence. *Trends in Cognitive Sciences*, *6*(9), 382–386.

Swets, B., Desmet, T., Clifton, C., & Ferreira, F. (2008). Underspecification of syntactic ambiguities: Evidence from self-paced reading. *Memory & Cognition*, *36*(1), 201–216.

Traxler, M. J. (2007). Working memory contributions to relative clause attachment processing: A hierarchical linear modeling analysis. *Memory & Cognition*, *35*(5), 1107.

Traxler, M. J., Pickering, M. J., & Clifton, C. (1998). Adjunct attachment is not a form of lexical ambiguity resolution. *Journal of Memory and Language*, *39*(4), 558–592.

von der Malsburg, T., & Vasishth, S. (2011). What is the scanpath signature of syntactic reanalysis? *Journal of Memory and Language*, *65*(2), 109–127.

van Gompel, R. P., Pickering, M. J., Pearson, J., & Liversedge, S. P. (2005). Evidence against competition during syntactic ambiguity resolution. *Journal of Memory and Language*, *52*, 284–307.

van Gompel, R. P., Pickering, M. J., & Traxler, M. J. (2001). Reanalysis in sentence processing: Evidence against current constraint-based and two-stage models. *Journal of Memory and Language*, *45*(2), 225–258.

Vitu, F., & McConkie, G. W. (2000). Regressive saccades and word perception in adult reading. In A. Kennedy, R. Radach, D. Heller, & J. Pynte (Eds.), *Reading as a perceptual process* (pp. 301–326). Amsterdam, the Netherlands: Elsevier.

Wells, J., Christiansen, M., Race, D., Acheson, D., & MacDonald, M. (2009). Experience and sentence processing: Statistical learning and relative clause comprehension. *Cognitive Psychology*, *58*(2), 250–271.

APPENDIX 1
Parameter estimates of the linear mixed models

TABLE A1
Summary of the linear mixed model of the corrected first pass reading times of the pre-verbal region of the adverbial clause

	Coefficient	SE	t
First pass reading time, pre-verbal region			
c–ab	0.03	0.01	2.58
Working memory	−0.00	0.03	−0.02
c–ab × working memory	0.03	0.01	2.33

TABLE A2
Summary of the linear mixed models of the rates of occurrence of the three types of regression scanpaths

	Coefficient	SE	z
Rereading			
a–bc	0.05	0.03	1.74
b–c	0.01	0.05	0.19
Working memory	0.07	0.11	0.63
first pass reading time	0.05	0.04	1.05
a–bc × working memory	0.05	0.03	1.69
b–c × working memory	0.03	0.05	0.69
a–bc × first pass reading time	0.08	0.03	3.11
b–c × first pass reading time	0.02	0.05	0.48
Checking of the verb:			
a–b	−0.01	0.05	−0.23
ab–c	0.13	0.03	4.01
Rapid regressions from the verb:			
a–b	0.08	0.06	1.43
ab–c	−0.06	0.03	−1.98
working memory	−0.03	0.07	−0.39
first pass reading time	−0.15	0.05	−3.03
a–b × working memory	−0.03	0.06	−0.60
ab–c × working memory	−0.06	0.03	−1.98
No regression			
a–bc	−0.06	0.02	−2.73
b–c	−0.04	0.04	−1.00
First pass reading time	0.08	0.04	2.14

APPENDIX 2
Sample items

1. El profesor dijo que los alumnos se levantaran del asiento...
   a. cuando los directores entraron en la clase de música.
   b. cuando los directores entraran en la clase de música.
   c. si los directores entraban en la clase de música.
2. El ministro dispuso que las ayudas se repartieran...
   a. cuando los sindicatos protestaron por el aumento del paro.
   b. cuando los sindicatos protestaran por el aumento del paro.
   c. si los sindicatos protestaban por el aumento del paro.
3. Pablito impuso que sus golosinas se escondieran...
   a. cuando sus amigos llegaron a la fiesta de cumpleaños.
   b. cuando sus amigos llegaran a la fiesta de cumpleaños.
   c. si sus amigos llegaban a la fiesta de cumpleaños.
4. Los universitarios reivindicaron que las aulas se renovaran...
   a. cuando las facultades crecieron en el número de alumnos.
   b. cuando las facultades crecieran en el número de alumnos.
   c. si las facultades crecían en el número de alumnos.
5. El periodista pactó que sus informantes quedaran al margen...
   a. cuando las noticias molestaron al Gobierno de la provincia.
   b. cuando las noticias molestaran al Gobierno de la provincia.
   c. si las noticias molestaban al Gobierno de la provincia.
6. El cirujano dijo que las enfermeras se prepararan...
   a. cuando los heridos llegaron al hospital universitario.
   b. cuando los heridos llegaran al hospital universitario.
   c. si los heridos llegaban al hospital universitario.
7. El director del teatro sugirió que los actores saludaran...
   a. cuando los aplausos empezaron en el palco de honor.
   b. cuando los aplausos empezaran en el palco de honor.
   c. si los aplausos empezaban en el palco de honor.
8. Los sindicatos reclamaron que el Gobierno aportara soluciones...
   a. cuando los parados aumentaron en el sector de construcción.
   b. cuando los parados aumentaran en el sector de construcción.
   c. si los parados aumentaban en el sector de construcción.
9. Cicerón ordenó que su esclavo favorito tomara nota...
   a. cuando los senadores aplaudieron con admiración y entusiasmo.
   b. cuando los senadores aplaudieran con admiración y entusiasmo.
   c. si los senadores aplaudían con admiración y entusiasmo.