Are interpretations of syntactic ambiguities under working memory load "good-enough"? Evidence from eye movements

Nick Cooper

St John's College, University of Oxford

Thesis submitted for the degree of Doctor of Philosophy

Trinity Term, 2017

Table of contents

Acknowledgements

First, I want to thank my supervisor Kate Nation. Kate's support, enthusiasm and belief in me have ensured that I have reached this point. Thank you, Kate, for all of your guidance, both in academic work and in allowing me the freedom to develop. (Sorry for all the long sentences over the years…) Thanks to the wider LCD/ ReadOxford family, and especially Holly for help with eye-tracking at the start, to Titus for advice on scanpath analysis, and to Niina for making the time fun (and for stats questions that meant I had to look a lot of things up). Thanks also to the Somerville College and St John's College alumni offices, and to the Cognitive Neuropsychology Centre, for help in finding older participants. I would also like to express my gratitude to the Economic and Social Research Council for their financial support, and St John's College, not only for their generous financial support, but for providing a wonderful place to study and live. Thank you also to my college advisor Dorothy Bishop for helpful advice and for making my Twitter procrastination at least partly work-related.

Outside of work, thanks to everyone at Oxford University Student Union (especially my fellow sabbatical officers) for an exciting, challenging year. Friends from OUSU, the MCR, ex-Somervillians, and others have made this an enjoyable few years. In particular, thank you to Marina and Robin for coffees that helped me get work done, and Matt, Jack, James and Obers for beers that didn't. (Thanks also to G&Ds and Combibos Coffee, for all the times they let me get away with one coffee for hours).

Finally, thank you to my parents for a lifetime of love, encouragement, and support. (Yes – it's written!). And of course, Katie – who has, in one way or another, been with me throughout this process, and who is daft enough to have agreed to stay with me for longer. Thank you for everything. Let's now spend some weekends outside.

Abstract

Are interpretations of syntactic ambiguities under working memory load "good-enough"? Evidence from eye movements

Nicholas Cooper

St John's College

Thesis submitted for the degree of Doctor in Philosophy, Trinity Term 2017

Syntactically ambiguous sentences offer an insight into how sentences generally are processed, by examining how readers recognise and reanalyse the ambiguity. However, it is only more recently that the comprehension product of syntactic analysis has been adequately tested, demonstrating that ambiguities are not always fully processed. This work has led to the good-enough approach to language processing and comprehension (e.g., Ferreira & Patson, 2007), which argues that sentence processing is merely *good enough* for the current task, and that our comprehension may not exactly match the content of what has been read. The work presented in this thesis set out to examine what it means for syntactic ambiguity processing to be *good enough*, by monitoring patterns of eye movements as people read sentences containing a temporary syntactic ambiguity. Comprehension questions probed the extent to which the syntactic ambiguity had been resolved. Across six experiments, it was demonstrated that both online sentence processing and comprehension are influenced by the presence of an extrinsic memory load, the presence or absence of comprehension questions, the length of texts being read, and the age of participants. Eye movement patterns were more superficial if the task permitted it; similarly, syntactic ambiguities were misinterpreted more commonly as the task demands increased. The results support a good-enough, adaptive sentence processing system, where initial misinterpretations can linger in the product of syntactic analysis, and which is affected by task demands and individual differences.

Chapter 1

General Introduction

The English language offers sentences that are both meaningful and syntactically correct but that are difficult to interpret. Consider sentences (1.1) – (1.3) (Bever, 1970; Christianson, Hollingworth, Halliwell, & Ferreira, 2001; Ferreira & Clifton, 1986):

(1.1) The horse raced past the barn fell.

(1.2) The defendant examined by the lawyer turned out to be unreliable.

(1.3) While the man hunted the deer ran into the woods.

When reading these sentences, we are metaphorically led "up the garden-path" of the incorrect interpretation. Upon reaching the words *fell*, *by the*, and *ran*, respectively, it becomes apparent that our initial interpretation of the sentence was incorrect: we need to resolve the syntactic ambiguity to reach the correct meaning. So called "garden-path" sentences (e.g., Frazier & Fodor, 1978) are surprisingly common in everyday writing (Clifton & Staub, 2008). Accordingly, establishing how we process syntactic ambiguity has been an active topic in psycholinguistic research for over 40 years (Altmann, 1989; Clifton, Staub, & Rayner, 2007; Sanz, Laka, & Tanenhaus, 2013).

Several questions about syntactic analysis remain the subject of debate. First there is no consensus on how ambiguous sentences (and, indeed, syntax in general) are processed. For instance, are all possible interpretations considered in parallel, or are they processed serially, one at a time? To what extent can non-syntactic information influence syntactic analysis, either in the initial construction of an interpretation, or in reanalysis? Do we always ensure we have fully processed a sentence's syntactic structure, or is the

process sometimes underspecified? Relatedly, what happens to an initial misinterpretation that is no longer being actively considered? Is it lost, or does it linger, interfering with future processing? How well can this process be observed in real time (for instance, by tracking eye movements) – and importantly, is this processing linked to the eventual comprehension of a text? What is the impact of this syntactic processing on cognitive resources more generally – and conversely, to what extent does syntactic processing rely on general resources being available? This final question relates both to individual differences between readers, and to how readers respond to differing task demands.

This chapter will start with a discussion of the methodology and terminology of measuring eye movements during reading. This is followed by a summary of research showing what readers do when presented with garden-path sentences, before turning to the question of *why*. I will compare the range of models put forward to explain syntactic analysis (and reanalysis), from serial approaches such as the garden-path model (e.g., Frazier & Fodor, 1978), to parallel competition-based models (e.g., MacDonald, Pearlmutter, & Seidenberg, 1994), to more recent race-based models that build on both (e.g., Logačev & Vasishth, 2016a; van Gompel, Pickering, & Traxler, 2000, 2001). However, models of sentence processing often assume that comprehension is ultimately attained, a view that has only recently been actively tested – and that has been demonstrated to be over-simplified (Christianson, 2016; Christianson et al., 2001). This leads to theories that suggest sentences are not always processed veridically, with one well-tested theory being the good-enough approach to sentence processing (Christianson et al., 2001; Ferreira, Bailey, & Ferraro, 2002; Ferreira & Patson, 2007). On this account, (re)analysis is often not fully completed, in part due to task demands and processing limitations. The nature of these limitations (such as internal and extrinsic task demands, and the role of working memory) will be discussed. Finally, literature on the challenges

facing sentence processing research, including the consideration of individual differences, will be reviewed. This will establish the basis for a series of experiments intended to elucidate whether syntactic processing is simply good enough, and what *good enough* exactly means.

1.1. Eye-tracking in sentence processing

A range of methodologies have been used to investigate sentence processing: while some studies have presented sentences in the auditory modality (e.g., Caplan, DeDe, Waters, Michaud, & Tripodis, 2011; DeDe, Caplan, Kemtes, & Waters, 2004), many have focused on reading. This thesis will focus on understanding sentence processing during reading, and will therefore discuss almost exclusively studies that have presented sentences visually. It is noteworthy, however, that many of the theories discussed have been built on research using auditory sentence presentation. Given this overlap, research using spoken language is discussed where appropriate.

For research using written language, the simplest methodology is to present sentences in their entirety, and leave participants to read them freely (testing only the overall reading time, possibly alongside comprehension at a later point). One problem with this approach is that it is difficult to ascertain how long participants have spent reading critical words or regions, such as regions where ambiguities arise or are detected. To resolve this, a common methodology in sentence processing research (particularly in the field of syntactic ambiguity processing) is self-paced reading (Jegerski, 2014; Mitchell, 1984). In this paradigm, a participant reads a sentence word-by-word, pressing a button to see the next word each time. Depending on the research questions, participants can sometimes continue to see previously-read words while new words are revealed, but more commonly

see only one word at a time. The dependent variable is the duration from the appearance of a critical word, to the point where the participant presses the button to display the next word (*reading time*). It is presumed that longer reading times index additional processing, suggesting that the participant has encountered more difficulty in incorporating the new word. For example, take sentence (1.3a), copied from (1.3), but this time with an optional comma after the word *hunted*:

(1.3a) While the man hunted(,) the deer ran into the woods.

A reasonable prediction is that *ran* will be read faster if the comma is present rather than absent. If the comma is absent, the reader will usually add *the deer* as the object of *While the man hunted...*, and will not expect the presence of a verb like *ran* (theoretical support for this prediction is discussed later).

Self-paced reading is suitable for investigating a range of questions in the sentence processing field. It does not though come without its downsides. First, it is not a natural way of reading (Metzner, von der Malsburg, Vasishth, & Rösler, 2016): it is rare to read sentences word-by-word in this way. As such, it is not clear whether results from self-paced reading studies accurately reflect how we process sentences in everyday reading. In fact, effects seen when allowing free reading of text have not always been seen in self-paced reading studies (Rayner & Frazier, 1987), creating a difficulty in interpreting results. Self-paced reading may be particularly challenging for people who are less fluent readers, or less familiar with computer-based tasks, such as older adults.

Second, re-reading earlier parts of a sentence is critical. Approximately 15% of our eye movements during reading are *regressive*, looking back to earlier parts of the text (Rayner, 1998). This is likely to be more common if sentences are ambiguous or otherwise difficult to read. If previous words are not displayed, re-reading is prohibited. While this is to an extent the purpose of self-paced reading, it is a limitation that

questions further the extent to which self-paced reading results are applicable outside of the laboratory. This links to a third issue: self-paced reading carries an additional demand on working memory (Gordon, Hendrick, Johnson, & Lee, 2006). By restricting access to earlier words in a sentence, participants are forced to remember these instead. Longer reading times and difficulties with comprehension may reflect difficulty with recalling earlier parts of the text, instead of difficulty with incorporating the new word into the sentence[1]. Finally, presenting a sentence word-by-word also prevents advance sight of upcoming words on the outside of the field of vision, so-called *parafoveal preview*. Parafoveal preview is not only important for efficient reading (by allowing prediction of upcoming words, and accurate programming of eye movements; Rayner, 1998), but is often used as a way of skipping upcoming words entirely, which is also not possible during self-paced reading (cf. Staub, Grant, Clifton, & Rayner, 2009).

How can these difficulties be overcome, while simultaneously analysing reading on a word-by-word basis in real time? The methodology used most commonly to do this is to track eye movements during free reading (Rayner, 1998, 2009). This allows exploration of processing in real time, considering both fixation durations, and forward and regressive eye movements. The link between eye movements and linguistic processing has been espoused for over 40 years (Frazier & Rayner, 1982; Just & Carpenter, 1980; Mehler, Bever, & Carey, 1967; Rayner, 1977, 1978), based on the presumption that wherever the eyes are fixated on is what the brain is processing at that time. So for example, if a person is looking at the word *ran* in *While the man hunted the deer ran into the woods*, it is assumed that they are processing that word at that time. This presumption is not without its challenges (cf. Meseguer, Carreiras, & Clifton, 2002;

[1] This limitation is less of a concern if the methodology allows previous words to be visible, but this creates a methodological problem of its own: participants may develop a strategy of pressing buttons quickly, in order to display the entire sentence and read more freely. This would render reading times uninterpretable.

Mitchell, Shen, Green, & Hodgson, 2008; Reichle, Warren, & McConnell, 2009; Vasishth, von der Malsburg, & Engelmann, 2013; von der Malsburg & Vasishth, 2013). Nevertheless, the benefits for monitoring reading processes in real-time make eye-tracking an invaluable tool for exploring sentence processing in detail.

Eye-tracking studies of reading can take various formats, from the simple presentation of entire sentences or passages, to more complex methodologies that ascertain the extent to which we benefit from parafoveal preview (for reviews, see Rayner, 1998, 2009). Within the basic paradigm of allowing unrestricted reading while eye movements are measured, several key features of the methodology and outputs are worth discussing. First, eye movements comprise a series of *fixations* (where the eyes remain static on part of the text), with intervening *saccades* (movements between fixations). Fixations tend to last for approximately 225-250ms, and saccades for approximately 30ms (Rayner, 1998, 2009). New information is not picked up during saccades, and so it is only during fixations that there is high sensitivity to visual input. For this reason, the most common dependent variables are based on the duration of fixations on key parts of the text (but see von der Malsburg & Vasishth, 2011, 2013). Over the years, a series of common measures have been derived (Clifton et al., 2007; Rayner, 1998, 2009). These measures tend to focus on critical regions of the text – those where it is expected to observe increased reading durations, or an increased rate of regressive eye movements (based on the premise that these are the areas where processing difficulty will occur). The exact measures used (and their definitions) vary between studies, but based on Clifton et al. (2007), this thesis will adopt the definitions given in Table 1.1[2].

[2] Single fixation and first fixation duration are presented for completeness, but are not used in detail in the experimental chapters as they are less appropriate and meaningful for longer regions and texts (Clifton et al., 2007).

Table 1.1.

Definitions of eye movement measures.

Measure	Definition
Single fixation duration	The duration on a region if exactly one fixation was made
First fixation duration	The duration of the initial fixation on a region, if one was made
First pass duration (also known as gaze duration)	The duration of all fixations on a region before exiting in either direction, if a fixation was made
Go-past duration (also known as regression path duration)	The duration from entering a region from the left before exiting the region to the right for the first time (including time spent during regressions to the left)
Second pass duration	The duration spent on any revisits to a region after going past it, if a revisit was made
Total reading duration	The sum of durations of all fixations in a region, if at least one fixation was made
Regressions in (to a given region)	The proportion of trials on which a regression was made into a region, following a first pass of the region
Regressions out (of a given region)	The proportion of trials on which the first eye movement out of the region following a first pass was regressive
Skipping rate	The proportion of trials where a fixation was not made on the region during the first pass (i.e. the region was *skipped*)

An important distinction is made between early and late eye movement measures. *Early* measures of reading such as first pass duration are considered to tap lexical and early syntactic processing, whereas *late* measures such as second pass duration tap reprocessing of sentences (Clifton et al., 2007; Rayner, 1998; see also Vasishth et al., 2013). Measures such as go-past durations and total reading durations are a blend of both types. By revealing reanalysis, late measures will be of particular interest in this thesis. The early/late distinction provides a further advantage over self-paced reading, which cannot offer this level of granularity (cf. Liversedge, Paterson, & Pickering, 1998). That said, the distinction between the two is not clear cut, and early and late measures are often significantly correlated with each other (Vasishth et al., 2013). The overlap between the two is in part because both eye movement control and measurement of eye movements

can be subject to errors (Reichle & Drieghe, 2015). Saccades may not accurately lead to fixations on the intended part of the text, requiring microsaccades to correct this; similarly, errors in calibration as an experiment progresses may lead to inaccuracies, especially in terms of whether a fixation is or is not in a critical region. Appropriate calibration and validation during the experiment mitigates against this reasonably well, but it remains a reason to consider multiple eye-movement measures.

The duration-based eye movement measures tend to follow a positively-skewed distribution that is approximately log-normally distributed (Staub, White, Drieghe, Hollway, & Rayner, 2010). These distributions are sensitive to several features of text, with three being of particular importance: length, frequency and predictability (Rayner, 1998, 2009). Words that are longer, less frequent in the language, and less predictable based on earlier text, tend to be read for longer. Length and lexical frequency are fairly explanatory. There is more debate about what is meant by predictability (see Luke & Christianson, 2016; Smith & Levy, 2011, 2013; Staub, Grant, Astheimer, & Cohen, 2015 for recent examples). It is usually measured in terms of cloze probability: how frequently people choose that word to fill the gap in that sentence (for example, *The girl went outside to fly her…* tends to be filled by *kite*, meaning it has a high cloze probability).

1.2. Difficulties of reading garden-path sentences

It has long been recognised that the structure of some sentences makes them more difficult to process and understand than others (for a review of earlier work, see Mitchell, 1994). Bever (1970) famously introduced the example *The horse raced past the barn fell*, a sentence that is grammatically correct but difficult to process. Bever proposed that there was only an abstract relation between whether a sentence followed grammatical rules, and

whether it is perceived as *acceptable*. He described a series of strategies that a reader may follow to determine acceptability; the idea that we often use heuristics rather than precisely following grammaticality will be important throughout this thesis.

Frazier and Rayner (1982) provided eye-tracking evidence to detail the processing difficulties caused by sentences containing temporary syntactic ambiguities. They compared "early closure" sentences such as (1.4) to "late closure" sentences such as (1.5).

(1.4) Since Jay always jogs a mile seems like a short distance to him.
(1.5) Since Jay always jogs a mile this seems like a short distance to him.

(1.4) is called early-closure because the reader needs to close the first clause after *jogs*, and start a new clause attaching *a mile* to *seems like…* In (1.5), the initial clause closes later, with *a mile* attached to *jogs*. Frazier and Rayner recorded eye movements to determine reading times for specific parts of the sentences. They found significantly longer reading times on early-closure sentences than on their late-closure equivalents. The longer reading was especially seen on the disambiguating word *seems*, with this being the critical point where it becomes apparent that the initial misinterpretation of (1.4) (that Jay always jogs a mile) is incorrect[3]. This inflated reading time was visible from the first fixation on that word. Frazier and Rayner also found longer reading times on revisits (but not initial visits) to the ambiguous region *a mile* in early closure sentences, suggesting that participants returned to this region for longer in order to disambiguate the sentence.

The term *garden-path effect* is used to describe the pattern of longer first pass reading durations on the disambiguating region of garden-path sentences, and/or more regressions from that region to earlier regions. These longer reading times and/or increased rate of regressions are thought to index increased processing (e.g., Clifton,

[3] The term "initial misinterpretation" is used throughout this thesis, referring to when an early closure garden-path sentence is initially built as a late closure construction (e.g., that Jay always jogs a mile).

Traxler, Mohamed, Williams, Morris, & Rayner, 2003; Clifton et al., 2007; Rayner, 1977, 1978, 1998; Rayner, Carlson, & Frazier, 1983), although as discussed more later, this view may be oversimplified (Altmann, 1994; Altmann, Garnham, & Dennis, 1992; Christianson, Luke, Hussey, & Wochna, 2016; Mitchell et al., 2008; von der Malsburg & Vasishth, 2011, 2013). The garden-path effect has been replicated repeatedly, with longer first pass, go-past and total reading durations, and more regressions out of the disambiguating region (for a review, see Clifton et al., 2007). Similar findings have been seen using self-paced reading, the visual world paradigm (where participants hear sentences and have to match them to one of several pictures on a screen; e.g., Spivey, Tanenhaus, Eberhard, & Sedivy, 2002; Tanenhaus, Spivey-Knowlton, Ebergard & Sedivy, 1995; for a review, see Huettig, Rommers, & Meyer, 2011), and by measuring event related potentials (ERPs; e.g., Kliegl, Dambacher, Dimigen, Jacobs, & Sommer, 2012; Gouvea, Phillips, Kazanina, & Poeppel, 2010; Vissers, Chwilla, & Kolk, 2007).

It has also been found across a range of garden-path sentences (Clifton et al., 2007). Examples of different structures are given in Table 1.2. There are good reasons to expect some of these ambiguities to be processed differently from others (Frazier & Clifton, 1996; von der Malsburg & Vasishth, 2013), or indeed for different sentences within any given structure to be processed differently. For instance, the sentence *The bodyguard of the governor that will be retiring after the troubles is very rich* is ambiguous (is the bodyguard or the governor retiring?), but it is possible to parse it and move on without even detecting the ambiguity. Although this may not produce the "correct" interpretation (i.e., the one intended by its author), it will produce a consistent parse. In contrast, when reading *While the father changed the baby played with its toys*, it is obligatory to at least partly reanalyse the sentence upon reaching *played*, or else it will not seem to be grammatically possible. This distinction is discussed in detail later.

Table 1.2.

Examples of garden-path sentence structures.

Type of sentence	Example	Use
Subordinate clause object v main clause subject	While the father changed(,) the baby played with its toys.	Christianson et al. (2001); Frazier and Rayner (1982); Pickering and Traxler (1998)
Subject vs object extracted relative	The director that the movie pleased won a prize.	Traxler, Morris and Seely (2002); Traxler, Williams, Blozis, and Morris (2005)
Main clause v reduced relative	The defendant examined by the lawyer turned out to be unreliable.	Ferreira & Clifton (1986); Kemper, Crow, and Kemtes (2004); Trueswell, Tanenhaus, and Garnsey (1994)
Direct object noun phrase v sentence complement	The lawyers think his second wife will claim the family inheritance (belongs to her).	Frazier and Rayner (1982); Garnsey, Pearlmutter, Myers, and Lotocky (1997); Sturt, Scheepers, and Pickering (2002)
Attach prepositional phrase to verb or noun	The people who lived near the canal blamed the toxic waste dump for many years (/for their leukaemia).	Britt, Perfetti, Garrod, and Rayner (1992); Speer and Clifton (1998)
Adverb high or low	The carpenter sanded the shelves he attached onto the kitchen wall yesterday morning.	Altmann, van Nice, Garnham, and Henstra (1998); van Gompel, Pickering, Pearson, and Liversedge (2005)
Relative clause to first or second noun phrase	The bodyguard of the governor that will be retiring after the troubles is very rich.	Traxler, Pickering, and Clifton (1998); van Gompel et al. (2005)
Argument or adjunct	The shrubs were planted by the greenhouse (/apprentice).	Liversedge, Pickering, Branigan, and van Gompel (1998)
Noun phrase or subject	The model embraced the designer and the photographer laughed.	Hoeks, Vonk, and Schriefers (2002); Staub, Clifton, and Frazier (2006)

The existence of the garden-path effect is well supported, but there is considerable debate about the role of semantic, lexical and discourse factors within syntactic ambiguity resolution (Altmann et al., 1992; Britt et al., 1992; Clifton et al., 2003, 2007; Ferreira & Clifton, 1986; MacDonald et al., 1994; Rayner et al., 1983; van Gompel & Pickering, 2007). The theoretical implications of this debate are discussed later in this chapter; nevertheless, several factors have been found to influence the extent of the garden-path effect. One factor is the length of the ambiguous region before disambiguation occurs. Longer ambiguous regions produce greater garden-path effects than shorter regions[4], as measured by judgments of grammaticality (Ferreira & Henderson, 1991), eye movements (e.g., Frazier & Rayner, 1982), and comprehension question accuracy (Christianson et al., 2001; Christianson, Williams, Zacks, & Ferreira, 2006).

Other factors include whether the initial verb (e.g., in (1.3), *hunted*) appears more frequently in a transitive or intransitive usage (Ford, Bresnan & Kaplan, 1982; Garnsey et al., 1997; MacDonald & Seidenberg, 2006; Pickering, Traxler, & Crocker, 2000) and whether the verb is in the simple past tense as in (1.3), or in a past progressive tense (*was hunting*), with the latter reducing the garden-path effect (Frazier, Carminati, Cook, Majewski, & Rayner, 2006). Finally, garden-path effects have been influenced by the plausibility of the initial interpretation (e.g., in (1.3), of the man hunting the deer; Pickering & Traxler, 1998). Effects may even be influenced by whether the sentence is preceded by a context that biases towards (or away from) the garden-path interpretation (Altmann et al., 1992; Britt et al., 1992; Spivey-Knowlton, Trueswell, & Tanenhaus, 1993; but cf. Ferreira & Clifton, 1986; Clifton et al. 2003; Rayner, Garrod, & Perfetti, 1992; Rayner & Sereno, 1994). For all of these, the debate is less about whether the

[4] It is not thought to be the length, *per se*, that is crucial, but the number of linguistic components involved (Grodner & Gibson, 2005; Staub, 2007; van Dyke & Lewis, 2003; Warner & Glass, 1987).

factors influence reading times at all, but more about whether these non-syntactic factors influence initial sentence processing (as indexed by *first pass* reading durations), or if they only influence reanalysis. More widely, determining the role of non-syntactic processes is often shaped by debate concerning the architecture of the cognitive processes involved. It is important to therefore begin evaluating potential models for the cognitive structure of syntactic processing.

1.3. Garden-path models of sentence processing

Garden-path sentences are fairly specific syntactic constructions; nevertheless, examining how these sentences are processed is thought to illuminate syntactic processing mechanisms more generally (Altmann, 1989; Frazier & Fodor, 1978). The most prominent explanations of syntactic processing have been the family of *garden-path* models (Frazier & Fodor, 1978; Frazier & Rayner, 1982). The garden-path model assumes a two-stage syntactic processor, first assigning new words to lexical or phrasal nodes, and second combining these to form a grammatical sentence. The first stage occurs according to two main strategies (Frazier & Rayner, 1982): *Minimal Attachment* – to use the fewest nodes possible – and *Late Closure* – to attach each incoming word to the currently active node where grammatically licensed to do so[5]. This can be illustrated by repeating sentence (1.3):

(1.3) While the man hunted the deer ran into the woods.

In (1.3), these two strategies would work to initially attach *the deer* as the direct object of *hunted*. *To hunt* is optionally-transitive (one can hunt [SOMETHING], or simply hunt),

[5] Minimal attachment is considered the primary strategy; if there is a conflict between the two, the interpretation that fits with minimal attachment will be preferred by the syntactic processor.

and this verb can therefore adopt a direct object; attaching *the deer* in this way is therefore licit (of course, it is also plausible – which is relevant in later discussions). Attaching *the deer* also prevents the addition of extra nodes that would be required if a new clause were started at *the deer* (despite this being the ultimately correct interpretation), in line with minimal attachment. The model has subsequently been supplemented with an *active filler strategy* (e.g., Frazier & Clifton, 1989), where the processor aims to fill gaps using the first available object. For instance, in *Which girl do you believe John loves a lot?*, the *Which girl…* is not currently attached to anything, and so fills the gap of who is being *believed*. This leads to an interpretation that would be answered, for example, *I believe [a certain girl]*. In turn, this causes processing difficulty at the word *John*, where the sentence must be reinterpreted (to be answered *I believe John loves [a certain girl]*).

The garden-path model has several key features. The model requires its two-stage architecture: a single-stage theory would predict that resources would have to increase proportionately to the number of words being processed (Kimball, 1973), while in practice, sentence length alone is a poor index of complexity (Frazier & Fodor, 1978; Grodner & Gibson, 2005). Two stages are therefore necessary – one constrained by the complexity of the individual nodes, and a second that can combine these nodes without such a restriction (Frazier & Fodor, 1978). This architecture is serial: the model only actively pursues one interpretation at any given time, and competing interpretations only come into play if the initial analysis is abandoned. This is based on the principle that language is incremental (Kamide, Altmann, & Haywood, 2003; Marslen-Wilson, 1975; Pollatsek, Reichle, & Rayner, 2005): each new word is added to the structure that has been built from previous words. If an error is detected (as in garden-path sentences), the serial parser has to go back to reanalyse the sentence, which is often costly in time and

effort (Clifton et al., 2007; Ferreira & Clifton, 1986; Frazier & Rayner, 1982). The model is also deterministic: for a given syntactic structure, it will always initially produce the same interpretation. This interpretation will be the one that fulfils the Minimal Attachment and Late Closure strategies, as this requires the least computational resources.

Finally, the garden-path model postulates a *modular* system (Fodor, 1983; Frazier, 1987): the syntactic processor is domain-specific, and relies only on syntactic information without reference to other modules. When building an initial syntactic structure, non-syntactic input such as semantic or lexical information is not considered; these factors should only influence reanalysis and therefore only be observable in later re-reading measures. This means that difficulty with garden-path sentences is attributed to difficulty in *syntactic* processing (with establishing which units go together), rather than, for instance, difficulty in domain-general processes such as memory (Frazier & Fodor, 1978).

There is considerable support for the tenets of the garden-path model (Ferreira & Clifton, 1986; Frazier & Clifton, 1989; Frazier & Rayner, 1982). The principle of Late Closure is supported by the garden-path effect in early closure sentences (Frazier & Rayner, 1982). When reading *Since Jay always jogs a mile seems…* , the verb *seems* cannot be integrated into the interpretation that has been built and reanalysis must be instigated, consistent with the garden-path model. Minimal Attachment is supported by evidence that readers do try to reduce the number of nodes when analysing sentences. Rayner and Frazier (1987) found longer reading times for sentences such as (1.6) compared to sentences such as (1.7).

(1.6) Karen knew the schedule was wrong.

(1.7) Karen knew that the schedule was wrong.

The initial interpretation of (1.6) (which is synonymous with *Karen was familiar with the schedule*) involves the fewest nodes: having *the schedule* as the direct object of

knew is simpler than creating a separate clause with *the schedule* as its subject. The simplest interpretation is therefore adopted initially, and if this is incorrect, further processing is required to revise it, consistent with the garden-path model. Interestingly, Rayner and Frazier's study found this effect despite it not being found in Holmes, Kennedy and Murray's (1987) self-paced reading study using similar stimuli, highlighting the importance of considering methodological differences when interpreting sentence processing data.

If the parser detects an error, the next question is: how does it reanalyse the sentence to attain comprehension? As mentioned, Frazier and Rayner (1982) found that having made a first pass of the disambiguating region of early-closure sentences, participants would spend longer on their second pass of the ambiguous noun phrase region (in the sentence about Jay jogging, (1.4), this would be *a mile*). They took this as evidence for a selective reanalysis strategy, where a reader can return to the site of ambiguity and focus their reprocessing there. They contrasted this against alternative theories that the parser may instead simply regress backwards until the ambiguity is resolved (Kaplan, 1972), or return to the start of the sentence and re-read it all (cf. Lewis, 1998). Selective reanalysis would again be the prediction of the garden-path model: having adopted an incorrect parse, the processor returns to the node where ambiguity arose, and reanalyses the structure accordingly. Frazier and Rayner's work supported selective reanalysis; nevertheless, later research that has focused specifically on regressions has identified alternative patterns of regressions (Christianson et al., 2016; Meseguer et al., 2002; Mitchell et al., 2008; von der Malsburg & Vasishth, 2013). Regressions may serve more purposes than simply to correct errors; for example, in simply verifying previous input, or to pass time while the processor completes its analysis. This work, and differences in regression patterns, is discussed further later.

1.4. Role of non-syntactic factors

Despite significant theoretical and empirical support for the garden-path model, there has been debate about how non-syntactic information can influence syntactic processing, and specifically syntactic ambiguity resolution (Altmann, 1989; Altmann et al., 1992; Crain & Steedman, 1985; Ferreira & Clifton, 1986; MacDonald et al., 1994; Trueswell et al., 1994; van Gompel & Pickering, 2007). The garden-path model is clear: an encapsulated syntactic processor will only process syntactic information, and so lexical, semantic and discourse-level information will not influence initial sentence processing. This does not preclude semantic information being used subsequent to an initial parsing (i.e., after first pass), in order to check or adjust the decision of the parser (cf. Frazier & Rayner, 1982). This modular account stands in contrast to "interactive" accounts (e.g., Marslen-Wilson & Tyler, 1980), in which syntactic processing takes place in interaction with semantic, lexical and other processes.

Early evidence supported the principle of modularity, finding that non-syntactic information only played a role in reprocessing, and not first pass processing (e.g., Ferreira & Clifton, 1986; Rayner et al., 1983). Other research has found an early effect of semantic factors on syntactic processing of garden-path sentences (e.g., Clifton et al., 2003, Experiment 2; Spivey-Knowlton et al., 1993; Trueswell et al., 1994; for reviews, see Clifton & Duffy, 2001; Rayner & Clifton, 2002). For example, Trueswell et al. (1994) found that the animacy of the noun in a reduced relative clause[6] had immediate effects of syntactic ambiguity resolution in the eye movement record, despite this being non-syntactic information. Marslen-Wilson, Tyler, Warren, Grenier, and Lee (1992) also

[6] Animacy here determines the plausibility of the two interpretations of a reduced relative sentence. For example, in *The thief/room searched by the police was quite unpleasant*, it is implausible that a room will be doing the searching (as opposed to being searched), and so garden-pathing is less likely than for *thief*.

found that beneficial prosody can overcome garden-path effects in auditory sentence presentation (similarly to punctuation such as a comma in written language; Christianson et al., 2001).

This debate has also extended to the question of whether garden-path effects can be reduced or even overcome by adding felicitous context before the temporarily ambiguous sentence (e.g., Altmann, 1988; Altmann et al., 1992; Altmann & Steedman, 1988; Britt et al., 1992; Crain & Steedman, 1985; Spivey-Knowlton et al., 1993). Again, the garden-path model purports that discourse context could only affect reanalysis, and could not influence first pass effects. In support, proponents of the model found that adding a context sentence (or passage) that biased towards the correct interpretation of the garden-path sentence (see (1.8) below) helped with *reanalysis*, as indexed by shorter total reading durations, but did not affect the initial first pass effect (e.g., Ferreira & Clifton, 1986; Frazier & Rayner, 1982; Rayner et al., 1983, 1992; Rayner & Frazier, 1987).

Research led by Altmann and colleagues found otherwise, demonstrating that adding a felicitous referential context could eliminate first pass garden-path effects. Altmann et al. (1992) found no first-pass garden path effect in the disambiguating region of sentences such as (1.9a), compared to control sentences such as (1.9b), if the sentence was preceded by a supportive context, such as (1.8) (disambiguating regions are marked in bold).

(1.8) An off duty fireman was talking to two women. He was telling them how serious the situation had been when their house caught fire. The fireman had risked his life to rescue one of the women while the other had waited outside.

(1.9a) He told the woman that he'd risked his life for **to install** a smoke detector.

(1.9b) He asked the woman that he'd risked his life for **to install** a smoke detector.

Altmann et al. interpreted these findings in terms of *referential theory* (Altmann & Steedman, 1988; Crain & Steedman, 1985). This theory states that a reader constructs an interpretation that requires the fewest presuppositions. When there is no preceding context (such as if (1.9a) were read in isolation), it is presumed that there is only one woman, as this is more parsimonious than presuming there are multiple women, one of whom he risked his life for. This leads to processing difficulty, and hence a garden-path effect. In contrast, when preceded by (1.8), the reader can establish a link to the referential context (that there are two women, one of whom was rescued), and thus avoid being garden-pathed, even on first pass.

The conflicting findings for the role of semantic and discourse factors can be attributed in part to differences in methodology and stimuli between experiments (for discussion, see Altmann, 1994; Clifton & Ferreira, 1989; Clifton et al., 2003, 2007; Rayner & Sereno, 1994; Steedman & Altmann, 1989). Of particular relevance is that some manipulations have a stronger effect than others: Rayner, Warren, Juhasz and Liversedge (2004) found that implausible sentences (*John used an axe to chop the large carrots…*) were *re*-read for longer than plausible ones, but with no effect on first pass durations. In contrast, anomalous sentences (*John used a pump to inflate the large carrots…)* produced increased first pass durations on the word *carrots*, suggesting that it takes a substantial change in plausibility to affect first pass processing. It is also possible that first pass durations may not cleanly isolate initial processing, and initial syntactic parses may be checked against other constraints almost immediately (Vasishth et al., 2013; see also Clifton et al., 2007). Nevertheless the results from studies manipulating semantic and discourse effects suggest that the original garden-path model may be insufficient to provide a complete explanation of syntactic ambiguity resolution, and hence of syntactic processing.

One option to overcome this is to extend the garden-path model by maintaining its fundamental structure, but either postulating additional components, or suggesting that not all syntactic processing decisions are conducted similarly. Rayner et al. (1983) proposed the addition of a thematic processor (see also Ferreira & Clifton, 1986; Frazier, 1985), which puts forward possible thematic structures (that is, the structure of relations within the sentence), and then decides between them. For instance, the thematic processor may consider whether a noun phrase is the agent or patient of a verb phrase in a garden-path sentence. This module would remain distinct from the syntactic processor, but the semantic effects outlined above may simply reflect separate processing by these two modules, which are able to interact, albeit serially.[7] The thematic processor may come into play almost immediately when the syntactic processor fails to build a plausible interpretation.

Alternatively, the garden-path model has been extended by suggesting that the main syntactic analysis module only processes certain syntactic relations, with others left to a "construal" process (Carreiras & Clifton, 1993; Frazier & Clifton, 1996). The construal model suggests that primary phrases (those containing arguments between a subject and object) are processed in accordance with the garden-path model. In contrast, non-primary phrases (such as prepositional phrases) are not attached in a fixed way immediately, but are instead loosely attached to the whole phrase being considered, with a precise decision being delayed until further information is received (i.e., the parsing is underspecified). While the first process will not be influenced by semantic factors, the construal process may be. In support of this, Speer and Clifton (1998) found faster reading of arguments than adjuncts in *The people blamed the toxic waste dump for their*

[7] These effects could also reflect methodological insufficiencies: self-paced reading cannot offer any distinction between early and late processes; if semantic processing is almost immediately engaged after syntactic processing, the distinction may be there but not visible (Rayner & Frazier, 1987).

leukaemia (argument)/*for many years* (adjunct). This could explain the effects of both discourse (Britt et al., 1992) and semantics (Swets, Desmet, Clifton, & Ferreira, 2008). By putting forward two pathways for syntactic processing, the construal model may also explain the different eye movement patterns on different structures of garden-path sentences (von der Malsburg & Vasishth, 2013).

1.5. Alternative models of sentence processing to the garden-path model

The conclusion of the work described above is that semantic and discourse factors can influence syntactic processing, with some evidence to support a role even in the initial building of a syntactic structure. The garden-path model has been extended to try to incorporate these findings; the alternative to this is to propose a new model of sentence processing. This section considers the two main families of alternative models, although others have been proposed (for a review, see e.g., van Gompel & Pickering, 2007).

1.5.1. Constraint-based models

One radically different view is to abandon the serial, modular garden-path model, in favour of an interactive, parallel architecture (Bates & MacWhinney, 1989; MacDonald et al., 1994; MacDonald and Seidenberg, 2006; McRae, Spivey-Knowlton, & Tanenhaus, 1998; Spivey & Tanenhaus, 1998; Trueswell et al., 1994; Tabor, Juliano, & Tanenhaus, 1997). Constraint-based models simultaneously take syntactic and non-syntactic information into account in a single stage. Within a connectionist network, weights between inputs and outputs are determined, or *constrained*, simultaneously and optimally by multiple factors (Bates & Macwhinney, 1989). These factors may be syntactic, semantic such as plausibility or animacy, discourse-based provided by context,

or lexical such as the frequency of words in different sentences (MacDonald & Seidenberg, 2006).

Unlike in the garden-path model, multiple interpretations are analysed in parallel. Consider again sentence (1.3), *While the man hunted the deer ran into the woods*. In a constraint-based model, competition occurs at the ambiguous noun phrase *the deer*. This phrase could be the direct object of *hunted* (the initial interpretation, eventually discovered to be incorrect), or the start of a new clause. The garden-path model always chooses the first interpretation, turning to the second only during reanalysis. Constraint-based models, in contrast, choose from possible interpretations by combining the various constraints optimally; for instance, the plausibility of *the deer* being the direct object of *hunted*. A garden-path effect indicates that the early-closure interpretation (the man hunting the deer) is considered more favourably based on all available information. Eventually the reader encounters *ran*, weightings change, and the alternative interpretation is henceforth weighted more strongly.

Constraint-based models can generally account for findings of semantic interactions with syntactic ambiguity resolution (e.g., Pickering & Traxler, 1998), as well as the influence of discourse-level factors on syntactic processing (Altmann & Steedman, 1988; McRae et al., 1998; but cf. Ferreira & Clifton, 1986). Lexical, syntactic, semantic and contextual information can be integrated and evaluated simultaneously, with garden-path effects resulting simply from the infrequency of certain ambiguous structures in language. Not all constraint-based models have been fully specified and tested, making explicit predictions difficult (van Gompel & Pickering, 2007), but several have been (e.g., McRae et al., 1998; Tabor et al., 1997), and several have been extended to explain eye movements (e.g., Just & Carpenter, 1992; Tanenhaus, Spivey-Knowlton, & Hanna, 2000).

Evaluating the respective merits of the garden-path model and constraint-based models involves determining if the parser operates in serial or in parallel. Determining this is difficult (Lewis, 2000; Staub, 2007; van Gompel & Pickering, 2007). For example, Staub (2007) found that when participants construct and then abandon the analysis that ultimately turns out to be correct *before* being garden-pathed, they show fewer regressions when being garden-pathed. This suggests that the previous construction is maintained; this is inconsistent with the garden-path model, where abandoned interpretations would no longer be considered. However, some serial models do allow for multiple analyses at the point of ambiguity, even if only one is maintained at any given time (e.g., Lewis & Vasishth, 2005; Traxler et al., 1998; van Gompel et al., 2000, 2001), and these models could account for Staub's data. Answering the serial-parallel question does not therefore answer the one vs. multiple analyses question.

The more pertinent and easily testable question when comparing these models is thus if there is evidence for competition between interpretations during initial sentence processing. Most constraint-based models predict competition between interpretations at each stage of reading a sentence, even if the nature of this competition varies (Green & Mitchell, 2006). As such, it has been suggested that these models would predict significant processing difficulty during the ambiguous parts of a sentence when two interpretations are equally plausible (e.g., van Gompel et al., 2000). For example, take sentence (1.10):

(1.10) The hunter killed the dangerous poacher with the rifle not long after sunset.
(1.11) The hunter killed the dangerous leopard with the rifle not long after sunset.

In (1.10), the words *hunter* and *poacher* were pre-tested by van Gompel et al. to confirm that they are balanced for plausibility – it is equally plausible that *the rifle* could belong to either of them. As such, a slowdown should be observed when reading this ambiguous

region, owing to competition between the interpretations. In contrast, the principle of minimal attachment in the garden-path model would always favour the interpretation of attaching *the rifle* to the hunter, as this is the interpretation with the fewest nodes (compared to attaching *the rifle* to the noun phrase *the dangerous poacher*). As such, the garden-path model predicts there would be no difference in reading times for (1.10) compared to (1.11), where it is implausible that the *leopard* will have the rifle.

Evidence for a slowdown (and hence for competition) is limited, and intriguingly, several studies have actually found faster reading in these ambiguous regions (the so-called *ambiguity advantage*; Traxler et al., 1998; van Gompel et al., 2001, 2005). van Gompel et al. (2000) argued that this is a challenge for constraint-based models, claiming that these models would predict a processing cost compared to ambiguous sentences that have a clear preference in terms of plausibility (cf. (1.11)). This criticism does not necessarily apply to all constraint-based models: Green and Mitchell (2006) ran simulations demonstrating that the absence of a slowdown is compatible with the predictions of one constraint-based model[8] (McRae et al., 1998), and may be compatible with others (e.g., MacDonald & Christiansen, 2002; Tabor et al., 1997). The debate is not over: replies to Green and Mitchell have questioned whether the envisaged competition would be too resource-intensive for most sentences – where there are often multiple or even numerous interpretations of an ambiguity, and certainly more than just two (Clifton & Staub, 2008; van Gompel & Pickering, 2007). In summary, these data challenge the fixed, serial garden-path model, but may also be incompatible with many constraint-based models.

[8] Green and Mitchell argue that this is because processing in this model is not at its highest when there is a conflict between two options (as van Gompel et al suggest), but where there is a conflict between prior biases that are embedded in the model, and new material.

1.5.2. Unrestricted race model

van Gompel and colleagues (Traxler et al., 1998; van Gompel et al., 2000, 2001, 2005) argued that the fixed strategies of Minimal Attachment and Late Closure do not account for the ambiguity advantage results in sentences such as (1.10) and (1.11). The garden-path model predicts that readers will always attach *with the rifle* to *poacher/leopard*, but this was not consistent with their eye movement data, which showed no such pattern of preference. As noted above, the ambiguity advantage also presents a challenge for at least some versions of constraint-satisfaction accounts. An alternative account is the *unrestricted race model* (Traxler et al., 1998; van Gompel et al., 2000, 2001, 2005). In this model, possible interpretations compete in a race; non-syntactic information acquired from previous words in the sentence can be used immediately to determine the time it takes each interpretation to be built (it is this sense in which it is *unrestricted*). van Gompel et al. (2000) suggested this is able to explain data that show graded garden-path effects[9] in response to non-syntactic manipulations (Garnsey et al., 1997; Trueswell et al., 1994). In contrast to constraint-based models, the model remains serial as only one syntactic interpretation is maintained after each race (put simply, there is one "winner"). The model is also non-deterministic: the time to complete analysis of each interpretation is stochastic, and so the same structure may be analysed in different ways on different occasions; individual differences can also be accounted for by this non-determinism.

The unrestricted race model can still account for the overall garden-path effect: in the event of being garden-pathed, there remains a processing cost associated with reanalysis. Furthermore, effects of animacy on syntactic processing (e.g., Trueswell et al.,

[9] Here, *graded* means that a bias for one interpretation over another does not result in preference for that interpretation every time; instead, Garnsey et al. (1997) found a moderate correlation between plausibility of a noun phrase being a direct object or a subject, and reading times.

1994) can be explained: animacy information from previous words is able to facilitate the building of an interpretation. The model can also explain two other important results. First, in Staub (2007), reanalysis was facilitated when the reader had previously constructed the correct interpretation in an earlier part of the sentence, only to abandon this upon being garden-pathed. Second, Mohamed and Clifton (2011) found that embedding ambiguous sentences in a biasing context slowed down reading times in the disambiguating region (contrary to the garden-path model), but had no effect on reading times on the ambiguous region. The unrestricted race model can explain these findings, by allowing variability in the syntactic parser's initial choice, and because the model can immediately use information from before the site of disambiguation. Both findings are more difficult to reconcile with the garden-path or constraint-based models: the garden-path model predicts a fixed parser decision that will be unaffected by context, while most constraint-based models would predict a slowdown in ambiguous regions that was not seen.

1.6. Underspecification, heuristics and good-enough processing

Swets et al. (2008) noted that despite the advantages of the unrestricted race model over other models, it still carries the assumption (along with most other serial models) that one interpretation will eventually be chosen, and that a fully specified parsing of the sentence will always be formed. Indeed, most syntactic parsing models have typically assumed that the extra effort needed to reanalyse an ambiguous sentence is rewarded by a complete parsing of the sentence, and so grammatically-correct comprehension. This view has not gone unchallenged. The idea that people use heuristics rather than always relying on strict pursuance of grammatical rules is not new (Altmann, 1989; Bever, 1970; Townsend & Bever, 2001), and is consistent with the use of heuristics more generally (e.g., Aaronson & Ferres, 1986; Bartlett, 1932; Gigerenzer, Todd, & the ABC Research Group, 1999; Sanford & Sturt, 2002; Simon, 1956). Furthermore, it has been demonstrated that when reading, people tend to avoid reanalysis where at all possible (e.g., Fodor & Frazier, 1980; Sturt, Pickering, Scheepers, & Crocker, 2001). This suggests that if an initial parse does not perfectly reflect what has been written, reanalysis may not be triggered if the costs of doing so are too high.

The literature has recognised that comprehension is not all-or-nothing, as seen in the construction-integration model of comprehension (Butcher & Kintsch, 2012; Kintsch, 1988, 1998; Kintsch, Welsch, Schmalhofer, & Zimny, 1990; van Dijk & Kintsch, 1983). This comprises three levels of comprehension. The first level, the *surface* level, represents the exact words that have been read, but is rapidly lost. Next is the *textbase* level, where the propositional content of the text is represented; for example, who has done what to whom in the sentence. In the longer term, for instance on a later day, this textbase model is integrated with existing knowledge at the *situation model* level. At each

level, the representation of the text becomes less veridical, and at the situation model level, existing schemas and knowledge will influence text recollection. The distinction between these levels has been tested by assessing whether paraphrases and/or inferences are accepted on a later recognition test (e.g., Fletcher & Chrysler, 1990). For instance, it has been demonstrated that after 40 minutes, paraphrases are likely to be accepted in a recognition test (see Kintsch et al., 1990). This suggests that after a short duration, the surface structure of a text (and with it, the precise syntactic structure) may degrade and be influenced by existing knowledge and heuristics.

Despite this work on the structure of comprehension, it is only recently that participants' comprehension of syntactically ambiguous sentences has been directly tested with questions tapping initial misanalyses (Christianson, 2016; Christianson et al., 2001; Kaschak & Glenberg, 2004; Swets et al., 2008; van Gompel, Pickering, Pearson, & Jacob, 2006). For example, Christianson et al. (2001) found that longer sentence reading times for garden-path vs. control sentences did not reflect perfectly resolved comprehension. Participants answered *Yes* (incorrectly) to questions like (1.13) after reading sentences such as (1.12a) on 57% of trials, despite answering the same questions correctly on 88% of occasions when the questions followed sentences that were disambiguating by a comma, such as (1.12b)[10].

(1.12a) While the father changed the baby that was cuddly played with its toys.

(1.12b) While the father changed, the baby that was cuddly played with its toys.

(1.13) Did the father change the baby?

[10] The numbers here are for comprehension accuracy for reflexive absolute transitive items in Experiment 3b of Christianson et al. (2001). The relevance of citing these items is discussed in the next paragraph.

The likelihood of making comprehension errors was also moderated by linguistic variables, such as verb type and sentence length. Christianson et al. (2001) found participants made more errors on questions relating to optionally-transitive verbs such as *hunted* in *While the man hunted the deer ran into the woods,* than reflexive absolute transitive verbs such as *changed* in (1.12a). This may reflect the fact that answering *Yes* with an optionally-transitive verb is not strictly incorrect, but is not licensed by the information given in the sentence. Making the sentences longer (in (1.12a), by adding *that was small and cute* after baby) also produced more errors. It is thought that with longer texts, there is more commitment to the initial misinterpretation, making reanalysis more difficult (cf. Christianson et al., 2001; Ferreira & Henderson, 1991, 1998).

van Gompel et al. (2006) questioned whether these results are a by-product of the methodology. Arguably, asking questions such as (1.13) may produce a strategy that affects the processing and/or recollection of sentences. This concern seems unfounded though, as compatible findings of incomplete processing have been identified using various other paradigms. For example, van Gompel et al. (2006) used syntactic priming, giving participants garden-path sentences (or equivalent sentences with a comma) to read, before asking them to complete a sentence production task. They found evidence of syntactic priming: after reading garden-path sentences, participants were more likely to produce transitive sentences than after reading comma sentences, suggesting that these misinterpretations had lingered and were not fully extinguished. Additionally, Patson, Darowski, Moon, and Ferreira (2009) tested comprehension of similar subject-object ambiguous sentences (like (1.12a)) by asking participants to provide a paraphrase of what they had read; for syntactically ambiguous sentences, participants tended towards a paraphrase that blurred both the initial misinterpretation (*the father changed the baby*) and the correct one (*the father changed* [his own clothes]). These results mirror similar

findings for other sentence constructions, such as passive sentences (e.g., *The dog was bitten by the man*; Ferreira & Stacey, 2000), and reduced relatives (*The coach smiled at the player tossed a frisbee;* Tabor, Galantucci, & Richardson, 2004). They also fit with the wider literature on underspecified comprehension (e.g., the Moses illusion, where people presented with *How many animals did Moses take on the Ark?* do not point out that this was actually Noah; Barton & Sanford, 1993; Sanford & Sturt, 2002).

There has been less work tapping incomplete syntactic processing using eye movements. One exception is Slattery, Sturt, Christianson, Yoshida and Ferreira (2013), who tracked eye movements as people read short passages such as (1.14a) and (1.14b):

(1.14a) While the boy washed(,) the dog that was hot hid behind the trees. The boy always washed himself before eating lunch.

(1.14b) While the boy washed(,) the sun that was hot hid behind the trees. The boy always washed himself before eating lunch.

Slattery et al. manipulated the plausibility of these sentences: in (1.14a) it is believable that the boy could be washing the dog, but this is not true for *sun* in (1.14b). Slattery et al. found the usual garden-path effect on *hid behind…* (although, interestingly, only in go-past and total reading durations, and not first pass durations). They also found that for plausible items such as (1.14a) only, participants spent longer reading *himself before* in the comma-absent condition. This indicates surprise at this reflexive term: the representation that the boy is washing the dog (and not himself) appears to have lingered. Slattery et al. (2013) argued that this misinterpretation was not completely pruned before moving to the second sentence.

Ferreira and colleagues (Christianson et al., 2001, 2006; Ferreira & Patson, 2007) proposed the *good-enough processing* account to explain many of these findings. They

argued that we often settle on heuristic-based interpretations that may not be veridical (for example, they may contain traces of abandoned misinterpretations), but are nevertheless usually good enough for the task at hand. This may be particularly important to avoid costly re-reading and reprocessing where resources are otherwise occupied, or when task demands are high. There are several versions of the good-enough framework (Christianson et al., 2001; Ferreira & Patson, 2007; Karimi & Ferreira, 2016; Slattery et al., 2013; Swets et al., 2008). Some have focused on underspecification of syntactic representations during initial processing (e.g., Swets et al., 2008), but more often, comprehension failures are attributed to lingering misinterpretations[11]. Christianson et al. (2001) also found that participants had few problems with answering questions about the main clause, such as *Did the baby play with its toys?* Together, these results suggest that people do reanalyse the sentence, but the initial interpretation of the father changing the baby is also maintained. This is especially true for longer sentences, where the initial misinterpretation is committed to for longer (Christianson et al., 2001).

Christianson (2016) outlined the general principles of a good-enough account: based on an overall goal to facilitate communications (Hale, 2011), the syntactic processor aims to accept a viable interpretation as quickly as possible, taking into account conflicting information. This process may not always be completed. In the event of being garden-pathed, a new interpretation may be built before there is time to fully extinguish the initial misinterpretation. Karimi and Ferreira (2016) elaborated on this, suggesting that we have two systems – one algorithmic and reliant on structural information, and one heuristic. Both attempt to form a consistent parse, but the heuristic system is faster – and may influence the algorithmic system in turn. As such, a simpler, perhaps less veridical

[11] This differentiation may in part come from the use of different types of garden-path sentence, as discussed in section 1.1 (see von der Malsburg & Vasishth, 2013); this is expanded on in Chapter 2.

interpretation may be built and adopted ahead of a strictly faithful parse. Most of the time, this is *good enough* for the task at hand. On occasion however, it might lead to comprehension errors.

The good-enough framework is not the only theory that aims to explain how language processing is not always veridical to the exact inputs received. For example, noisy-channel models (e.g., Gibson, Bergen, & Piantadosi, 2013; Levy, 2008; Levy, Bicknell, Slattery, & Rayner, 2009), based on surprisal and Bayesian estimation techniques, suggest that a reader's representation of linguistic input can be disrupted by noise from distractors, other processing constraints, and the person's prior expectations. Similarly, the Now-or-Never bottleneck hypothesis (Christiansen & Chater, 2016) suggests that cognitive demands force us to rely on predictions, and lead our representations to become "lossy": by acting quickly, detail is lost in favour of more abstract representations. Also, van den Broek, Lorch, Linderholm, and Gustafson (2001) introduced the idea of *standards of coherence*, suggesting that text comprehension is guided by higher-level goals, and that reading times may vary based on what standards are applied. In this thesis, I adopt the good-enough framework for two reasons. First, it builds on similar research focusing on a good-enough approach that has used early-closure garden-path sentences (Christianson et al., 2001, 2006; Patson et al., 2009; Slattery et al., 2013). Second, it is not entirely clear how the alternative approaches differ in their predictions (Christianson, 2016; Ferreira & Christianson, 2016). These alternative approaches (and how they fit with the results described in this thesis) will be discussed further in Chapter 6.

In general, the good-enough approach is an appropriate framework for discussing how comprehension is not always veridical during syntactic ambiguity resolution. Ferreira and Patson (2007) argued that good-enough processing results from limitations in

cognitive resources and an adaptive wish not to overload these. If this is the case, there should be links between working memory and good-enough processing. The next three sections discuss three sources of evidence that are consistent with this: individual differences in working memory, effects of task demands and cognitive load, and effects of ageing on sentence processing.

1.7. Syntactic processing and link to working memory

There has been a longstanding link between working memory and sentence processing (e.g., Baddeley & Hitch, 1974; Baddeley, 1986; Conrad & Hull, 1964; Just & Carpenter, 1992; for reviews, see Caplan & Waters, 2013; Gordon & Lowder, 2012; van Dyke & Johns, 2012). Of particular relevance to this thesis, performance on complex working memory tasks is associated with longer reading times on syntactically-complex sentences. Kemper et al. (2004) found that participants with low working memory capacity spent longer reading garden-path sentences, mostly because they made more regressions to earlier words in the sentence (although this is not always the case, and it is sometimes higher-span participants who take longer, after more careful analysis; Nicenboim, Logačev, Gattei, & Vasishth, 2016; von der Malsburg & Vasishth, 2013). There is also a recognised link between performance on complex span tasks and language comprehension more generally (e.g., Daneman & Merikle, 1996).

To examine the role of working memory, a common paradigm is to look at reading under dual-task conditions, to see how the second task affects the resources available for reading (depending on how "resources" is conceptualised). Early studies found that concurrent word list recall or subvocalisation impaired sentence processing (e.g., Baddeley, 1986; Foss & Cairns, 1970; Slowiaczek & Clifton 1980; Wanner &

Maratsos, 1978). More recently, a concurrent task that taps working memory has been shown to affect reading patterns and/or comprehension (Gordon, Hendrick, & Levine, 2002; Kemper & Herman, 2006; but cf. Caplan & Waters, 1999; Waters, Caplan, & Hildebrandt, 1987). When the secondary task involved recalling words that were similar to the texts being read, comprehension was impaired (Fedorenko, Gibson, & Rohde, 2006; Gordon et al., 2002) – consistent with cue-based parsing theories of memory. Dual-task conditions also led to longer reading times on disambiguating words in self-paced reading (Fedorenko et al., 2006; Kemper & Herman, 2006). The exact nature of the two tasks is critical for comparisons, and differences between these may underpin conflicting data and theories (Fedorenko et al., 2006). Nevertheless, these results provide further support for a link between memory resources and sentence processing.

The fact that sentence processing and working memory are associated in some way is therefore clear. However, the nature of the link between the two is much debated (e.g., Caplan & Waters, 1999, 2013; Gordon & Lowder, 2012; Just & Carpenter, 1992; Lewis & Vasishth, 2005; MacDonald & Christiansen, 2002; van Dyke & Johns, 2012). A key debate has been between three main theories explaining the link. These in turn argue a) that sentence processing enlists general working memory resources that are used also for other tasks (Just & Carpenter, 1992); b) that sentence processing requires separate, dedicated resources and is therefore unaffected by concurrent tasks (Waters & Caplan, 1996), or c) that the concept of working memory capacity should be eliminated, and seen instead as an emergent property of language experience and skills (e.g., Macdonald & Christiansen, 2002). An extensive discussion of all three positions is beyond the scope of this review, but the three families of theory have different perspectives on two key issues: the role of individual differences (as measured by working memory span tests), and whether dual-task conditions affect initial syntactic processing.

Just and Carpenter's research has supported the idea of a shared, domain-general working memory capacity across sentence processing and other tasks; as such, individual differences in working memory affect sentence processing ability, and a dual-task will reduce available resources for sentence processing, limiting performance (Daneman & Carpenter, 1980; Daneman & Merikle, 1996; Just & Carpenter, 1992; King & Just, 1991). Waters and Caplan's argument for separate resources is supported by findings that individual differences on a memory span task do not predict online sentence processing times (Waters & Caplan, 1996, 2001, 2003), and that dual-task conditions do not affect *online* sentence processing, as the two tasks rely on different resources (Waters et al., 1987). Waters and Caplan do not deny a role for working memory, but only in what they call *post-interpretive* processes, after initial syntactic processing has taken place.

Finally, stemming from parallel distributed processing, Macdonald and Christiansen (2002) reinterpreted previous evidence, claiming that results of better reading in people with higher "working memory capacities" are actually determined by greater reading experience and biological differences in connection strengths. Macdonald and Christiansen provide results from connectionist modelling to support this view. They argue that individual differences in standard working memory measures (such as the reading span task; Daneman & Carpenter, 1980) reflect language processing generally, and that performance varies as a result of language experience (combined with biological factors). It is not entirely clear how well the connectionist model can completely eliminate the concept of working memory as a merely emergent property though, as its initial parameters (the "biology") are set when the model is built, and the creator of the model therefore still has the power to alter "capacity" (Caplan & Waters, 2002).

While these three accounts all have evidence both to support and to question their approaches, there are alternative accounts available. It has been suggested that both

sentence processing and subsequent comprehension are underpinned by a cue-based retrieval mechanism (Lewis & Vasishth, 2005; Lewis, Vasishth, & van Dyke, 2006; van Dyke & Johns, 2012; van Dyke & Lewis, 2003; van Dyke & Mcelree, 2006), based on findings that working memory may in fact be very limited in capacity (e.g., Mcelree, 2006). On this view, difficulties with comprehension can result from interference between similar cues when trying to retrieve information. Linguistic items are envisaged as bundles of features: more reliable cues can lead to faster retrieval (Mcelree, Foraker, & Dyer, 2003), while more similar distractors are therefore thought to produce more interference. This would help to explain the results of the dual-task studies outlined above, especially where similarity between materials creates added interference. Furthermore, if working memory is framed as a person's susceptibility to interference (Bunting, Conway, & Heitz, 2004; Lustig, May, & Hasher, 2001), this provides an insight into the nature of individual differences.

1.8. Effects of task demands on sentence processing

There is a paucity of research into how task demands affect sentence processing, although the studies that have been conducted have identified a significant role for task effects on reading, and in some studies, on eye movements during reading. Three types of task effect are discussed briefly here: the reason participants are given for reading; the presence and type of questions asked; and the way in which a text is presented (for instance, in single sentences or in longer passages). Of course, the issue of dual-task conditions, discussed above, is also of relevance to this section.

1.8.1. Goal of reading

Several studies have manipulated task instructions in order to change the task goal. For example, are reading times or eye movements different if people are asked to read for comprehension, to proofread for errors, or simply to enjoy the material? The answer seems to be yes (e.g., Aaronson & Ferres, 1986; Daneman, Reingold, & Davidson, 1995; Kaakinen & Hyönä, 2010; Kaakinen, Lehtola, & Paattilammi, 2015; Rayner, Sereno & Raney, 1996; Schmalhofer & Glavanov, 1986; Schotter, Bicknell, Howard, Levy, & Rayner, 2014; van den Broek et al., 2001; White, Warrington, Mcgowan, & Paterson, 2015). For instance, van den Broek et al. (2001) found that when the goal was to read for study purposes, participants tended to produce more inferences and predictions (when asked to think aloud) than when the goal was to read for enjoyment; they also performed better on a later free recall task. Aaronson and Ferres (1986) modelled reading times using a number of predictor variables, and found that participants who were asked to retain sentences focused more attention on structural analysis; participants with a comprehension goal focused on the meaning of sentences. Kaakinen and Hyönä (2010) manipulated these task factors in an eye movement study where they found that word length and frequency effects on first pass durations were attenuated when the goal was comprehension compared to proofreading.

Another interesting study in this area was conducted by Lewis, Shvartsman and Singh (2013), who manipulated reading goals by offering different payoffs that either favoured accuracy or speed. They found that eye movements were adaptive to different payoffs, with longer fixation durations when accuracy was prioritised over speed – a result that mostly concurred with data from an adaptive eye movement control model. Taken together, these results suggest that reading behaviour is adaptable to task demands and the goal of reading, consistent with the good-enough approach.

1.8.2. Presence and nature of questions

Linked to the issue of task demands, another important manipulation is whether participants are asked questions – and what those questions tap. Eye movement studies of reading tend to ask fairly simple questions on a certain proportion of trials to ensure attention. However, beyond the work on good-enough processing (e.g., Christianson et al., 2001, 2006), little work in the area of syntactic ambiguity processing has asked questions that assess final representations of sentences. Importantly, if sentence processing is adaptive to current demands, the presence of questions by itself (and what those questions ask) should influence reading behaviour. A few studies have considered this issue (Kaakinen et al., 2015; McConkie, Rayner, & Wilson, 1973; Radach, Huestegge, & Reilly, 2008; Rothkopf & Billington, 1979; Swets et al., 2008; Wotschack & Kliegl, 2013), generally finding that the presence and type of questions do considerably change reading behaviour. Radach et al. (2008) found that the presence of questions resulted in significantly longer fixation durations, reflecting more careful reading in order to respond to questions. This resembles the changes in eye movements and comprehension that result from different reading goals (e.g., comprehension vs. proofreading). Wotschack and Kliegl (2013) extended this finding, demonstrating variability according to the nature of the question asked. Participants asked difficult questions read sentences for longer and made more regressions than participants given easy questions. Again, these results are demonstrative of an adaptive processing system that is sensitive to the task in hand.

1.8.3. Text presentation

Although the body of work is small, there is evidence that the length of the text to be read matters (Kuperman, Drieghe, Keullers, & Brysbaert, 2013; Radach et al., 2008;

Whitford & Titone, 2014; Wochna & Juhasz, 2013). Radach et al. (2008) compared single sentences presented in isolation to longer texts and found that first pass durations were longer for isolated sentences, but total reading durations were longer when reading passages. This suggests that longer texts induce an initially more superficial read that is supplemented by additional re-reading where necessary. Kuperman et al. (2013) found similar results when comparing gaze durations from one corpus in which isolated sentences were read, to gaze durations from a different corpus in which the reading material was longer passages. That said, the tasks were different in the two corpora (no questions were asked in the passage reading task from which the second corpus was collated), making a more direct comparison a topic for further research.

1.9. Effects of ageing on reading

Older adults (for example, adults aged over 65) show differences in their reading, both in terms of eye movements and comprehension – although it is a matter of debate whether this is a quantitative slowdown with age, or a qualitative shift in reading behaviour (Kliegl, Grabner, Rolfs, & Engbert, 2004). On easy sentences, older adults may show few differences from younger adults (Kliegl et al., 2004), but as sentences become more difficult, differences become more apparent. The effects of ageing on reading are reviewed in more detail in Chapter 4; here, I highlight why research with healthy older people might help inform our understanding of a good-enough, adaptive sentence processing approach.

Older adults' eye movements when reading are generally marked by an increase in the number of fixations, and an increase in the duration of those fixations (e.g., Kemper et al., 2004; Kemper & Liu, 2007; Kliegl et al., 2004; Payne, Grison, Gao, Christianson,

Morrow, & Stine-Morrow, 2014; Rayner, Castelhano, & Yang, 2009; Rayner, Reichle, Stroud, Williams, & Pollatsek, 2006; Stine-Morrow, Miller, & Hertzog, 2006). Older adults have also been shown to make more regressions (e.g., Kliegl et al., 2004), and greater "wrap-up" effects – the phenomenon of fixating for longer at syntactic boundaries, in order to finish processing a clause and move on (Payne & Stine-Morrow, 2012; Stine-Morrow, Shake, Miles, Lee, Gao, & McConkie, 2010). Furthermore, Rayner et al. (2006, 2009) found that older adults are more likely to skip words, but also to spend longer re-reading to compensate, a reading strategy referred to by the authors as *risky*. This risky strategy does not though leave older adults any more likely to mistake upcoming words for high-frequency neighbours, questioning to what extent this demonstrates a risky reliance on parafoveal lexical prediction (Warrington, White, & Paterson, 2016).

Beyond this greater tendency to skip words, a qualitative shift with age is also seen in higher-level processing (DeDe, 2014, 2015; von der Malsburg, Kliegl, & Vasishth, 2015). von der Malsburg et al. (2015) looked at older readers' eye movement scanpaths, and found that the regularity of scanpaths was less affected by syntactic complexity, compared to scanpaths of younger adults. This suggests that with increasing age, in-depth syntactic processing may be relied upon less, in favour of more semantic- or discourse-based inferences. Similarly, DeDe (2014) found that older adults are more prone to predicting upcoming syntactic structure using lexical cues; they are also more disrupted if the cues turn out to be misleading, suggesting that an overall strategy of using cues and heuristics is appropriate when it works, but more costly with age where it produces an incorrect interpretation.

Some studies have shown that older readers have poorer comprehension of sentences, especially when sentences contain temporary syntactic ambiguities

(Christianson et al., 2006; Kemtes & Kemper, 1997). Following their work with younger adults, Christianson et al. (2006) compared younger and older readers on comprehension of subject-object ambiguities. They found that older adults were more likely than younger adults to incorrectly accept the garden-path misinterpretation of optionally-transitive items *(While the man hunted the deer…),* but showed no difference in performance on questions following reflexive absolute transitive items *(While the father changed the baby…*). These results demonstrate that good-enough processing is more common in older adults, who seem more likely to adopt an incorrect interpretation if it has already been built and remains plausible. The absence of an effect in the reflexive absolute transitive items aligns with the view that older adults are more likely to rely on a semantic-based heuristic: the man hunting is compatible with the man hunting the deer, while in contrast, the father changing (himself) and changing the baby are unlikely to be occurring simultaneously. Christianson et al. (2006) argued that older adults are less able to reinstantiate the exact structure of sentences upon being asked the question. In line with this, comprehension performance was correlated with working memory capacity. This is consistent with evidence that older adults are poorer at recalling the surface structure of texts, even if they are less impaired with being able to make inferences from text (Radvansky, Copeland, & Zwaan, 2003; Radvansky, Zwaan, Curiel, & Copeland, 2001; Shake, Noh, & Stine-Morrow, 2009).

The link to working memory capacity in Christianson et al.'s (2006) study mirrors a general finding that older adults' reduced working memory capacity (and/or efficiency) may underpin differences in their reading behaviour. As with research on working memory generally, the causal assumption in this view does not enjoy universal support (cf. Caplan et al., 2011; DeDe et al., 2004; Kemper & Liu, 2007; Stine-Morrow et al., 2006a; Waters & Caplan, 2001, 2003). It is therefore important to consider whether

effects of ageing may underpin the age-related differences seen in reading. An initial consideration is differences in visual processing and oculomotor control. Owsley (2011) reviewed work in this area, and found that ageing has negative effects on several aspects of vision and visual processing, most notably in acuity and contrast sensitivity. This has been demonstrated in reading studies, with older adults more reliant on coarse, low spatial frequencies, and younger adults more able to make use of the fine detail of letters (Jordan, McGowan, & Paterson, 2014; McGowan, White, & Paterson, 2015; Paterson, McGowan, & Jordan, 2013). The additional effort required for visual processing may have an impact on older adults' reading strategies (DeDe & Flax, 2016; Gao, Levinthal, & Stine-Morrow, 2012; Tun, McCoy, & Wingfield, 2009).

It is also possible that differences with ageing are due to other cognitive factors, which may explain away any link between age-related reading changes and working memory. These factors include a decline in the efficiency of inhibitory processes (Carlson, Hasher, Connelly, & Zacks, 1995; Hamm & Hasher, 1992; Zacks & Hasher, 1997; but see also Kemper & McDowd, 2006; Kemper, McDowd, Metcalf, & Liu, 2008). On this view, differences and difficulties with reading are linked to a poorer ability to block irrelevant information and hence to maintain comprehension. This body of research has identified that older adults show more difficulty with inhibiting distractors (Carlson et al., 1995; Connelly, Hasher, & Zacks, 1991; Healey, Hasher, & Campbell, 2013; Zacks & Hasher, 1997) – although if working memory is interpreted in terms of a cue-based retrieval, and not a capacity-based system (Gordon & Lowder, 2012; van Dyke & Johns, 2012), separate roles of working memory and inhibition may not be so distinct.

There are several other cognitive changes associated with ageing. These are not all negative: older adults have clearly acquired more reading experience, knowledge and vocabulary – and these may shift their reading strategies accordingly (Macdonald &

Christiansen, 2002; Rayner et al., 2006; Salthouse, 2003; Stine-Morrow, Shake, Miles, & Noh, 2006; Verhaeghen, 2003). However, older adults also have more difficulty adapting their behaviour, in response to deadlines (Laver, 2000), or with prioritising speed over accuracy (e.g., Mitzner, Touron, Rogers, & Hertzog, 2010; Stine-Morrow et al., 2006a). Stine-Morrow and colleagues have described this as a reduced efficiency in re-allocating resources to different levels of reading, for instance, surface processing vs. deeper, inference-making. In turn, this may make older adults less adaptive to task demands, and indeed to the cognitive changes that come with ageing (Smiler, Gagne, & Stine-Morrow, 2003; Stine-Morrow, Gagne, Morrow, & DeWall, 2004; Stine-Morrow & Miller, 1999; Stine-Morrow, Miller, Gagne, & Hertzog, 2008; Stine-Morrow et al., 2006a). This could explain why age-related difficulties only tend to appear when sentences are more complicated: most sentences are fairly simple, and so ageing effects are not as apparent. With difficult sentences, more adaptation is required – and it is here where differences with reading can be observed (Salthouse, 2004).

While changes in inhibition, vocabulary and language processing seem to play a role, many of the results still tally with the general finding that working memory is poorer in older adults (e.g., Bopp & Verhaeghen, 2005; DeDe et al., 2004; Gordon, Lowder, & Hoedemaker, 2016; Hull, Martin, Beier, Lane, & Hamilton, 2008; Just & Carpenter, 1992; Payne & Stine-Morrow, 2012; Payne et al., 2014; Stine-Morrow & Miller, 2009; Waters & Caplan, 2003), and there is still considerable support for a link to reading. There is evidence that working memory is linked to changes in eye movements, with ageing effects on eye movements during reading being dependent on working memory capacity (Kemper & Liu, 2007). Working memory may also protect against ageing effects: older adults with high working memory capacity seem to show less difficulty in resource allocation (e.g., Payne et al., 2014; Stine-Morrow, Morrow, & Leno, 2002).

There is disagreement about the exact role of working memory on online sentence processing, even if it has a role in comprehension (Caplan & Waters, 1999; Caplan et al., 2011; DeDe et al., 2004; Waters & Caplan, 2001, 2003). In Chapter 4, the distinctions between these views are discussed in more detail; for now, though, there is at least some evidence for an association between working memory and differences with ageing.

The link to working memory has also been investigated by using dual-task designs to investigate age-related shifts in processing or in reading strategies. Results have been mixed, with some evidence of an age-related decline (e.g., Goethe, Oberauer, & Kliegl, 2007; Kemper & Herman, 2006; Riby, Perfect, & Stollery, 2004), but conflicting results of no such decline (Baddeley, Chincotta, & Adlam, 2001; Brebion, 2003), or even an improvement with ageing (Brebion, 2001; Daneman, Hannon, & Burton, 2006; Kemper et al., 2004; Kemper, Herman, & Lian, 2003; Kemper, Schmalzried, Herman, Leedahl, & Mohankumar, 2009; Stine-Morrow, Miller, & Leno, 2001). As discussed later in Chapter 4, the variable results are likely to stem from differences in methodology, but the overall conclusion is that older adults are not necessarily impaired at dual-task studies. Instead, it may be that older adults take a conservative approach to cognitive tasks, and aim for accuracy over speed (Daneman et al., 2006; Rabbitt, 1979; Starns & Ratcliff, 2010). Most of the time, the use of heuristics will ensure this without considerable cost; however, as seen in Christianson et al. (2006) and other studies, heuristics may not always produce an accurate result, and age-related limitations may make excessive re-reading or reprocessing too onerous. In this sense, it could be argued that the decline in self-regulated language processing seen in older adults is *good-enough* (Stine-Morrow & Miller, 2009): a generally successful strategy is overstretched by syntactically complex sentences, leading to more comprehension errors.

1.10. Outstanding questions and scope of this thesis

This review has presented research demonstrating that contrary to the predictions of early sentence processing models, comprehension of syntactically ambiguous texts is not always veridical, and syntactic processing is not always completed. Furthermore, the fidelity of representations to the actual sentence content is influenced by factors such as individual differences in working memory (including due to ageing), task demands, and dual-task conditions. Superficial reading could be an adaptive strategy in the face of demands on cognitive resources. There are still several outstanding questions, which are to be addressed in this thesis. This section considers four such questions.

1.10.1. What is meant by *good enough* syntactic processing?

There are several versions of the good-enough account (e.g., Christianson, 2016; Christianson et al., 2001; Ferreira & Patson, 2007; Slattery et al., 2013; Swets et al., 2008), and it remains unclear what is actually meant by sentence processing being *good enough* (Christianson, 2016; Traxler, 2014). The most supported version is that when an initial misinterpretation of a sentence is built, it is not sufficiently pruned and lingers to disrupt comprehension (Christianson et al., 2001; Slattery et al., 2013). But why is this the case, and what influences the extent to which it lingers?

The extent of lingering varies based on verb type, the length of sentences, and individual differences in working memory capacity (Christianson et al., 2001, 2006). Furthermore, there is evidence that good-enough processing in the eye movement record is affected by the plausibility of the initial misinterpretation (Slattery et al., 2013). There are a number of factors that may also be of interest – several of which feed into the questions considered below. Importantly, it is not clear how long misinterpretations

linger for. Slattery et al. (2013) found evidence on eye movements in a second sentence, but does the extent of lingering depend on the length of passages, and do misinterpretations linger in longer-term representations of text? A greater understanding of these issues may also shape our understanding of the architecture of the syntactic processor, as any theory would have to account for misinterpretations continuing to disrupt processing after they are ostensibly abandoned.

1.10.2. What is the effect of task demands on syntactic ambiguity processing?

As discussed earlier, research into task demands is relatively sparse, despite their perceived influence on reading. It is even less clear how task demands affect good-enough processing of syntactically ambiguous sentences. How do different goals, methodologies and task demands affect the extent to which misinterpretations linger? Swets et al. (2008) found that readers show strategic underspecification of syntactic representations as a function of the difficulty of questions they are expecting (or instead of this being underspecification, it may be that no structure is specified at all, with questions simply answered with a guess; Logačev & Vasishth, 2016b). Swets et al. used a sentence construction where reanalysis was not necessary to form a consistent parse. Plausibly, the pattern of results may be different if reanalysis is obligatory (as discussed earlier; cf. von der Malsburg & Vasishth, 2013). Furthermore, if good-enough processing is adaptive to working memory constraints, to what extent will it be affected by cognitive load, imposed by a concurrent task?

The issue of task demands may also extend to manipulating the length of passages being read (Radach et al., 2008). This may also be visible in the eye movement record: Slattery et al. (2013) did not find a garden-path effect on first pass durations when participants read short passages, even though this effect is prevalent for single sentence

presentation (Clifton et al., 2007). There is therefore significant interest in exploring good-enough effects while presenting syntactically ambiguous sentences in differing formats.

1.10.3. How do individual differences affect syntactic ambiguity processing?

Another consideration is how much good-enough effects are subject to individual differences between groups, and even between individuals. This may be differences in working memory capacity, given the aforementioned link to syntactic processing difficulties (Christianson et al., 2006; King & Just, 1991; Nicenboim et al., 2016; but cf. Caplan & Waters, 1999; MacDonald & Christiansen, 2002), or to differences in cognitive control (January, Trueswell, & Thompson-Schill, 2009; Novick, Trueswell, & Thompson-Schill, 2005). When considering the architecture of syntactic processing, the focus is usually on finding similarities between readers. However, it may be of just as much help to consider differences between people (Farmer, Misyak & Christiansen, 2012), and using this information to test hypotheses from possible theoretical architectures. It is therefore important to look at individual differences between participants, and to compare findings from theoretically meaningful cross-population comparisons. To this end, I will investigate whether there are differences in eye movements, comprehension and dual-task conditions between older and younger adults, based on the age-related differences in reading behaviour reviewed earlier. I will also look into individual differences between younger adults, to consider the various approaches that may be taken after encountering a syntactic ambiguity.

1.10.4. What are the links between eye movements, comprehension, and domain-general resources?

There is extensive evidence that eye-tracking data can inform on questions about the nature of reading processes and the representations that support reading behaviour. Critically, eye movement data can reveal real-time processing of syntactically ambiguous or complex sentences. Lacking however, is a comparison of eye movement patterns to eventual comprehension. Many studies have looked at one of these, focusing on either online processing as tapped by eye movements, or offline comprehension following questions or recall. Surprisingly few have considered both in the same experiment. A direct comparison between both measures should provide a better insight into syntactic processing, from initial detection of ambiguities, to eventual understanding (or lack of understanding).

This also links to a wider question of whether cognitive resources more broadly impact on reading. The debate in the literature has already been set out, but to understand the link better requires simultaneous measurement of eye movements, eventual comprehension, and of individual differences. The disruption to syntactic processing induced by dual-task conditions (Fedorenko et al., 2006; Gordon et al., 2002; Kemper & Herman, 2006), and links to working memory and executive control processes (Novick, Hussey, Teubner-Rhodes, Harbison, & Bunting, 2014; von der Malsburg & Vasishth, 2013) indicate that the extent to which misinterpretations are permitted to linger may be influenced by general processing constraints – but further testing of this would help to elucidate these links.

1.11. Summary of thesis

This thesis sets out to answer the four questions above: what it means for processing to be *good enough*, how this is influenced by task demands and individual differences, and how good-enough processing manifests itself in both eye movements and comprehension. Of particular interest is how good-enough processing changes while under different levels of cognitive load – either because of the nature of the reading task, or because of demands from a concurrent task.

Six experiments looked at eye movements and comprehension concurrently, to bring together two areas of research that have often been conducted separately. All six experiments follow the same fundamental paradigm. Participants read garden-path sentences such as *While the father changed(,) the baby played with its toys*, while their eye movements were monitored. After reading sentences, comprehension of these sentences was tested, usually with a question such as *Did the father change the baby?* This was to investigate whether the garden-path misinterpretation had lingered.

Chapter 2 focuses on the effects of load, by manipulating whether participants had to complete a concurrent n-back task. The dual-task condition tapped verbal working memory resources, to see how this affected online sentence processing (as measured by eye movements), and eventual comprehension. Experiment 1 used single garden-path sentences, while Experiment 2 embedded the sentences in a two-sentence passage. By having more to read, would participants change their strategy towards the reading and/or n-back tasks? More generally, Experiment 2 set out to investigate differences in reading short passages, compared to isolated sentences, to see whether the passage format would lead to more superficial reading of the garden-path sentence, and affect comprehension. Finally, in Experiment 3, the difficulty of the n-back task was raised, to assess whether it

was the mere presence of a dual-task that could lead to disruption in reading, or whether the size of the load also affected the rate of good-enough processing.

Chapter 3 focuses on the role of task demands. In Experiment 4, the text format was changed once more, so garden-path sentences were now embedded in a longer, 4-sentence passage. There was also a manipulation of bias in the sentence following the garden-path sentence, which was either neutral about the garden-path misinterpretation, or which perpetuated the error. The aim of this experiment was to look at whether the longer passages and the bias manipulation led to an increase in good-enough errors, and the effects of these factors on eye movements. Experiment 5 looked at what happens if the comprehension questions were removed, and comprehension was instead tested after a delay using a surprise paraphrase recognition task. The first aim was to see whether eye movements would show less thorough reading, as participants were not necessarily expecting to have their comprehension tested. The second was to look at good-enough effects after a 10 minute delay: would misinterpretations linger over this period, and cause disruption in the ability to distinguish genuine paraphrases of what had been read, from foils? A reading span measure was also used to see how individual differences in memory affected both eye movements and recognition task performance.

Chapter 4 also focuses on the topic of individual differences, by comparing the younger adults in Experiment 1 to older adults (aged 65+) in Experiment 6. While previous research has considered how ageing affects good-enough comprehension errors (Christianson et al., 2006), and eye movements (e.g., Kliegl et al., 2004), less work has considered both concurrently. Furthermore, I was interested in the impact of the dual-task condition on older adults, and the insight this would provide into the effects of ageing on resource allocation, as well as on sentence processing.

Chapter 5 looks at exploratory analyses of entire scanpaths, to see whether the effects that were seen (or, in some cases, not seen) in conventional eye movement measures are apparent across eye movements more generally. Building on work by von der Malsburg and Vasishth (2011, 2013), the analyses reconsider the data from Experiments 2, 3 and 5 to see how syntactic ambiguity, the concurrent n-back task, and the presence or absence of questions affects eye movement patterns across scanpaths. These analyses are exploratory, but offer an insight into individual differences in reading behaviour, opening up potential for future, more focused research.

Finally, Chapter 6 summarises the results presented in the thesis, and discusses them in the context of the good-enough framework to sentence processing. In this General Discussion, the four questions outlined above are reconsidered in the light of these results, allowing a more detailed picture of what it means for sentence processing to be *good enough*. This chapter also considers the wider theoretical and methodological implications of this research, as well as future directions that stem from the experiments presented in this thesis.

Chapter 2

Effects of load on syntactic ambiguity processing

As discussed in Chapter 1, people tend to slow down and re-read when sentences containing temporary syntactic ambiguities are initially misinterpreted (Clifton et al., 2007) but this does not necessarily mean they will correctly answer comprehension questions about these sentences (Christianson et al., 2001). This has been attributed to a "good-enough" approach to syntactic processing (e.g., Ferreira & Patson, 2007), according to which we process sentences only to the extent necessary to achieve our current task. Comprehension may not always be veridical if it would require extensive effort to fully process the text. In this chapter I focus on how extrinsic load would affect comprehension and eye movements. Three experiments set out to elucidate what it means for processing to be good enough for a given set of task demands, how good-enough processing of syntactically complex sentences affects both eye movements and comprehension accuracy, and the role of working memory in syntactic processing.

2.1. Background

2.1.1. Good-enough processing

Chapter 1 highlighted the considerable body of research that has been conducted on "garden-path" sentences containing a syntactic ambiguity. There are still gaps in our understanding of how syntactically ambiguous sentences are processed, and what happens to the competing interpretations during processing and subsequent comprehension. Serial "garden-path" models (Ferreira & Clifton, 1986; Frazier & Fodor, 1978; Frazier & Rayner, 1982) argue that only one interpretation is considered at any point; the processing difficulty in garden-path sentences arises from the need to return and reanalyse the

sentence. This results in the "garden-path effect" of longer reading times (e.g., Clifton et al., 2003, 2007; Rayner, 1998; but see also e.g., Altmann, 1994; Mitchell et al., 2008).

However, garden-path models presume that the extra processing effort is rewarded by a complete parsing of the sentence. Indeed, earlier work on syntactic ambiguity resolution rarely asked questions that would explore whether participants had completed reanalysis, and comprehended the sentence correctly (Christianson, 2016). This view has not gone unchallenged, however. More recent findings, introduced in Chapter 1, show that the product of syntactic analysis can be incomplete or incompletely processed (Christianson et al., 2001; Kaschak & Glenberg, 2004; Swets et al., 2008; van Gompel et al., 2006). To reiterate the key finding, Christianson et al. (2001) found poorer comprehension on questions such as (2.1c) for garden-path sentences such as (2.1a), compared to sentences that were disambiguating by a comma, such as (2.1b):

(2.1a) While the father changed the baby that was cuddly played with its toys.
(2.1b) While the father changed, the baby that was cuddly played with its toys.
(2.1c) Did the father change the baby?

Ferreira and colleagues' *good-enough processing* account (Christianson et al., 2001, 2006; Ferreira & Patson, 2007; Slattery et al., 2013; Swets et al., 2008) argued that we often settle on heuristic-based interpretations that may not be veridical (for example, they may contain traces of abandoned misinterpretations), but that are nevertheless usually good enough for the task at hand. This may be particularly important to avoid costly re-reading and reprocessing where resources are otherwise occupied, or when task demands are high.

It is worth repeating that Ferreira, Christianson and colleagues' good-enough framework is not the only theory to explain how language processing may not always be veridical to the exact inputs received. Other examples, described in more detail in Chapter 1, include noisy-channel models (e.g., Gibson et al., 2013; Levy, 2008; Levy et al., 2009), the Now-or-Never bottleneck hypothesis (Christiansen & Chater, 2016), and van den Broek et al. (2001) concept of *standards of coherence*. In this chapter, the focus will remain on the good-enough framework, given the similarity between the experiments presented here and previous work focusing on the good-enough approach (Christianson et al., 2001, 2006; Patson et al., 2009; Slattery et al., 2013). Nevertheless, I will return to these alternative approaches in more detail in Chapter 6.

2.1.2. Links to working memory

Ferreira and Patson (2007) argued that good-enough processing results from limitations in cognitive resources and an adaptive wish not to overload these. If this is the case, there should be links between working memory and/or executive control, and good-enough processing. As discussed in Chapter 1, three sources of evidence are consistent with this. First, performance on complex working memory tasks has been linked to comprehension generally (Daneman & Merikle, 1996), but specifically to differences when reading syntactically-complex sentences (Kemper et al., 2004; but cf. Caplan & Waters, 2002; Traxler, Long, Tooley, Johns, Zirnstein, & Jonathan, 2012). Kemper et al. (2004) found that participants with low working memory capacity spent longer reading garden-path sentences, mostly due to more regressions. von der Malsburg and Vasishth (2013) also found that memory span was linked to increased re-reading in critical regions of sentences containing an attachment ambiguity, suggesting that lower-span participants were more likely to leave the sentence underspecified.

Furthermore, Novick et al. (2014) looked at links to the n-back task, in which participants are presented with a stream of stimuli and are asked to remember the last *n* items (with *n* usually between 1 and 4), before being periodically tested for recall or recognition of the n^{th}-to-last item. The task therefore requires storage of recent items, deletion of older items, and recall of the correct order in which items were presented. Novick et al. found that participants who improved on an n-back task with training improved commensurately in their comprehension accuracy on questions following garden-path sentences such as (2.1a). The "responders" to n-back training also showed significantly reduced go-past reading durations when reading syntactically ambiguous sentences. Novick et al. concluded that the n-back training improved cognitive control mechanisms, with the benefits observed in syntactic ambiguity resolution. Training on other tasks was not found to have a similar effect, suggesting that mechanisms guiding n-back task performance are linked to those guiding syntactic ambiguity resolution.

Second, in dual-task paradigms, a concurrent task tapping working memory has been shown to affect reading and/or comprehension (Gordon et al., 2002; Kemper & Herman, 2006; but cf. Caplan & Waters, 1999; Waters et al., 1987). When the secondary task involved recalling words that are similar to the texts being read, comprehension was impaired (Fedorenko et al., 2006; Gordon et al., 2002). Dual-task conditions also led to longer self-paced reading times on disambiguating words (Fedorenko et al., 2006; Kemper & Herman, 2006).

Third, age-related declines in working memory have been linked to stronger garden-path effects in older readers. Compared with younger adults, Christianson et al. (2006) found that older adults were more likely to answer *Yes* incorrectly to comprehension questions tapping the initial misanalysis (such as (2.1c) above), despite reading these sentences for longer. More generally, working memory span correlated with

comprehension performance, suggesting that good-enough errors are associated with limitations in working memory. This pattern of longer overall reading times, longer fixation durations and more regressions in older readers is similar to results in younger readers under dual-task conditions (Kemper & Herman, 2006), and in younger participants with low working memory spans (Kemper et al., 2004). These differences with ageing have been linked to differences in lexical processing (Rayner et al., 2006, 2009), but more importantly, to older readers' reduced engagement in syntactic processing and reanalysis, and reliance on heuristics and world knowledge (Christianson et al., 2006; DeDe, 2014; von der Malsburg et al., 2015). This is discussed further in Chapter 4. Taken together, these studies demonstrate that reductions in working memory resources are linked to increased good-enough processing. Superficial initial reading could be an adaptive strategy, in the face of demands on cognitive resources.

While it is generally accepted that there is a link between working memory and sentence processing, there is less consensus on how this link is conceptualised (e.g., Caplan & Waters, 1999, 2013; Gordon & Lowder, 2012; Just & Carpenter, 1992; Lewis & Vasishth, 2005; MacDonald & Christiansen, 2002; van Dyke & Johns, 2012). While too extensive to repeat in detail here, relevant is the theory that a cue-based retrieval mechanism underpins sentence processing (van Dyke & Johns, 2012; van Dyke & Lewis, 2003; van Dyke & Mcelree, 2006). The cue-based retrieval approach would attribute comprehension errors during syntactic ambiguity resolution to interference between similar cues during retrieval. This may help to explain the results of both dual-task studies, and how individual differences in working memory affect sentence processing (Bunting et al., 2004). Novick et al. (2014) noted that improvements in cognitive control processes tapped by their n-back training may result from improvements in retrieval processes, as conceptualised in cue-based theories of memory.

The debate about the role of working memory in online sentence processing is not concluded. For instance, a self-paced reading experiment found that while individual differences in working memory had an overall effect on reading times (regardless of syntactic ambiguity), working memory did not moderate the size of the reading time increase for syntactically-ambiguous sentences (Evans, Caplan, Ostrowski, Michaud, Guarino, & Waters, 2015). Evans et al. concluded that the influence of working memory is limited to post-comprehension processing (Caplan et al., 2011). Regardless of how one interprets the association between sentence processing and working memory, there is considerable evidence that good-enough processing is more apparent under conditions of reduced working memory resources.

Of course, the link to working memory depends on the answer to the wider question: what does it mean for sentence processing to be merely *good enough*? One view is that syntactic representations are underspecified (e.g., Swets et al., 2008). Alternatively, representations may be intact, but the process of pruning previous misinterpretations may not have been adequately completed. This was seen in Slattery et al. (2013): as outlined in Chapter 1, they found that for plausible[12] items such as (2.2), participants spent longer reading 'himself before' in the comma-absent condition:

(2.2) While the boy washed(,) the dog that was hot hid behind the trees. The boy always washed himself before eating lunch.

The longer looking times indicate surprise at the reflexive term; the representation that the boy is washing the dog (not himself) appears to have lingered. Slattery et al. (2013)

[12] Here, plausible refers to whether it is believable that the second noun phrase could be attached to the verb (for example, in *While Robin shaved the moustache/dome just kept growing*, it is plausible that Robin could be shaving the moustache, but unbelievable that he would be shaving the dome).

argued that this misinterpretation was not completely pruned before moving to the second sentence. They attributed this to the process of reanalysis not having been completed. Slattery et al. also suggested that individual differences in working memory capacity may explain this finding, if those with higher working memory capacities can bind interpretations and their surface structures more accurately (Christianson et al., 2006).

2.1.3. The role of task demands

If good-enough processing is related to general constraints, task demands should influence reading strategies – and should presumably influence eye movements while reading (Karimi & Ferreira, 2016; von der Malsburg & Vasishth, 2013). A person who is reading in order to fully comprehend a text and who is without distraction may read a text more carefully (i.e., less superficially) than a reader seeking gist, or a reader facing a concurrent cognitive load. The task could also affect whether they will reach an accurate final specification of syntactic structure (having pruned the initial misinterpretation), or if they arrive at a good-enough specification with lingering misinterpretations. As discussed in Chapter 1, there is considerable evidence that eye movement patterns can differ as a result of task demands – for instance, if participants are asked to read for comprehension, or to proofread (Kaakinen & Hyönä, 2007, 2010). There is also evidence from Swets et al. (2008) that the likelihood of underspecification is influenced by question difficulty (although this study used self-paced reading, which may also have influenced behaviour). The issue of how questions affect reading strategy is elaborated on in Chapter 3.

Task demands are not limited to the explicit instructions given to participants. Faced with a passage of text, the first sentence of that passage might be read differently to if the same sentence is read on its own. It might be read more superficially, with an expectation that upcoming text will provide information to aid comprehension. Consistent

with this, Slattery et al. (2013) found no evidence of a first pass garden-path effect on the disambiguating region of their passages – in (2.2), this region is the word *hid* and the words that immediately follow it[13]. In contrast, a first pass effect is highly replicated when the sentences are presented in isolation (cf. Clifton et al., 2007). Similarly, Radach et al. (2008) found different patterns of eye movements when participants read passages compared to single sentences, with shorter first pass reading durations but longer total reading durations on critical words. They also found poorer comprehension on questions following sentences embedded in passages, compared to questions following sentences. These findings highlight a pressing need to explore syntactic processing using a range of stimuli, and under a range of task demands.

2.1.4. Measuring eye movements

An additional point, exemplified by Radach et al. (2008) and Slattery et al. (2013), is the utility of eye movement methodology for questions about the nature of reading processes and the representations that support reading behaviour. Eye-tracking looks at fixation durations, and forward and regressive eye movements, offering a more detailed insight into online processing than methodology such as self-paced reading that has characterised most work exploring good-enough effects to date. The distinction between *early* measures of reading such as first pass duration, and *late* measures such as total reading duration and second pass duration (Clifton et al., 2007; Rayner, 1998; see also Vasishth et al., 2013) is useful since later measures more readily reveal reprocessing, which may reflect a need to recover from earlier good-enough processing. Self-paced reading is also thought to impose greater demands on working memory than naturalistic

[13] Similarly, Ferreira and Clifton (1986) only found a marginal first pass effect when presenting another type of syntactically complex sentences (reduced relative clauses) in passages.

reading (Gordon et al., 2006). A similar demand is placed on participants if they are instructed not to re-read material, as was the procedure in Christianson et al.'s (2001, 2006) experiments[14]. This amplifies demands when asked comprehension questions that probe the initial misanalysis. In turn, this might lead to an artificial and perhaps incorrect index of good-enough effects. Eye-tracking studies allowing unrestricted reading are therefore preferable. However, as Christianson et al. (2001) note, most eye-tracking studies (including Slattery et al., 2013) have failed to ask the sort of comprehension questions needed to allow links to be made between online reading behaviour and eventual comprehension. Furthermore, the fundamental assumption that eye movements index ambiguity detection and reanalysis is questioned by evidence that regressions may serve purposes other than simply reanalysis (Christianson et al., 2016; Mitchell et al., 2008). Adding questions therefore provides a clearer index of post-comprehension processes.

2.1.5. The current experiments

It is apparent that people can fail to make a complete syntactic parse of sentences, relying instead on good-enough, heuristic-based interpretations. This varies based on individual differences, the nature of the text being read, and external task demands. Lacking however is investigation of these factors in a design that allows detailed moment-by-moment analysis of reading behaviour alongside comprehension questions that probe the output of that behaviour – there has been little attempt to explore eye

[14] Christianson et al. (2001, 2006) presented sentences in their entirety, except in Experiment 1A of the 2001 study (which used rapid serial visual presentation [RSVP]). This means that participants *could* have re-read the sentence if they wanted (there was no eye-tracking and so this can't be assessed). However the methods section for the 2001 paper states that participants were instructed to read the sentence and then press the button to terminate the trial "without rereading the sentence". While there was no difference in results between the RSVP in Experiment 1A, and the sentence presentation in Experiment 1B (indicating that insufficient reading time was not the cause of good-enough effects), this instruction not to re-read may have influenced reading patterns.

movements and comprehension concurrently (except e.g., Kemper et al., 2004; Novick et al. 2014; von der Malsburg & Vasishth, 2013). This is unfortunate as such evidence is critical to inform what is meant by good-enough processing, and how this relates to syntactic ambiguity resolution, and to sentence processing more generally.

Three experiments set out to remedy this by (a) recording people's eye movements as they read syntactically ambiguous sentences under conditions of varying working memory load and (b) addressing performance on comprehension questions relating to the ambiguity. Building on work by Christianson and colleagues (e.g., Christianson et al., 2001, 2006; Slattery et al., 2013), garden-path sentences were compared to unambiguous sentences containing a comma. The use of stimuli with subject-object ambiguities allowed a comparison to von der Malsburg & Vasishth's (2013) research using attachment ambiguities such as (2.3).

(2.3) The maid of the princess who scratched herself in public was terribly humiliated.

In sentences such as (2.3), there is ambiguity in whether the central clause should be attached to *the maid*, or *the princess*. However, there is no obligation to reanalyse the ambiguity – once a reader has one interpretation, it is syntactically valid and can be maintained. In contrast, upon reaching *played* in *While the father changed the baby that was cuddly played with its toys*, a reader is forced to instigate reanalysis (for more discussion, see von der Malsburg & Vasishth, 2013; see also Frazier & Clifton, 1996; Swets et al., 2008). There is therefore reason to predict a different pattern of results with the sentence structure used in these experiments (for instance, larger effects in re-reading), warranting further investigation.

To explore the impact of processing demands, the reading task was combined with a between-subjects manipulation of cognitive load (cf.Fedorenko et al., 2006; Gordon et

al., 2002). A concurrent n-back task was chosen as the load manipulation for two reasons. First, Novick et al. (2014) demonstrated that n-back training was the most effective predictor of syntactic ambiguity processing ability. This therefore makes the n-back task the most appropriate concurrent task to cause disruption. Second and more generally, n-back tasks have face validity as a working memory task (cf. Jaeggi, Buschkuehl, Perrig, & Meier, 2010; Kane, Conway, Miura, & Colflesh, 2007). The n-back task required verbal working memory resources both as participants read and processed the sentences, and subsequently as they answered the comprehension questions. This enabled me to tap the impact of increased load on both online and offline processing, testing the assertion that load will increase the rate at which misinterpretations linger due to incomplete reanalysis. In this sense, load may act like lower working memory capacity in von der Malsburg & Vasishth (2013), creating a greater tendency to read less carefully on first pass.

However, added load should not simply encourage participants to read sentences quickly and leave them underspecified: the presence of comprehension questions in this study makes this unlikely, as it is important to aim for comprehension accuracy as well. This is especially important for the sentence construction used in these experiments: an initial misinterpretation *must* at least partly be reanalysed in order to attain consistent comprehension; this is unlike, for example, attachment ambiguities, where reanalysis is optional and underspecification will still produce a viable interpretation (cf. Swets et al., 2013; von der Malsburg & Vasishth, 2013). Previous research has not combined a manipulation of extrinsic load with eye-tracking to examine the effects of task demands on both comprehension and eye-movements during sentence processing.

Experiment 1 compared eye movements and comprehension of garden path sentences under a concurrent extrinsic load to a no-load condition. To further investigate

how task demands determine good-enough processing, Experiments 2 and 3 used this paradigm for the reading of passages, similarly to those used by Slattery et al. (2013). This allowed examination of lingering effects of abandoned misinterpretations on eye movements. Experiment 3 also increased the complexity of the extrinsic task to determine whether it is the mere presence of extrinsic load – or the size of that load – that affects processing. This has been seen previously using a manipulation of visual noise: while moderate noise had only a slight effect on young adults' reading, a higher level of noise led to a substantial change in reading patterns (cf. Gao et al., 2012; Gao, Stine-Morrow, Noh, & Eskew, 2011).

Good-enough processing was predicted to be influenced by the presence of the concurrent task, and by the demands of that task. This led to the prediction that adding load would produce more superficial reading followed by a need for extensive re-reading, similarly to eye movement patterns seen in older readers. In turn, I expected there to be more comprehension errors indicative of good-enough processing with load, again like older readers. These effects were predicted to be stronger in Experiment 2, if passages serve to induce more superficial reading, and stronger still when increasing load in Experiment 3. Finally, I predicted that even no-load participants would read sentences more superficially in Experiment 2, where the sentences were embedded in a passage rather than presented in isolation. This should manifest itself in the eye movement record, and also as more good-enough comprehension question errors.

2.2. Experiment 1

Eye movements were monitored as participants read single sentences that either were syntactically ambiguous, or contained a disambiguating comma. After each sentence, they answered a question that explored the temporary ambiguity; questions were designed to tap the good-enough effects seen in previous non-eye-tracking experiments (e.g., Christianson et al., 2001). One group of participants completed the reading task alongside a concurrent 2-back memory task, designed to engage verbal working memory resources while still allowing participants to read the sentences.

No-load participants (i.e., those without the dual-task) were expected to show similar effects to previous research: the classic eye movement signature of longer reading durations on the disambiguating region for garden-path sentences compared to comma sentences, and poorer question accuracy for garden-path sentences. For the first time, this experiment allowed the combination of an extrinsic load manipulation and a direct comparison of skilled readers' eye movement behaviour and comprehension when questions directly tapped the initial misanalysis. The secondary task was expected to produce the superficial first pass reading and subsequent re-reading similar to those seen in older readers (DeDe, 2014; Kliegl et al., 2004); this prediction is considered further in Chapter 4. An interaction with load should be stronger in later eye movement measures since these measures tap reprocessing of a text, instigated when insufficiencies in initial processing demand additional or more systematic processing to parse the text (Karimi & Ferreira, 2016).

If additional load disrupts processing, this might reduce accuracy on questions that follow ambiguous sentences – as seen in older readers (Christianson et al., 2006), and some dual-task studies (Fedorenko et al., 2006). The links between n-back performance and syntactic ambiguity resolution in Novick et al. (2014) suggest that similar

mechanisms underpin both. Kemper and Herman (2006) however failed to find an effect of a secondary task on question accuracy, albeit in a task where the questions were not tapping the initial misanalysis directly. The to-be-recalled words used in this n-back task were also not related to the content of sentences in the reading task; where the items in the two tasks of a dual-task paradigm are unrelated, there is less interference, and in turn, a less detrimental effect on sentence comprehension (Gordon et al., 2002). Furthermore, if participants under load show added reprocessing (in the eye movement record) as predicted, they may use this extra processing to overcome the additional difficulty. Given this set of findings, a small but significant decline in accuracy with load was predicted.

2.2.1. Method

2.2.1.1. Participants. Fifty-six students or recent graduates of the University of Oxford (all aged 18-30, mean age = 20.4 years; 45 female) took part in this experiment. It was requested that all had normal or corrected-to-normal vision, no history of visual or reading impairments, and were not simultaneous bilinguals (i.e., English had to be their only first language). Participants were sourced either from current psychology undergraduates as part of a research participation scheme (n = 42), or via emails and posters to university students (n = 14). For Experiment 1 therefore, most participants were psychology students; the remaining 14 came from across humanities and science disciplines. All received £5 or course credits as compensation. In terms of sample size, the intention was to source approximately thirty participants per condition (both in Experiment 1 and later experiments), in line with the average sample size in similar studies (Christianson et al., 2001, 2006; Slattery et al., 2013). As outlined in the Results section below, analyses were conducted using linear mixed effects modelling; unlike t tests and analyses of variance, there is no simple or agreed-upon method of conducting a

power analysis for these analyses, especially where interactions are predicted. The choice of a comparable sample size to similar studies was therefore considered the most appropriate way to ensure sufficient power to detect effects.

Participants were allocated into two groups. One group completed the experiment with a concurrent working memory task (*2-back* group, $n = 28$) and the other participated in an otherwise identical condition but without an additional task (*no-load* group). All participants gave informed consent in accordance with ethical approval from the University of Oxford's Medical Sciences Interdivisional Research Ethics Committee.

2.2.1.2. Materials and Procedure.

Eye-tracking methodology. An Eyelink 1000 eye-tracker (SR Research) recorded right eye movements at 1000 Hz during the reading task. Participants sat 55cm from the screen. Pupil and corneal reflection thresholds were adjusted according to the eye-tracker instructions. The eye-tracker was then calibrated and validated to <.5° using three fixation crosses. Calibration was repeated during the experiment when this threshold was lost. The task was programmed using EyeTrack (http://www.blogs.umass.edu/eyelab/).

Reading task. On each trial, there was an initial drift correction to control for slight movements – participants fixated on a black square located where the first character of the sentence would be. Participants then read a sentence, before pressing a button on a controller to advance to the next screen. Sentences were displayed in size 18 Courier New font on a single line in the (vertical) centre of the screen, with approximately 3 characters subtending one degree of visual angle. The sentence disappeared after 10s; in less than 1% of trials, participants did not finish reading the sentence before it timed out.

There were 32 experimental items (see Appendix A). Most items were derived from those used in Christianson et al. (2001), with some minor amendments to ensure they were not anomalous to a British audience. As Christianson et al. only used 12 items

with reflexive absolute transitive verbs (elaborated on below) and 16 were required here, 4 new items were created and added. All took a form such as: *While the father changed(,) the baby played with its toys.* Participants saw 16 of these as a *garden-path* sentence (with no comma), and 16 as a *comma* sentence (with the disambiguating comma). This within-participant variable is referred to as Structure. Whether participants saw a given item with or without a comma was counterbalanced across participants. Experimental items were interspersed with 40 filler sentences that resembled the experimental sentences (for instance, by starting with the word *While*) but did not contain a syntactic ambiguity.

The stimuli contained both optionally-transitive and reflexive absolute transitive verbs (cf. Christianson et al., 2001), and the sentences varied in length; these manipulations are not discussed further for two reasons. First, their main purpose was to increase the diversity of stimuli and reduce repetition, rather than for theoretical purposes; and second, neither Experiments 2 nor 3 included these manipulations, and it is therefore simpler not to present them for this experiment either. Including these variables into the statistical models did not alter any of the effects reported in Experiment 1, and only produced one additional significant effect, which is noted in the Results section. However, the distinction between optionally-transitive and reflexive absolute transitive verbs is discussed further in Chapter 4, where the results from Experiment 1 are compared to those from a group of older readers on this task. The full analyses (including these variables) are available in Appendix B.

Comprehension was assessed with a yes-no question that followed each sentence on a new screen. The 32 experimental items were followed by questions such as *Did the father change the baby?* for which the correct answer was always no. 16 of the 40 filler sentences were also followed by a question for which the correct answer was always yes.

Participants indicated their response to questions using the controller. The next sentence then appeared.

Working memory load task. Participants in the 2-back condition completed a concurrent task designed to tap verbal working memory resources. They were instructed that while reading the sentences, they would occasionally see words entirely in capital letters which they should try to remember. To-be-remembered probe words were embedded in filler sentences only, as in (2.4):

(2.4) While the pilot FLEW gracefully, the kites flapped in the wind.

Twenty-four probe words were distributed throughout the experiment; probe words were unrelated to experimental items in the main task. Participants were instructed to remember the last two words seen in capitals at all times. On eight unpredictable occasions during the experiment, the screen displayed a two-alternative forced-choice question ("What was the second-to-last word you saw in capitals?") followed by a choice of two words. One word was the correct answer, and the other a foil that had either recently featured in one of the sentences or that was a similar word to the correct answer. Participants made their decision using the controller.

The order of presentation of stimuli in the reading task was not random to facilitate programming the n-back task.[15] Two orders of presentation were created: half of participants in each condition saw items in one order, and the other half in the other order.

[15] If sentences had been presented randomly, there would be no straightforward way of programming the 2-back probe questions as the program would not know the identity of the last word displayed in capitals.

2.2.2. Results

Eyelink software (http://www.blogs.umass.edu/eyelab/) was used to analyse eye-tracking records. Two participants were removed due to excessive tracker loss (both in the 2-back condition), leaving 54 participants in the final analysis. Trials were removed when participants blinked or otherwise did not fixate on the disambiguating verb (5.2 % of trials). Fixations above 800ms or below 80ms were also deleted.

Data were analysed using linear mixed-effects models (LMEs) in the *lme4* package (Bates, Maechler, Bolker, & Walker, 2015) in R. LMEs have several advantages over analyses of variance for psycholinguistic data with crossed random effects (Baayen, Davidson, & Bates, 2008; Jaeger, 2008). The models simultaneously account for the fact that participants are a random sample of the population, and that experimental items are a random sample of the population of possible sentences. As well as modelling differences between participants and between items (for example, some participants read faster than others), they also account for how each independent variable affects each participant and item differently (for example, that load affects some participants' reading times more than others, or removing the comma affects the reading of some items more than others).

All models had Structure (garden-path vs. comma) and Load (no-load vs. 2-back) centred and entered as fixed-effects. To protect against Type I errors, a fully maximal random effect structure was used, of random intercepts and within-subject random slopes of Structure on participants and on items, and of Load on items (cf. Barr, Levy, Scheepers, & Tily, 2013). Where a model failed to converge, random slope interactions and effects were removed one-by-one until convergence.

Models analysing continuous dependent variables (fixation durations) were fitted using restricted estimations of maximum likelihood (REML), with reading durations log-

transformed to improve model fit due to positive skew (Baayen & Milin, 2010)[16]. There was no initial data trimming; instead and as recommended by Baayen and Milin (2010), data points were removed in REML models where their absolute standardised residuals in the first model fit were greater than 2.5 standard deviations away from the mean, with models then re-fit. It is the re-fit models that are reported here. Models with binary dependent variables were fitted as binomial logistic generalised models, following *z* distributions. There is no agreed-upon way to calculate degrees of freedom, and hence *p* values, from LME *t* distributions; here, as is common in similar studies, statistical significance at the .05 level is assessed by $t/z > 2$ (Baayen et al., 2008).

2.2.2.1. Did 2-back participants focus on the n-back task? Recall performance was high ($M = 84.6\%$, $SE = 2.5\%$) and significantly above chance according to a one sample *t* test, $t > 2$. This demonstrates that those in the 2-back condition concentrated and performed well on the concurrent task.

2.2.2.2. Does working memory load affect comprehension question accuracy? Figure 2.1 shows mean accuracy across Structure and Load. Both no-load and 2-back participants showed an effect of Structure, with little difference between the two Load groups. An LME model confirmed that accuracy was significantly poorer for questions following garden-path sentences [$M = 80.3\%$, $SE = 1.4\%$] than comma ones [$M = 85.3\%$, $SE = 1.2\%$], $z = 2.26$, but there was no difference between the two Load groups, and no interaction between Load and Structure, both $z < 1$.

While the manipulation is not discussed in this experiment, a significant difference was found between accuracy for items containing optionally-transitive verbs

[16] Note that means and standard errors referred to in this thesis (including those used for figures) are based on arithmetic means of raw reading durations rather than on log-transformed data. This is consistent with most eye-tracking research and is arguably more intuitive to understand and interpret.

[M_{CO} = 76.4%, SE_{CO} = 2.1%; M_{GP} = 71.3%, SE_{GP} = 2.2%] and items with reflexive absolute transitive verbs [M_{CO} = 93.6%, SE_{CO} = 1.4%; M_{GP} = 89.2%, SE_{GP} = 1.5%], $z > 2$. There was no interaction between Structure and verb type. This is discussed further in Experiment 2, and again in Chapter 4.

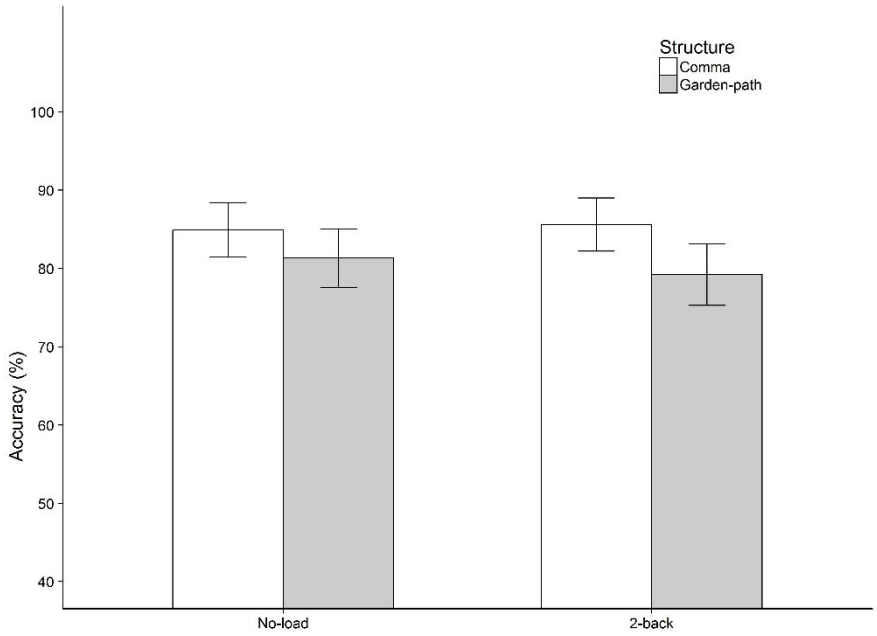

Figure 2.1. Mean accuracy for comprehension questions in Experiment 1, by Load and Structure (error bars indicate 95% confidence intervals). White bars indicate questions following comma sentences.

2.2.2.3. Does working memory load affect eye-movement patterns? While there are many potential dependent variables that can be collected from eye-movement data, it is important not to choose every measure and consequentially vastly increase the chance of a Type I error (cf. von der Malsburg & Angele, 2017). Accordingly, while data are presented from a range of measures (see Table 2.1), three dependent measures were focused on during analysis, in line with similar research (e.g., Slattery et al., 2013). These were: *first pass duration* (the duration of all fixations during the first visit to a region, if one was made), *go-past duration* (the duration from entering a region from the left before exiting a region to the right, including time during any regressions to the left), and *total*

reading duration (the total amount of time spent in a region, on all visits). I will also briefly discuss *second-pass duration*: the duration of any revisits into a region after going past it, if a revisit was made (if no revisit was made, the data point is excluded). First pass duration is considered an *early* measure, thought to reflect lexical processing. Second pass reading duration is deemed *late*, reflecting higher-level reprocessing of the text; go-past duration and total reading duration include first-pass duration as well as re-reading time, and so can be considered a combination of both (Clifton et al., 2007). For all measures, fixation durations are for the critical region containing the disambiguating verb (*played* in *While the father changed(,) the baby played with its toys*.)

Table 2.1 gives descriptive data from a range of dependent measures. Linear mixed effects models were run on the measures not discussed in detail below. They found a main effect of Structure on first fixation durations, $t = 2.40$, regressions in, $z = 4.29$ and regressions out, $z = 3.88$, all in the expected direction with longer reading durations and more regressions in the garden-path condition. There was also a marginal effect of Group on regressions in, $z = 1.94$, and a significant effect on regressions out, $z = 2.18$, with more regressions of both types in the 2-back condition. There were no effects on skipping rate, and no other significant effects or interactions.

First pass duration is shown in Figure 2.2. The no-load group showed a clear garden-path effect, but the 2-back group did not. An LME model for first pass duration supported an overall effect of Structure, $t = 2.10$, but no difference with Load, $t < 1$; the interaction between Structure and Load was not significant, $t = 1.57$. The two groups' results were considered separately *post hoc*: while the no-load group showed a clear garden-path effect with shorter first pass durations on comma sentences [$M = 236$ms, $SE = 5$ms] than garden-path sentences [$M = 257$ms, $SE = 5$ms], $t = 2.38$, the load eliminated

any such difference in the 2-back participants, $t < 1$. The difference between the groups should however be treated with caution given that the interaction was not significant.

Table 2.1.

Additional descriptive statistics for Experiment 1, by Structure and Load (standard errors in brackets). All measures are as defined in Table 1.1; other measures are described in-text.

	No-load		2-back	
	Comma	Garden-path	Comma	Garden-path
First fixation duration (ms)	209 *(3)*	224 *(4)*	222 *(5)*	222 *(5)*
First pass duration (ms)	236 *(5)*	257 *(5)*	244 *(7)*	249 *(7)*
Go-past duration (ms)	285 *(11)*	415 *(30)*	323 *(14)*	467 *(32)*
Second pass duration (ms)	335 *(27)*	333 *(23)*	289 *(21)*	412 *(28)*
Total reading duration (ms)	370 *(13)*	429 *(15)*	373 *(12)*	512 *(18)*
% first pass regressions out	11.2 *(1.7)*	17.8 *(2.1)*	17.5 *(2.1)*	28.7 *(2.4)*
% trials with regression in	37.6 *(2.6)*	47.7 *(2.7)*	45.0 *(2.7)*	58.5 *(2.7)*
% skipping	17.9 *(1.9)*	13.6 *(1.7)*	17.9 *(1.9)*	18.6 *(1.9)*

For go-past durations, a garden-path effect is visible in Figure 2.2 for both groups. An LME model supported this, finding a main effect of Structure, $t = 3.57$, but no main effect of Load, nor an interaction, both $t < 1$. The effect of Structure was as expected, with longer go-past durations for garden-path sentences [M = 441ms, SE = 39ms] than comma sentences [M = 304ms, SE = 16ms]. Table 2.1 gives the proportion of trials on which a regression was made before exiting the disambiguating region (i.e., where go-past duration was greater than first-pass duration), excluding any trials where the disambiguating region was skipped.

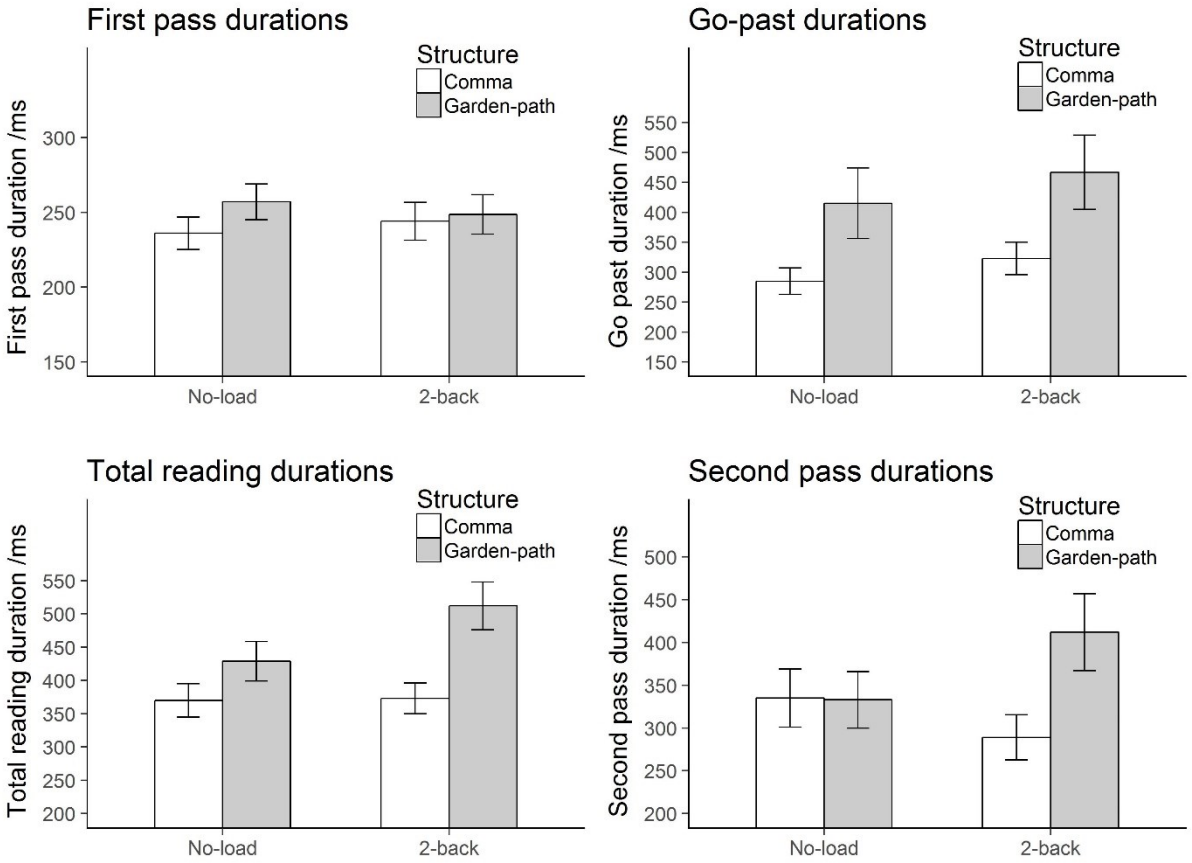

Figure 2.2. First pass, go-past, total reading and second pass durations in Experiment 1, by Load and Structure. Error bars indicate 95% confidence intervals; white bars are for comma sentences.

The final measure is total reading duration on the critical verb, shown in Figure 2.2. A garden-path effect is present in both groups, but is more pronounced in the 2-back group. An LME model confirmed a main effect of Structure, $t = 5.82$, qualified by a significant interaction with Load, $t = 2.58$; there was no main effect of Load, $t = 1.14$. The size of the garden-path effect, as measured by the difference between mean total duration for garden-path sentences and for comma sentences, is significant in both groups, but is notably larger in the 2-back group [$M = 139$ms, $SE = 33$ms], $t = 5.53$, than the no-load group [$M = 59$ms, $SE = 25$ms], $t = 3.29$.

To explore this finding further, second-pass durations were also examined. Note that this measure is only based on trials where a second pass was made (47% of all trials where a first pass was made; see Table 2.1), and so caution is needed in its interpretation. However, there was a main effect of Structure, $t = 2.25$, qualified by a significant interaction with Load, $t = 2.99$; there was no main effect of Load, $t < 1$. As shown on Figure 2.2, this interaction demonstrated that while the 2-back group showed significantly longer second-pass durations on garden-path sentences [$M = 412$ms, $SE = 28$ms] than comma sentences [$M = 289$ms, $SE = 21$ms], $t = 3.57$, there was no effect of Structure in no-load participants. For the no-load group, time spent re-reading (if this was done) was therefore broadly equivalent for comma and garden-path sentences. This supports the result from total reading durations that the locus of the difference between the load groups was in re-reading.

2.2.3. Discussion

Experiment 1 examined how an extrinsic memory load affects comprehension and eye-movements as people read sentences containing a temporary syntactic ambiguity. This combination of eye-tracking and comprehension questions was designed to elucidate what is meant by *good enough* in relation to sentence processing (Christianson et al., 2001; Ferreira & Patson, 2007). Experiment 1 replicated previous findings of increased looking times and decreased comprehension accuracy for garden-path sentences; this provides a valid baseline against which to compare performance under load. With the addition of load, people had to re-read for longer to detect the ambiguity: a smaller effect in first pass reading times was replaced by a stronger effect in re-reading, as indexed by go-past durations and total reading durations. This need for re-reading to complete

processing resembles the eye movement patterns seen in older readers (DeDe, 2014, 2015; Kliegl et al., 2004; Rayner et al., 2006, 2009). In contrast to older adults, load did not affect comprehension accuracy, contrary to the hypothesised effect.

The reduction in accuracy for questions following garden-path sentences was less pronounced than in previous studies (e.g., the 60% error rate seen in Christianson et al. (2001) on garden-path questions). One explanation is the change in methodology: participants were not given any instructions about not re-reading the sentence (cf. Christianson et al., 2001, 2006; Patson et al., 2009). Not permitting re-reading may increase working memory load as it requires more recall of earlier text, in turn leading to more difficulty in processing, and eventually, in comprehension (measured by question accuracy or by other methods, such as paraphrasing; Patson et al., 2009). High accuracy was also seen in Engelhardt's (2014) eye-tracking study, which used sentences and questions similar to those used here (like this experiment, Engelhardt also allowed unrestricted reading). It is also noteworthy that the participants in this experiment were highly skilled readers; comprehension was barely hindered, despite the complexity of the sentences. High levels of reading skill may also have affected the strategies employed when given an extrinsic load, a point returned to later.

No-load participants showed a significant garden-path effect on first pass durations, as observed in previous research. The garden-path effect was less clear in 2-back participants: a *post hoc* analysis did not find a significant effect in this group, although the absence of an overall Structure-by-Load interaction should be reiterated. The 2-back group did spend more time reprocessing, as indexed by later eye movement measures. This suggests a distinction: no-load participants generally resolved the ambiguity at first pass, showing less need for reanalysis; in contrast, the participants under load required a re-read. This pattern resembles both eye movements observed in the

elderly, and also younger adults with low working memory spans (Kemper et al., 2004). Together, these results indicate a link between the availability of working memory resources and eye movement behaviour during syntactic processing.

Importantly though, as load did not affect question accuracy, there is no evidence that the 2-back group showed any eventual impairment in comprehension. Why are there no downstream effects on comprehension (unlike older readers in Christianson et al., 2006)? One explanation is that the secondary task did not cause sufficient disruption, stemming from the combination of three possible factors: first, a group of highly skilled participants; second, sentences that while temporarily ambiguous were not too taxing to reprocess; and third, a concurrent task that was not resource-intensive. As a result, 2-back participants showed an initial slowing of reading all sentences: a strategy that is good enough most of the time, and that reflects the resource demands of the concurrent task. However, when an ambiguity is eventually observed, 2-back participants engaged in significantly more reprocessing to fully resolve it. This reprocessing was necessary: as discussed earlier, ambiguities such as those in the sentences used here cannot be left underspecified when comprehension is being tested (unlike, for example, attachment ambiguities; cf. von der Malsburg & Vasishth, 2013). The high performing readers in Experiment 1 are likely to have recognised this, but without any restrictions imposed on re-reading, they were able to use reprocessing time to maintain comparable comprehension to participants reading without load.

An outstanding question is whether the observed effects are specific to reading single sentences, whose relative simplicity may have minimised the effect of the ambiguity and of the (relatively simple) concurrent task. Embedding the sentence into a short passage may alter reading strategies: the first sentence may be read less carefully on the expectation that the second sentence will provide additional context, alongside any

possible additional processing demands of reading and recalling a longer block of text. The absence of a first pass garden-path effect in the 2-back group here was interesting (although note again the absence of a significant interaction), given that this effect is prevalent in sentence processing studies (cf. Clifton et al., 2007). However, first pass effects have been elusive in previous studies where complex sentences appear in passages rather than in isolation (Ferreira & Clifton, 1986; Slattery et al., 2013). First pass durations were also shorter in Radach et al.'s (2008) study when reading passages, with increased times shifting to total reading durations. The classic first pass garden-path effect therefore appears to be task dependent. A difference in eye movement behaviour when reading passages would also extend Slattery et al.'s (2013) findings of lingering effects in a subsequent sentence, effects demonstrating that the initial misinterpretation was not pruned before the second sentence was read. Slattery et al.'s (2013) eye movement data do not necessarily tell us about comprehension of those passages: reading a passage rather than an isolated sentence could either improve comprehension (since there is more material from which to gain comprehension), or impair it (if it produces more superficial reading of the text). Experiment 2 therefore set out to explore these questions.

2.3. Experiment 2

Experiment 2 looked at whether the effects seen when reading sentences in Experiment 1 would also be observed when participants read short passages containing a garden-path sentence, such as in (2.5):

(2.5) While the father changed(,) the baby that was cuddly played with its toys. The father had to finish changing his clothes to pick up and look after the baby.

Here, the first sentence is as seen in Experiment 1. The second sentence should entirely clarify the meaning and remove the initial misinterpretation: the explicit use of a reflexive makes clear that the father is getting changed, rather than changing the child. Other than the use of 2-sentence passages, Experiment 2 resembled Experiment 1 in design, being a 2x2 study of Structure (garden path vs. comma) and Load (no-load vs. 2-back).

The main rationale for considering passage stimuli is that most work in this area has presented ambiguous sentences in isolation (Radach et al., 2008; see Clifton et al., 2007 for a review of relevant research). Research using longer passages has tended to do so in order to consider the effects of referential context in an opening sentence (e.g., Altmann et al., 1992; Ferreira & Clifton, 1986). But the mere presence of other sentences *by itself* may affect reading behaviour, if it shifts reading strategies: on the one hand, more information should help to disambiguate an ambiguous sentence; on the other, a passage of text means more to read, and the first sentence may be read more superficially on the expectation of receiving further information in later text. This latter prediction should be even more apparent under cognitive load. Either way, there should be a difference from Experiment 1 in eye movement patterns and potentially in subsequent comprehension (irrespective of load). This is supported by the differences already discussed in first pass reading durations when reading passages compared to sentences

(Kuperman et al., 2013; Radach et al., 2008). This superficiality account is also supported by the lingering effects observed by Slattery et al. (2013) in eye movements on the second sentence (*changing his clothes…*). If the idea that the father is changing his clothes comes as a surprise, as Slattery et al.'s findings indicate, this indicates either a memory failure, or that the process of clearing up the misanalysis in the first sentence must have been incomplete.

There were three main research questions in Experiment 2. First, how would load affect reading in this experiment? Participants reading under a 2-back load were expected to again show more superficial initial reading with increased reprocessing of garden-path sentences, as seen in Experiment 1. If Slattery et al.'s (2013) lingering effects in the second sentence were replicated, this should be even more apparent in the 2-back group if the first sentence was read more superficially. Performance on the comprehension questions provided an opportunity to see whether the lack of load effects seen in Experiment 1 replicated.

Second, would eye movements when reading the first sentence be similar to the results of Experiment 1? This was not a within-item design, as the stimuli were marginally different from Experiment 1; nevertheless, it was still of interest to compare eye movements and comprehension in the two experiments. If readers expect that continuing on to read the second sentence will aid comprehension, this could obviate the need to spend longer reading the critical region - reducing or even eliminating the garden-path effect in first pass durations on the first sentence. A garden path effect should still occur in later measures, especially in total reading durations (cf. the longer total reading durations for passages in Radach et al. (2008)).

The final question was whether comprehension accuracy would be affected by the move from sentences to passages. Intuitively, one may predict that comprehension would improve as a result of the disambiguating power of the second sentence. In contrast, however, I predicted that the availability of a further line of text would cause people to read less carefully than in Experiment 1, with initial misinterpretations more likely to be maintained without full reinterpretation. This would lead to no improvement in comprehension performance relative to Experiment 1, or even a reduction in performance.

2.3.1. Method

2.3.1.1. Participants. Sixty students at either the University of Oxford (*n* = 53) or Oxford Brookes University (*n* = 7) participated (41 female; mean age 22.3 years, range 18 – 30 years), with all receiving £5 or course credits as compensation. Participants were again recruited using posters, an online recruitment system and via an undergraduate research participation scheme; once more, there was a range of participants from psychology (*n* = 12), other science subjects (*n* = 21), humanities (*n* = 13) and social sciences (*n* = 14). None of the participants took part in Experiment 1. Participants were allocated into one of two conditions, performing the reading task either with a concurrent working memory task (*2-back*) or without (*no-load*).

2.3.1.2. Materials and procedure. The same eye-tracker was used as in Experiment 1. Due to the passage design, calibration and validation instead used nine fixation crosses, and the accuracy threshold was always < .75° for all nine crosses, and always < .5° for the top left cross (where passages began). Calibration was repeated when

this threshold was lost. The task was programmed using Experiment Builder (SR Research).

The basic reading task resembled Experiment 1, with the following differences. Sentences appeared in passages, with the first sentence being similar to those used in Experiment 1. Temporarily ambiguous sentences were similar in structure to Experiment 1, with some minor stylistic differences; see Appendix A. All experimental items used reflexive absolute transitive verbs. The second sentence resolved the ambiguity using a reflexive term, similarly to the items in Slattery et al. (2013)'s second experiment. Sentences were displayed in size 14 Arial font on two lines towards the top of the screen, with approximately 3.2 characters subtending one degree of visual angle. Items were triple-spaced to prevent minor calibration issues from causing confusion with which line was being read. The line break fell after the second sentence had begun but with at least one word on the second line before the second critical region began (see (2.5) above).

The maximum reading time permitted for each passage was 40 seconds. Since it took longer to read passages than single sentences, the number of items was reduced: participants saw 8 garden-path passages, 8 comma passages and 32 filler passages. Again, it was counterbalanced whether participants saw a given item with, or without, a comma. All experimental and filler items were followed by a comprehension question on the next screen. Participants were given 10 seconds to respond Yes or No. Questions after experimental items tapped the syntactic ambiguity, and the correct answer was always No (as in Experiment 1).

The n-back task also resembled Experiment 1, with two exceptions. First, there were only twenty probe words and six questions, given the reduced number of items. Second, to facilitate programming of the experiment, the question was changed to produce a yes/no response. Questions resembled, for example, "Was the second to last

word in capitals TRAIN rather than CONDUCTOR?"). As participants in the 2-back condition could not easily be shown the items randomly, instead two orders of presentation were created. No-load participants saw items in a randomised order.

2.3.2. Results

Eye-tracking records were analysed using Data Viewer (SR Research). Two participants (one per load condition) were removed due to excessive tracker loss, leaving 58 participants in the final analysis. Data were analysed using linear mixed-effects models similarly to Experiment 1. Fixations above 1200ms were deleted; fixations below 40ms were merged with adjacent fixations or if not possible, deleted. It should be noted that these are marginally different exclusion criteria from those used in Experiment 1 (where fixations below 80ms and above 800ms removed), as a different setting was used when changing software. However, this only resulted in 2% more fixations being included. If these were notable outliers, the model criticism process should have prevented them from skewing the presented results. Furthermore, critical results (both in Experiment 2, and in Experiments 3 to 5, which also used the 40ms and 1200ms criteria) were re-run excluding any trials on which the first pass duration had been less than 80ms or more than 800ms: there was no change in the pattern of results.

2.3.2.1. Did 2-back participants focus on the n-back task? Participants in the 2-back group performed well and significantly above chance on the concurrent working memory task (M = 79.7%, SE = 3.0%), $t > 2$. This level of performance is similar to that seen on the n-back task in Experiment 1, suggesting that both the n-back task used here and the participants are comparable to Experiment 1.

2.3.2.2. Does working memory load affect comprehension question accuracy? Figure 2.3 (Panel A) shows the comprehension accuracy results for Experiment 2.

Accuracy was again significantly poorer for garden-path sentences (M = 75.0%, SE = 2.0%) than comma sentences (M = 88.3%, SE = 1.5%), z = 4.16. The garden-path effect looks more pronounced in the 2-back group, but an LME model found no interaction between Load and Structure, $z < 1$, nor any main effect of Load, $z < 1$.

As one of the purposes of Experiment 2 was to explore the difference between sentences and passages, the comprehension accuracy results for Experiments 1 and 2 were combined in one model with three fixed effects: Structure, Load and Experiment (Expt 1 vs. Expt 2), and their interactions. While the design was neither within-subject nor within-item (the items were similar but not identical in design), the comparison remained of interest for understanding differences in pattern. Accuracy in Experiment 2 was poorer than in Experiment 1; descriptive data are in Table 2.2. Beyond an overall effect of Structure, z = 4.04, the model found a main effect of Experiment, z = 3.49, and an Experiment-by-Structure interaction, z = 2.37, due to poorer accuracy and a stronger garden-path effect in Experiment 2. There were no other significant effects.

Experiment 1 used both optionally-transitive (OPT) and reflexive absolute transitive (RAT) verbs, with a clear difference in comprehension accuracy between the two; in contrast, Experiment 2 only used the latter. For a closer comparison, this model was run again using only the RAT items from Experiment 1. This reproduced the main effects of Structure and Experiment, but the interaction failed to reach significance, z = 1.50. It is unclear whether the absence of an interaction indicates the interaction was driven only by the inclusion of OPT verbs, or whether the absence here is due to reduced power (from the reduced number of observations). As participants in Experiment 1 performed considerably *worse* on OPT items than RAT items (see Table 2.2), it is unlikely that this interaction effect would be driven by these data, favouring the explanation of lack of power.

Table 2.2.

Percentage accuracy on comprehension questions (standard errors in brackets), by Experiment, Verb type [for Experiment 1 only] and Structure (comma vs. garden-path).

	Comma items	Garden-path items
Experiment 1 (OPT only)	76.5 *(2.1)*	71.3 *(2.2)*
Experiment 1 (RAT only)	93.6 *(1.2)*	89.2 *(1.5)*
Experiment 1 (all items)	85.2 *(1.2)*	80.3 *(1.4)*
Experiment 2 (all are RAT)	88.3 *(1.5)*	75.0 *(2.0)*

2.3.2.3. Does load affect eye-movement patterns? Three dependent measures were again analysed: first pass duration, go-past duration and total reading duration for two different regions. (2.5) is repeated below, with these regions underlined.

(2.5) While the father changed the baby that was cuddly played *with its* toys. The father had to finish changing his clothes to pick up and look after the baby.

Second pass duration was not calculated: with a passage design, an increase in the number of eye movements (and especially regressive eye movements, cf. Heustegge & Bocianski, 2010) makes the measure more difficult to interpret meaningfully. However, the proportion of trials with a *first pass regression out* was also looked at.

The first region was comparable to Experiment 1, and the results presented below are for this single-word region 1 for direct comparison with Experiment 1. However, due to the increased length of the stimuli, and following similar experiments (e.g., Slattery et al., 2013), eye movements were also analysed for an extended region 1, with the region extended to include a spillover region following the disambiguating verb. The final word of the sentence was not included, to avoid capturing wrap-up effects due to the period. Therefore, when there was only one word between the disambiguating verb and the final

word of the sentence, this intervening word was the spillover region; where there was more than one word, the spillover region comprised the two words that followed the disambiguating verb. Spillover regions were always at least four characters long. There was little difference between the two variants of Region 1, but any differences in the pattern of results are presented below.

Region 2 tested for lingering effects of the misinterpretation. Stimuli were designed such that there was always at least one word on the second line before this second region (in (2.5), the new line started with the word *finish*). This was to avoid a confound with the effects of the new line on eye movements, and to allow participants parafoveal preview of the second critical region (cf. Clifton et al., 2007).

Analysis of region 1. Firstly, as with Experiment 1, descriptive data for the one-word region 1 are presented in Table 2.3, alongside data from Experiment 3. Linear mixed effects models found no effects on first fixation duration or regressions out. There was a main effect of Structure on regressions in, $z = 3.40$, and a main effect of Load on skipping rate, $z = 2.22$. This latter effect will be discussed more in Experiment 3.

Figure 2.3 shows the eye-movement results for region 1. Unlike in Experiment 1, there was no sign of a garden-path effect on first pass durations in either load condition (see Figure 2.3, Panel B). Accordingly, the LME model found no significant effect of Structure on first pass duration, $t = 1.14$; neither the effect of Load nor the interaction were significant.

For go-past durations, a garden-path effect is visible in Figure 2.3 (Panel C) in both groups, as in Experiment 1. An LME model supported this, finding a main effect of Structure, $t = 2.60$, qualified by an interaction with Load, $t = 2.39$; there was no main effect of Load, $t < 1$. The interaction was not seen in the longer Region 1 (see below), and

was not evident from inspection of the means alone. However, closer inspection found several significant outliers in the no-load condition. By removing go-past durations longer than 3000ms (1.5% of the data), it became apparent that the nature of the interaction was a stronger garden-path effect in the 2-back condition than the no-load condition.

As per Figure 2.3 (Panel D), there was a main effect of Structure on total reading durations, $t = 5.74$, but no other effects. Analysis was also conducted on the proportion of trials on which a regression was made to an earlier part of the text upon reaching region 1 (that is, how often go-past durations differed from first pass durations). Descriptive data were given in Table 2.3. While a garden-path effect was present, this was not significant, $z = 1.63$ (and neither were the effect of Load, $z = 1.37$, nor interaction, $z = 1.57$).

Table 2.3.

Additional descriptive statistics for Region 1 eye movements in Experiments 2 and 3, by Structure and Load (standard errors in brackets). All measures are defined in Table 1.1.

	No-load		2-back		4-back	
	Comma	G-path	Comma	G-path	Comma	G-path
First fixation duration	229 *(8)*	236 *(7)*	215 *(7)*	226 *(8)*	246 *(9)*	246 *(9)*
First pass duration	287 *(15)*	276 *(11)*	239 *(11)*	255 *(12)*	280 *(11)*	281 *(11)*
Go-past duration	485 *(31)*	708 *(70)*	467 *(43)*	671 *(56)*	528 *(39)*	661 *(76)*
Total reading duration	548 *(26)*	751 *(35)*	530 *(25)*	680 *(32)*	518 *(24)*	753 *(35)*
% regressions out	32.4 *(3.6)*	32.6 *(3.6)*	32.8 *(4.0)*	45.6 *(4.3)*	38.7 *(3.8)*	36.6 *(3.8)*
% regressions in	55.3 *(3.8)*	66.9 *(3.6)*	61.9 *(4.2)*	75.0 *(3.7)*	50.3 *(3.9)*	70.9 *(3.6)*
% skipping (short R1)	29.2 *(2.9)*	28.3 *(2.9)*	40.2 *(3.3)*	39.3 *(3.3)*	37.2 *(3.1)*	30.6 *(2.9)*
% skipping (longer R1)	14.2 *(2.3)*	18.3 *(2.5)*	20.1 *(2.7)*	20.5 *(2.7)*	15.8 *(2.3)*	14.9 *(2.3)*

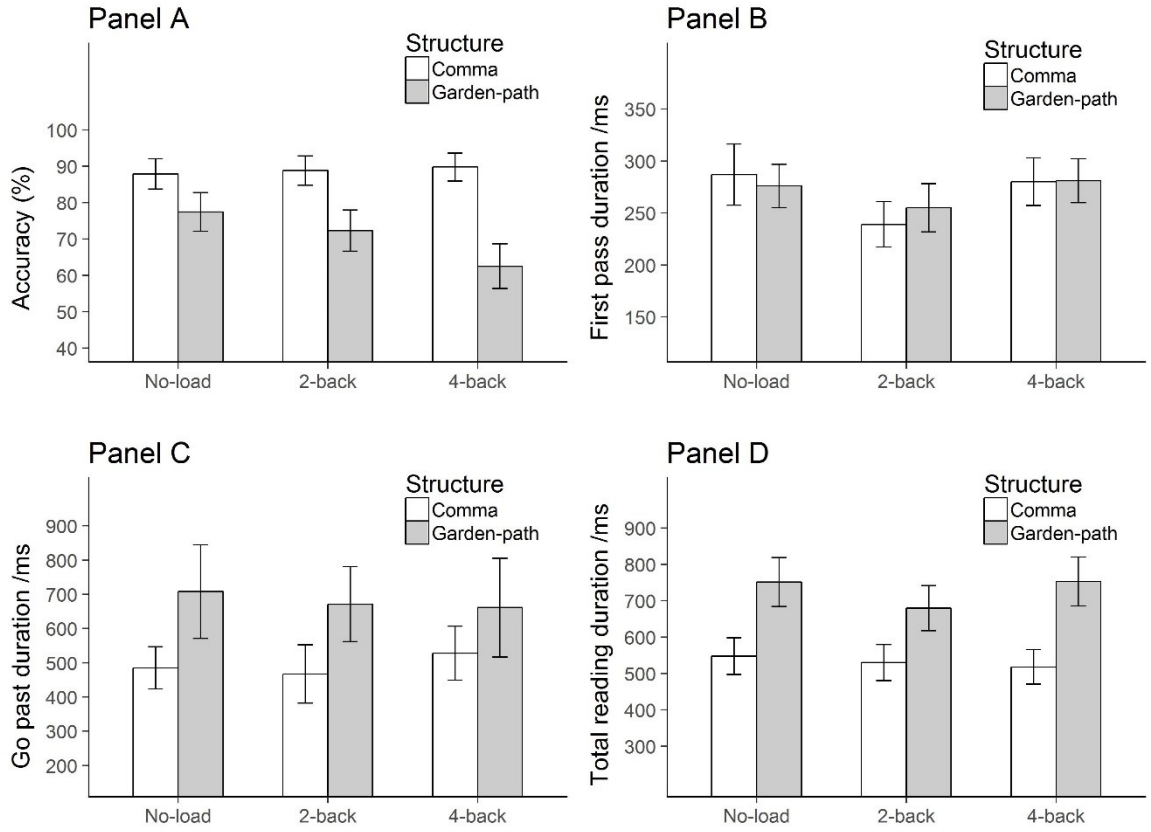

Figure 2.3. Results for Experiment 2 (*no-load* and *2-back*) and Experiment 3 (*4-back*), by Load and Structure (error bars indicate 95% confidence intervals): comprehension accuracy (Panel A); and Region 1 first pass durations (Panel B), go-past durations (Panel C) and total reading durations (Panel D). White bars indicate comma items.

Analysis of extended region 1. The pattern of results was broadly similar to the shorter region 1, with two exceptions. There was no Structure-by-Load interaction in go-past durations, $t < 1$ (even if go-past durations over 3000ms were excluded). Additionally, the main effect of Structure on regressions out was significant, with regressions made more often on garden-path items (60% of trials) than comma items (48% of trials), $z = 3.59$. The fact that this did not reach significance for the shorter Region 1 indicates that regressions were commonly made from the spillover region on garden-path items.

Analysis of region 2. First pass, go-past and total reading durations were calculated for Region 2 in the second sentence. The only visible effect was a small but non-significant garden-path effect in go-past durations, displayed in Figure 2.4 [M_{CO} = 702ms, SE_{CO} = 41ms; M_{GP} = 847ms, SE_{GP} = 69ms]. The size of the error bars on Figure 2.4 (which were similar for the other measures) demonstrated clear variability in these data. In line with this, LME models for the three measures produced no significant main effects or interactions.

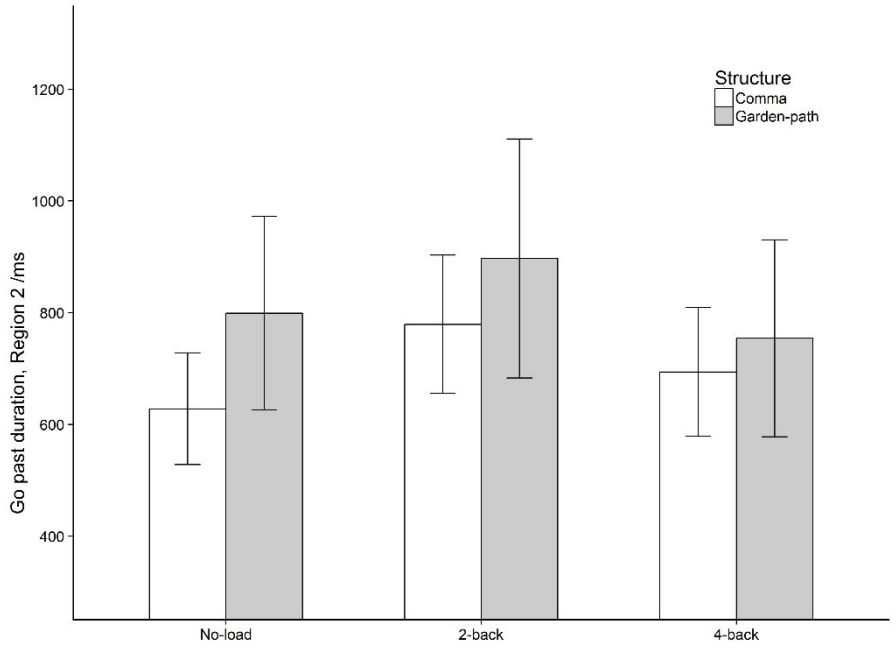

Figure 2.4. Region 2 go-past durations for Experiments 2 (*no-load* and *2-back*) and 3 (*4-back*), by Load and Structure (error bars indicate 95% confidence intervals).

2.3.3. Discussion

In Experiment 2, participants' eye movements and comprehension were measured for passages containing garden-path sentences, with a manipulation of extrinsic memory load. The results can be summarised as follows. The temporary syntactic ambiguity again resulted in reduced comprehension accuracy, relative to performance on the same

questions following comma items. Comprehension was not affected by load. Interestingly, accuracy was also poorer than in Experiment 1, despite the fact that the second sentence of the passage should make the meaning clear. For eye movements, Structure did not have a significant effect on first pass reading durations in the disambiguating region. This is unlike in Experiment 1 and much previous literature, but replicates results from other studies using passages as stimuli rather than isolated sentences (e.g., Slattery et al., 2013). However, the standard garden-path effect was observed in go-past and total reading durations. Other than a small but unreliable effect in go-past durations, the lingering effects seen by Slattery et al. (2013) were not observed in eye movements on the second sentence. Finally, the addition of a memory load had little effect on any eye movement measures – unlike in Experiment 1.

It was interesting that comprehension accuracy was worse than in Experiment 1 (even for no-load participants). This was not a within-subject or within-item design, and so the data are not perfectly comparable; nevertheless, participants were more prone to the good-enough comprehension errors relative to single sentence conditions in Experiment 1, despite the second sentence making the first sentence's meaning explicit. Also of relevance is that Experiment 2 contained only reflexive absolute transitive verbs where there is no room for ambiguity: unlike for optionally transitive verbs such as "hunted", if someone is "washing" without an explicit direct object, they are definitively washing *themselves* and not the incorrectly attached noun phrase. These data therefore demonstrated that syntactic ambiguities linger, and can override the clarification provided by the second sentence. The second sentence had no facilitatory effect on comprehension, despite its disambiguating content; if anything, its presence impaired comprehension. This suggests that initial misinterpretations were more likely to be maintained when

reading the two-sentence passages of Experiment 2, compared to identical sentences in isolation in Experiment 1.

Increased superficiality of reading in passages may also explain the absence of a clear first pass garden-path effect on the disambiguating verb. As discussed earlier, a first pass effect is a classic and well-replicated finding in experiments using early-closure garden-path sentences presented in isolation (cf. Clifton et al., 2007). The findings here join those of others in showing no garden-path effect in first pass duration when reading a complex sentence within a multi-sentence passage (e.g., Ferreira & Clifton, 1986; Slattery et al., 2013). These experiments, like Experiment 2 here, did find a garden-path effect in go-past duration and total reading duration. Together, these findings demonstrate that reading passages rather than single sentences shifts people from preferring to spend longer reading on first pass (in order to finish parsing and move on) to either instigating a regression (seen in increased go-past durations), or reading on and hoping to gain comprehension, even if this is unsuccessful (seen in increased regressions from the spillover region). Of course, an absence of a significant effect does not necessarily constitute evidence that the first pass effect is not present (for instance, the effect may be small and underpowered); this is discussed further in the General Discussion of this chapter.

In contrast to Slattery et al.'s (2013) findings, there was little evidence to support a lingering garden-path effect in the eye movement record, on the second sentence. A numerical trend towards an effect in go-past durations did not come close to significance. This does not necessarily challenge Slattery et al.'s findings. The absence can be explained by differences in methodology (especially the requirement to answer questions tapping the ambiguity, encouraging participants to read the first sentence more carefully than the second; Logačev & Vasishth, 2016a; Swets et al., 2008), in stimuli (there was a

juxtaposition of plausible and implausible items in Slattery et al., which may have encouraged more commitment to *plausible* misinterpretations), and in precise interest areas. The standard errors around the mean (especially in go-past durations) demonstrate clear variation in eye movement patterns, as seen in Figure 2.4; this variation may have masked any difference seen by Slattery et al., whose reported effect was small (a 40ms garden-path vs comma difference for plausible items, and no effect in go-past or total reading durations). Taken together, it may be that different participants employ different strategies for different items, and lingering effects in the eye movement record may not always be captured by standard measures (cf. von der Malsburg & Vasishth, 2011, who advocate complete scanpaths to be analysed; see also Chapter 5). This also adds support to work that questions the garden-path model's principle that syntactic processing is deterministic, i.e., that the parser will always make the same decision when faced with a given syntactic structure. Other models (e.g., van Gompel et al., 2000) propose that this decision is probabilistic, which would explain the variable results here and elsewhere.

Surprisingly, and in contrast to Experiment 1, imposition of a concurrent task had no effect on eye movement behaviour. This suggests that either the concurrent task did not impact online reading behaviour when garden-path sentences were presented as part of a passage rather than as an isolated sentence, or alternatively, that no-load participants also re-read more while reading passages compared to sentences. The absence of an effect on first pass durations points towards the latter possibility: both 2-back and no-load groups tended to read the text more superficially in a passage than in a single sentence, obscuring any differences produced by the concurrent task as seen in Experiment 1.

A relevant factor is the difficulty of the concurrent task. Experiment 1 found an effect of the load on online processing, but not on comprehension. It is thus unclear from these data whether the concurrent task has differential effects on online and offline

processing, or whether it was simply insufficiently disruptive to impede eventual comprehension. Experiment 2 found no significant effects of load at all. Would increasing the burden of the concurrent task produce a more substantial impairment on online processing and subsequent comprehension? Experiment 3 investigated this.

2.4. Experiment 3

Participants in Experiment 3 faced a concurrent 4-back task (i.e., they were asked to simultaneously recall four words while reading). The idea was to assess whether it is simply the presence of an extrinsic load that affects syntactic processing, or whether the size of the load would affect the extent of superficial processing. Teasing these apart was intended to elucidate the effects of task demands on good-enough processing, and the extent to which online and offline measures tap separate processes.

Participants under load in Experiments 1 and 2 read more slowly (and with more regressions), arguably to compensate for the interference from the 2-back task. Clearly, participants can cope with some degree of extrinsic load, but there may be a tipping point at which it is no longer possible to compensate by slowing down (cf. Gao et al., 2011, 2012 for a similar finding of a tipping point, using a manipulation of visual noise). At this point, comprehension should decline. I therefore predicted that the 4-back task in Experiment 3 would be above this tipping point, and that comprehension following garden-path items would be poorer than in either condition of Experiment 2. An increase in impact as the demands of an n-back task increase has been observed before (Mcelree, 2001) and the idea of a tipping point could explain the absence of any load effect on comprehension in Experiments 1 and 2. I also predicted that the more difficult concurrent

task would have a more detrimental effect on eye movements than in Experiment 2 (with a greater increase in superficial reading at first followed by a need for more re-reading).

2.4.1. Method and Results

2.4.1.1. Procedure. Experiment 3 used the same design, materials and method as the 2-back condition in Experiment 2. The only change was that the concurrent task was increased to 4-back: participants instead had to answer which of two options was the *fourth*-to-last word that they had seen in capitals. There were 35 participants drawn from the same population as Experiment 2 (23 female, mean age = 22.8 years, range 19 – 30 years), but again, not having taken part in Experiments 1 or 2. Eye-tracking records were analysed as in Experiment 2. Four participants were removed due to excessive tracker loss, leaving 31 participants in the final analysis.

2.4.1.2. Did participants focus on the n-back task? Replicating Experiments 1 and 2, performance on this task was high – and significantly above chance (M = 79.4%, SE = 2.6%), $t > 2$. Accuracy was almost identical to that in Experiment 2 (M = 79.7%, SE= 3.0%). This demonstrates that participants concentrated on the concurrent task, and that increasing load did not adversely affect performance.

2.4.1.3. Does working memory load affect comprehension question accuracy? Accuracy, shown in Figure 2.3 (Panel A) alongside Experiment 2 for ease of comparison, was significantly poorer for garden-path sentences (M = 62.5%, SE = 3.1%) than comma sentences (M = 89.8%, SE = 1.9%), z = 5.38. Figure 2.3 also shows accuracy to questions following garden-path items falling as load increases. A further model was run to compare these results to those obtained in Experiment 2, analysing the effects of Structure (garden path vs. comma) and Load (no-load, 2-back, 4-back). Load was added as a

continuous (and centred) variable to reflect the graded nature of the task[17]. There was an overall effect of Structure, $z = 5.89$, but also an interaction with Load, $z = 2.90$. There was no main effect of Load, $z = 1.00$. The interaction reflects the pattern in Figure 2.3 that increased load worsens accuracy only on garden-path items.[18]

2.4.1.4. Does working memory load affect eye-movement patterns? As in Experiment 2, two regions of interest were considered: Region 1 with the disambiguating verb, and Region 2 with the reflexive term in the second sentence. Data for Region 1 are shown in Figure 2.3 (Panels B – D), alongside those from Experiment 2. Firstly, linear mixed effects models found no effects on first fixation duration, regressions out or skipping rate. There was a main effect of Structure on regressions in, $z = 3.92$. The results of Experiment 3 were also combined with those of Experiment 2: beyond the effect of Structure on regressions in, there were no further significant effects. Notably, the main effect of Load on skipping rate in Experiment 2 was not seen when considering results across both experiments, and so this is not considered further.

LME models with one fixed effect (Structure) were fitted for the Experiment 3 data. These found similar patterns to Experiment 2: no significant main effect of Structure on first pass durations [comma $M = 280$ms, $SE = 12$ms; garden-path $M = 281$ms, $SE = 11$ms], $t < 1$, and a significant effect on total reading durations [comma $M = 518$ms, $SE = 24$ms; garden-path $M = 753$ms, $SE = 34$ms], $t = 4.17$. However, in Experiment 3, there was not a *significant* garden-path effect for go-past durations [comma $M = 528$ms, $SE =$

[17] Models were also run entering Load as a discrete variable with three levels using treatment coding, to ensure this approach was reasonable. The pattern and reliability of effects did not change from those reported here.

[18] Confirming this pattern, separate models found that accuracy worsened with increasing Load in garden-path items, $z = 2.49$, but not in comma items, $z < 1$. Separate models also revealed a Sentence by Load interaction between 0-back and 4-back participants, $z = 2.84$; the effect between 2-back and 4-back participants did not quite reach significance, $z = 1.91$.

40ms; garden-path M = 661ms, SE = 74ms], $t < 1$. Again, there was also no effect of Structure on regressions out of the shorter Region 1, $z < 1$.

Once again, this was repeated for an extended region 1 including a spillover region. The pattern of results were identical, except that the main effect of Structure on go-past durations was significant, $t = 4.61$, with the usual pattern of longer durations on garden-path items [M = 1457ms, SE = 102ms] than on comma items [M = 954ms, SE = 68ms].

4-back participants in Experiment 3 were also compared to the no-load and 2-back participants in Experiment 2. There was no Structure by Load interaction in first pass, $t = 1.40$, go-past durations, $t = 0.42$, or total reading durations, $t = 1.44$. (This pattern of results was identical for the longer Region 1.)

Finally, eye movement data for Region 2 (e.g., *changing his clothes*) were examined. There was no effect of Structure on first pass duration or go-past duration, both $t < 1$; go-past durations are shown on Figure 2.4 for comparison with results from Experiment 2. There was a marginal garden-path effect in total reading durations from Experiment 3, $t = 2.01$, with longer reading durations for garden-path items [M = 683ms, SE = 37ms], than for comma items [M = 600ms, SE = 28ms]. This effect, while of interest, should be treated with caution: a model comparing results from Experiment 2 and 3 failed to find a significant main effect of Structure, $t = 1.54$, or interaction between Structure and Load, $t < 1$.

2.4.2. Discussion

Increasing the load imposed by the concurrent task led to comprehension impairments, specifically for questions that followed garden-path items. In contrast, eye movement behaviour replicated that seen in Experiment 2, with the exceptions of a trend towards a lingering garden-path effect in total reading durations on region 2 (in the second sentence), and the absence of a garden-path effect in go-past durations on the shorter region 1.

The results demonstrated that under difficult task conditions, readers are more prone to making good-enough comprehension errors. Comprehension was only impaired at this increased level of difficulty; note however that combining Experiments 2 and 3, there is some evidence for a graded effect as load increased from none, to 2-back, to 4-back (as seen in Figure 2.3, Panel A). This is consistent with the difficulty of the 4-back task being past a tipping point: the added load makes it difficult to adopt the strategy of spending longer reprocessing the text. Without refuge to this, the participant has to rely on a good-enough parse, with stronger remnants of initial misinterpretations that have not been sufficiently pruned. Comprehension is impaired as a result – but only on syntactically complex items where the good-enough parse (based on the misinterpretation) proved to be incorrect.

2.5. General Discussion

Three experiments explored the processing of syntactic ambiguities, subsequent comprehension of those ambiguities, and how increasing the level of extrinsic cognitive load affected processing and comprehension. I summarise the main findings briefly before discussing them in the context of the good-enough processing framework (Christianson et al., 2001, 2006; Ferreira & Patson, 2007; Slattery et al., 2013).

The pattern of eye movement behaviour and performance on the comprehension questions across no-load conditions were broadly in line with previous research (Christianson et al., 2001; Clifton et al., 2007; Slattery et al., 2013). Participants showed longer reading times in critical regions when reading garden-path sentences compared to sentences with a disambiguating comma. Nevertheless, temporary ambiguities lingered and disrupted comprehension: identical questions were answered incorrectly more often when the item contained no disambiguating comma (compared to identical items with commas) – although these results were less pronounced than in previous research.

Beyond this, the purpose of this study was to explore how the presence of an extrinsic memory load impacted on the processing and comprehension of temporarily syntactically ambiguous sentences, and how this intersected with whether sentences were presented in isolation or in a short passage. There were three main findings. First, extrinsic load affected eye movements in Experiment 1, with longer looking times in later measures, although this pattern was not observed when reading passages in Experiments 2 and 3. Second, while the 2-back memory task in Experiments 1 and 2 had little impact on question accuracy, increasing load using a 4-back task in Experiment 3 did affect comprehension detrimentally. Importantly, this was specific to those questions that tapped an initial misinterpretation following ambiguous sentences. Third, embedding sentences

into passages affected eye movements: there was no evidence of the classic first pass effect. Surprisingly, despite the context of the second sentence providing clear information that should have resolved any lingering ambiguity, comprehension did not improve. In fact, comprehension declined in Experiment 2 compared to Experiment 1.

Experiment 1 demonstrated that a concurrent task can disrupt processing of sentences containing a temporary syntactic ambiguity. This need for additional reprocessing produced a pattern of eye movement behaviour similar to that seen in people with lower working memory spans, including older adults (DeDe, 2014, 2015; Kemper et al., 2004; Kliegl et al., 2004), suggesting a link between changes in reading strategies and changes in resource allocation (cf. Smiler et al., 2003). The shift to increased re-reading with load in Experiment 1 was not, however, observed to the same extent when reading passages in either Experiment 2 or 3. Although not expected, it is possible that passages led no-load participants to show a similar pattern (more superficial first pass; more time spent re-reading), masking any difference between groups. This would explain the lack of evidence of a first pass reading effect in Experiments 2 and 3 – an absence that has been seen previously (Slattery et al., 2013). It may also explain the surprising reduction in comprehension in Experiment 2 compared to Experiment 1.

The 4-back task in Experiment 3 resulted in a clear reduction in comprehension: questions that tapped garden-path item misanalyses were answered poorly, relative both to the same items without load, and to control items with commas. There was evidence for a graded reduction in comprehension accuracy as extrinsic load increased, when comparing no-load and 2-back participants from Experiment 2 with 4-back participants from Experiment 3. I had predicted that the 2-back task would disrupt processing, producing a less veridical parsing of the ambiguity, and hence more good-enough comprehension errors. Arguably, the 2-back task may not have been sufficiently taxing to

impact on comprehension for these young skilled readers, especially in Experiment 1 where the reading material was easier. Instead, the additional time spent re-reading was sufficient to drive comparable comprehension to those reading without concurrent load. In Experiment 3, however, where the load was more substantial, processing was sufficiently disrupted to affect comprehension. As engaging in extra reprocessing is costly, a good-enough parse is settled on more readily. Notably, comprehension accuracy in the 4-back condition was only significantly worse for questions following garden-path passages: the comprehension deficit was not general, but specific to syntactically complex texts. This is entirely consistent with a good-enough framework of syntactic processing, and concurs with evidence from older readers (Christianson et al., 2006). The presence of a comma blocks the initial misinterpretation from being built, leaving nothing to linger when tapped by the question; without a comma, the misinterpretation should be built, and if it is not pruned, it remains to disrupt comprehension.

The final key finding is the reduction in comprehension of ambiguous sentences presented within a passage (Experiment 2) compared to similar sentences presented in isolation (Experiment 1). This observation provides further support for the good-enough framework, showing that task demands influenced the extent to which temporary syntactic ambiguities were resolved. As mentioned earlier, this was not a within-subject or within-item design, although the use of linear mixed effects models will overcome these random effects to an extent. Nevertheless, it is apparent that the addition of the second sentence certainly did not increase comprehension accuracy. This is a surprising finding, as the second sentence in the passages in Experiment 2 was designed to explicitly clarify any lingering ambiguity. To repeat an example from earlier, (2.5):

(2.5) While the father changed the baby(,) that was cuddly played with its toys. The father had to finish changing his clothes to pick up and look after the baby.

When a person answers yes to the question *Did the father change the baby?* after reading (2.5), they either ignored the information in the second sentence (that the father is getting changed *himself*, heavily implying he is not changing the baby as well), or maintained an interpretation that contains elements of both the correct one and the initial misinterpretation (of washing the baby).

It is possible that participants did read the second sentence less carefully, especially seeing as questions only tapped the content of the first sentence. Alternatively, evidence consistent with the latter explanation of maintaining both interpretations is provided by Patson et al. (2009). They found that when asked to paraphrase similar garden-path sentences to those used here, participants often gave an account that supported *both* the father getting changed himself, and changing the baby (see also Malyutina & den Ouden, 2016). In general, the data here demonstrated that people form a good-enough interpretation of the first sentence, and do not reconsider this interpretation despite reading conflicting information in the second sentence. This resulted in errors when answering questions, consistent with initial misinterpretations being detected and reanalysed, but not always being satisfactorily pruned (cf. Slattery et al., 2013). In short, having an additional sentence led to more superficial reading: initial misinterpretations failed to be pruned on the expectation that a complete understanding would be gained from to-be-read information. These misinterpretations then interfere with comprehension.

To reiterate, however, comprehension across experiments was much better than in previous research (e.g., Christianson et al., 2001, 2006), even for participants in the load conditions. The young skilled participants in these experiments might not be impaired by the garden-path sentences – or indeed the concurrent task. Accordingly, they used effective strategies to overcome the (relatively easy) garden-paths in this study,

explaining the absence of load effects in comprehension accuracy in Experiments 1 and 2. Individual differences have been found to be better predictors of eye movements than properties of the text being read (Kuperman & Van Dyke, 2011); it may be that high reading skill is enough to overcome both syntactic ambiguity, and moderate load via a dual-task. The question of individual differences in covered more in later chapters.

Overall, the findings presented in this chapter are consistent with a good-enough framework of sentence processing. Increasing both the length of the text and the concurrent load led to a pattern of more superficial reading behaviour, as seen in the eye movement record. This led to poorer comprehension, although at the low levels of extrinsic load induced by the 2-back task, participants were able to overcome this by increasing re-reading. More generally, these results add to evidence questioning the extent to which the garden-path model can account for syntactic processing under all conditions (cf. Christianson et al., 2001; Swets et al., 2008; van Gompel & Pickering, 2007). The comprehension data show that ambiguities are not always resolved, even if participants re-read to compensate for load. What is less clear is why we sometimes spend longer on the first pass of disambiguating regions, and sometimes instigate a regression to help with disambiguation, as seen in increased go-past durations (cf. Clifton et al., 2007). The absence of a clear first pass effect in the 2-back condition in Experiment 1, and all conditions in Experiments 2 and 3, indicate that task demands (intrinsic *or* extrinsic to the reading task) may influence this decision. Experiment 1 also demonstrates that the presence of an extrinsic load can result in a more underspecified initial parse of an ambiguous sentence, followed by additional re-reading in order to complete resolution.

Having concentrated on how these results support the good-enough framework, a key question remains: what does it actually mean for syntactic processing to be *good enough*? Slattery et al. (2013) suggested that rather than syntactic representations

themselves being underspecified, it is the process of pruning misinterpretations that is merely good enough. On this view, when comprehension goes astray on questions like those used here, the syntactic representation is in itself intact, but the complete reprocessing required to prune the initial misinterpretation has not happened. In consequence, the garden-path is not fully overcome. Further support for this comes from Logačev and Vasishth's (2016b) modelling of data from Swets et al. (2008). Their modelling provided tentative support for the idea that when comprehension is impaired, it may not be that representations are underspecified, but instead were never specified at all. If there is no specification of the syntactic structure, we have no information to help us answer the question – and rely simply on a guess (see also Metzner et al., 2016, who suggest that the parser can either choose to explore different interpretations, or alternatively, resort to a good-enough strategy). The decision not to specify may be driven by a mechanism that is both probabilistic (i.e., not totally predictable for a certain structure) and task-dependent (Logačev & Vasishth, 2016a).

The findings here point to a similar conclusion: on a given trial, syntactic processing is completed to an extent that is dependent on both the difficulty of the text to be parsed, and the resources available (here, this is the concurrent task; elsewhere, it could also be individual differences). The 2-back condition imposed only moderate load, meaning that sufficient resources remained to attain reasonable comprehension – although additional re-reading time was required to achieve this, suggesting that the initial parse was insufficient. As subject-object ambiguities such as these are more difficult to leave unrepaired (unlike, for example, attachment ambiguities used in von der Malsburg & Vasishth (2013)), the 2-back participants have to allocate additional time to reanalysis, rather than leaving their representation of the sentence underspecified. However, when the load was increased in Experiment 3, the parser chose the faster option of a more

heuristic-based interpretation (cf. Karimi & Ferreira, 2016; Logačev & Vasishth, 2016a; Metzner et al., 2016). In turn this produced a cost: poorer comprehension.

Another question in this chapter was whether initial misinterpretations lingered, and if so why? Slattery et al. (2013) argued that they did and due to slow pruning rather than memory fallibility. The absence of any effect of the 2-back load on comprehension in these data stands against the idea of memory fallibility; however, the limited impact of the 2-back task may be because to-be-recalled items were not semantically similar to the content of experimental sentences, reducing the extent to which they caused interference (cf. Gordon et al., 2002). The notion of slow pruning is plausible: the initial misinterpretation is not deleted even in light of opposing information. Participants did not look longer in Region 2 on garden-path items, showing no surprise at this text that conflicts with the garden-path misinterpretation in the first sentence. Despite this, participants continued to make comprehension errors when questions tapped that garden-path misinterpretation. As Patson et al. (2009) found, readers seem comfortable with both the initial misinterpretation, and the seemingly contradictory correct parse.

Karimi and Ferreira (2016) offered an interesting perspective on what "good enough" actually means. They conceptualised good-enough processing as a reader aiming to reach an equilibrium, which is only significantly revised if it is considered to be incorrect in light of new information. Put another way, readers adopt heuristic-based processes. These are fast and efficient, until or unless they are forced into more laborious, algorithmic processing of sentence structure. This can explain the lingering of initial misinterpretations, and readers' satisfaction with an (ultimately incorrect) good-enough blend of syntactic representations. This would also explain the relative contributions of internal task demands (such as whether texts are presented as a sentence or a passage), external demands (in these experiments, the resource limitations imposed by the dual

task) and individual differences (not featuring in these experiments, but seen in ageing for example). All will influence the extent to which the reader relies on heuristic processes, the extent to which they therefore show more superficial reading, and in turn the risk of being susceptible to making more good-enough errors. These topics, as well as coherence between these results and alternative theories highlighted in Chapter 1, are also discussed more in later chapters.

More generally, Experiments 1 to 3 demonstrate the importance of combining eye-tracking with offline measures. Christianson et al.'s (2001) results stressed the importance of testing comprehension, and not assuming that accurate comprehension was indexed by increased reading durations and/or regressions (as was often the case before that study). The findings from eye-movements in this chapter offer further reassurance that the good-enough errors described by Christianson et al. are not an artefact of the presentation mode used in those studies (see also Patson et al., 2009). Importantly, however, relying on offline comprehension measures alone would have led to the false conclusion in Experiment 1 that the extrinsic memory load had no effect on the processing garden-path sentences. The pattern of eye movements seen when reading under load show very clearly that this was not the case. Similarly, in Experiment 3, the absence of an effect of added load in online processing does not negate its clear impact on eventual comprehension.

It is evident from the varying effects seen here, and in previous research, that online and offline measures tap different things. It is striking that comparatively little research has investigated the association between the two – that is, the extent to which differences in eye movement patterns are related to differences in comprehension. The findings presented here suggest that task demands influence the decision of whether to spend extra time reading a disambiguating region, or whether to regress to earlier parts of

the text to disambiguate what has been read. It is apparent from the garden path effect in go-past durations across all conditions, and the data presented in Tables 2.1 and 2.3 (of the proportion of trials with regressions out of, and into, the disambiguating verb) that regressions and re-reading are common when reading syntactically ambiguous sentences. Despite this, it is unclear whether these regressions are used to revise syntactic representations in the service of comprehension (e.g., Metzner et al., 2016), whether they serve to confirm what was read on the first pass (Christianson et al., 2016), or even whether they just allow extra time for processing (Mitchell et al., 2008). Based on the experiments presented here, good-enough processing is evident in both online and offline processing, but in different ways; gaining a better understanding of the link between the two (if one even exists) is crucial.

To conclude this chapter, the results here demonstrated that an extrinsic memory load disrupted eye-movement patterns in skilled readers when reading syntactically ambiguous sentences. Participants required additional re-reading of the text, resembling the pattern typically seen in older participants. Although recognition of the syntactic ambiguity was delayed by concurrent load, participants did not show a commensurate decrease in comprehension accuracy, except under the higher 4-back load. This suggests that, up to a certain point, skilled readers were able to use the added reading times in order to fully resolve the garden-path ambiguities. Eye movement behaviour and comprehension were also affected by the length of the text participants were asked to process. However the paucity of research relating eye-movements to comprehension – especially across ages – leaves open the question of how the processes tapped by online measures differ from those tapped by offline measures, and the impact that task demands and cognitive load have on each.

Chapter 3

Effects of task demands on syntactic ambiguity processing

3.1. Background

In Chapter 2, both eye movements and comprehension were affected by the task participants were performing, and the nature of the texts being presented (as isolated sentences or in short passages). Adding load via a concurrent n-back task in Experiment 1 led to more re-reading of the disambiguating verb of garden-path sentences. A more difficult 4-back task in Experiment 3 produced a marked disruption in comprehension, despite comprehension being virtually unimpaired with the 2-back task in Experiments 1 and 2. Furthermore, a garden-path effect on first pass durations when reading sentences in Experiment 1 was not seen when reading longer passages (containing similar sentences) in Experiments 2 and 3. There was also a reduction in comprehension when sentences were seen in passages, despite those passages containing further disambiguating information.

These results mirror evidence in the literature that isolated sentences are read differently to passages or paragraphs (Kuperman et al., 2013; Radach et al., 2008; Whitford & Titone, 2014). This may be because longer passages often contain more context, which facilitates prediction during first-pass reading (Whitford & Titone, 2014). Additionally, reading a longer passage requires more effort (e.g., in terms of memory load), and participants may speed up on first pass in order to account for this, returning to difficult words if necessary (cf. longer total reading durations on passages in Radach et al., 2008). If an unexpected, novel or syntactically illicit word appears, we may opt to

continue reading on the expectation that the ambiguity will be resolved in later text, rather than spend additional time reading the surprising word.

In Chapter 2, the results were interpreted as being consistent with Slattery et al.'s (2013) account that initial misinterpretations lingered and disrupted later processing. Nevertheless, Experiments 2 and 3 failed to replicate Slattery et al.'s findings of garden-path effects on the reflexive region of the second sentence. Three outstanding questions from Chapter 2 will be explored in this chapter. First, were the 2-sentence passages in Experiment 2 insufficiently complex to cause disruption for the participants who took part? Participants performed well on the comprehension questions, with notably higher accuracy than in previous research (Christianson et al., 2001, 2006), even if accuracy was lower than in Experiment 1. Therefore, it may be that for highly skilled readers, the addition of a second sentence was not sufficiently taxing, and this may explain the absence of the lingering eye movement effects seen in Slattery et al. (2013). This was explored in Experiment 4, where the two sentences of Experiments 2 and 3 were placed into longer, four-sentence passages.

The second question that was looked at in Experiment 4 was also posed by Slattery et al. (2013): how long are lingering effects seen for? Slattery et al. saw that misinterpretations from a garden-path sentence lingered when reading a second sentence – but here, I asked if this lingering would be seen if an additional intervening sentence was placed between these two sentences? Or would the misinterpretation decay before then? Also of interest was whether the content of the intervening sentence would affect eye movement records, and eventual comprehension. If the intervening sentence were biased towards the initial misinterpretation (by perpetuating the idea from *While the father changed the baby that was cuddly played with its toys* that the father is changing the baby, rather than getting changed himself), would this affect the extent to which the

misinterpretation lingered? While the longer passages tested longer-term lingering in the eye movement record, they did not test whether good-enough misinterpretations linger after a more significant delay. Experiment 5 assessed whether misinterpretations would linger over a longer period, testing recollection of experimental items after a 10 minute delay (rather than using questions immediately after each item). If lingering was still evident in comprehension, this would lend support to Slattery et al.'s conclusion that their results were more due to insufficient pruning of misinterpretations (causing these misinterpretations to linger, having been initially built), and less due to simple decay in memory over time (where we would still expect poorer comprehension for garden-path items over time, but should also see poorer comprehension of items that contained disambiguating commas following the delay).

Third and finally, by assessing comprehension after a delay and not after reading each text, Experiment 5 directly assessed how task demands affect eye movements. In the experiments in Chapter 2, participants knew that they would be asked specific questions after each item, and they may have read texts more carefully as a result (especially on disambiguating regions). By removing these questions, how would this affect eye movements during reading? Experiment 5 used the same stimuli as Experiments 2 and 3, allowing a direct comparison with these earlier experiments. A reasonable expectation is that participants would read texts more superficially if they were not expecting to be tested on those items. This may depend on individual differences: people with lower working memory spans may be more likely to underspecify and read superficially than those with higher spans (Nicenboim et al., 2016; von der Malsburg & Vasishth, 2013). By measuring memory span, the effect of individual differences could also be explored – both on eye movements, and on eventual comprehension.

Experiment 4

3.2. Introduction

Slattery et al. (2013) found evidence that initial misinterpretations from a first sentence lingered and disrupted eye movements on a second sentence that was inconsistent with those misinterpretations. They also discussed how long misinterpretations would linger for. In Chapter 2 (as well as in other studies: Christianson et al., 2001, 2006; Malyutina & den Ouden, 2016; Patson et al., 2009), initial misinterpretations disrupted comprehension, even when participants had no concurrent memory load. Experiments 2 and 3 did not replicate Slattery et al.'s findings of lingering in the eye movement record on a second sentence. As discussed earlier, this lack of replication could have reflected differences between Experiment 2 and Slattery et al.'s study. First, Slattery et al. found an interaction between syntactic ambiguity (i.e., Structure) and plausibility: the effect in the second sentence was only found in items where the initial misinterpretation was plausible (e.g., *While the boy washed the dog,* vs. *washed the sun*). The items used in Experiment 2 contained no such manipulation, and all items were designed to be plausible. Second, the presence of questions tapping the initial misinterpretation in Experiment 2 (not used in Slattery et al.'s study) may have altered reading strategies and reduced the effect. Finally, participants here were highly-skilled readers and performed well on the comprehension test. Having parsed the initial sentence, they may have cleared up the initial misinterpretation more effectively (and so processing of the second sentence was not disrupted). Furthermore, with only two lines of text to read, participants may have been able to reactivate the structure of the first sentence more faithfully, making the second sentence less surprising.

One way of exploring this final explanation is by extending the region between the temporarily ambiguous sentence, and the sentence containing the explicitly disambiguating reflexive term. As mentioned, the lingering effect in Slattery et al. (2013) was only seen if the first sentence was temporarily ambiguous *and* that ambiguity was plausible. On Slattery et al.'s view, the lingering misinterpretation of the temporary ambiguity remained to interfere with processing of the second sentence. They did acknowledge that this result could stem from memory fallibility rather than for syntactic reasons (or indeed, both accounts could be correct). If the participants in Experiment 2 did not find the 2-sentence passages difficult, and were able to faithfully recall the content of the first sentence when reading the second, adding intervening material between the sentences may make this task more difficult. This would be particularly true if that material was biased to perpetuate the initial misinterpretation. For example, take (3.1) (underlined areas are those where eye movements are measured):

(3.1) While Rebecca woke up(,) the neighbour that was in bed shouted out loudly. The neighbour was often disturbed by **Rebecca/sunrise** nowadays. Rebecca was fully awake now and heard the shouting next door.

The first and third sentences in (3.1) are taken from Experiment 2. The second sentence adds further material to read. If the word in bold is *sunrise*, this provides no evidence that Rebecca woke *herself* up, or that Rebecca woke up the neighbour (i.e., it is neutral as to whether the initial misinterpretation or the correct interpretation is accurate). In contrast, if the word is *Rebecca*, this biases the reader towards the (incorrect) interpretation that Rebecca woke up the neighbour. In this case, participants should show longer reading durations on *fully awake*, which refers back to the alternative (and correct) interpretation that Rebecca woke up herself. This would resemble Slattery et al.'s (2013) results, a pattern not seen in Experiment 2.

Another advantage to using longer passages is to gain further insight into how task conditions affect syntactic processing. Many experiments studying temporary syntactic ambiguities present sentences in isolation (cf. Radach et al., 2008); studies using longer passages tend to do so specifically to manipulate discourse context, rather than simply to look at the effect of embedding the sentence in a longer passage. This was part of the motivation for using two-sentence passages in Experiment 2 (and Experiment 3), as a comparison to sentences being presented in isolation, as in Experiment 1. One limitation of Experiments 2 and 3 was that simply adding one further sentence might have been insufficient to alter reading strategies extensively, especially if the participants are highly skilled readers. A four-sentence passage, in contrast, provides double the material to deal with. This may produce even more superficial reading on first pass than seen before, with revisits made to key parts of the text to attain comprehension (cf. Radach et al., 2008; Whitford & Titone, 2014). Nevertheless, if comprehension accuracy was poorer in Experiment 2 than Experiment 1 because of the longer passages, this finding should be replicated (if not extended) in the even longer passages of Experiment 4. With more material to read, participants are more likely to rely on good-enough representations[19] of the text, containing remnants of initial misinterpretations.

These interests led to three main questions posed in this experiment. First, would the lingering effects observed by Slattery et al. (2013) be seen, despite not being seen in Experiments 2 and 3? The effect would be predicted to appear in Experiment 4 if the longer passages meant that participants could not easily reinstantiate the surface content of the garden-path sentence, and relied instead on a good-enough representation that contained a lingering misinterpretation. I therefore expected to replicate Slattery et al.'s

[19] This would also be consistent with alternative, similar theories; for example, noisy-channel models may explain this as longer texts causing increased uncertainty about the text (cf. Levy et al., 2009).

finding of small but significant lingering effects in the reflexive region [in (3.1), *fully awake*].

Second, how would a manipulation of bias provided in that intervening sentence affect both eye movements (especially on the sentence containing a reflexive term), and comprehension? I predicted that if the intervening sentence biased towards the temporary ambiguity (by providing background context relevant to the misinterpretation, rather than the correct interpretation), this would produce more of the lingering effects seen by Slattery et al. Furthermore, this biased material should keep misinterpretations active, creating more interference when answering comprehension questions. Accordingly, participants should incorrectly answer Yes more often for garden-path items with an intervening sentence that perpetuates the misinterpretation, compared to when the intervening sentence is neutral. Similarly, the lingering eye movement effects reported by Slattery et al. (2013) on the reflexive term should be stronger in the biased condition (versus the neutral condition). This is because the initial misinterpretation is reactivated, making its interference stronger; as such, the presence of a reflexive term promoting the correct interpretation will be more confusing, and should be read for longer. However, there is little reason to expect any effects of bias on the comma items: except in a few errant cases, the punctuation should block the misinterpretation from being built in the first place.

Finally, how would comprehension in this experiment compare to no-load participants in Experiments 1 and 2? The second sentence of passages in this experiment resembled the first sentence of the passages in Experiment 2, and also a subset of the sentences ("long" RAT items) presented in isolation in Experiment 1. The fourth sentence here was the same as the second sentence used in Experiment 2. The difference was therefore a) the addition of a first sentence, which added content, but content that was

irrelevant to the comprehension questions asked, and b) the addition of an intervening sentence. Where these intervening sentences were neutral, they should not have affected comprehension of the items in any way other than by adding more material to read (in contrast, intervening sentences that perpetuate the ambiguity were *designed* to reduce comprehension). This offers a comparison of comprehension accuracy of no-load participants in Experiment 1 and 2, to accuracy on neutral items in this experiment: the prediction was for accuracy to be lower in this experiment than in Experiment 1, and to be poorer even than in Experiment 2, due to the longer stimuli.

3.3. Method

3.3.1. Participants

Sixty-four students at the University of Oxford took part in this experiment (41 female; mean age = 22.2 years, range 18 – 30 years); as in Chapter 2, the students were across psychology (n = 4), other science subjects (n = 21), humanities (n = 29) and social sciences (n = 10). All were recruited by email and posters, and received £5 or course credits as compensation.

3.3.2. Materials and procedure

3.3.2.1. Eye-tracking methodology. An Eyelink 1000 eye-tracker (SR Research) recorded right eye movements at 1000 Hz during the reading task. Participants sat 55cm from the screen. Pupil and corneal reflection thresholds were adjusted according to the eye-tracker instructions. The eye-tracker was then calibrated and validated to <.75° using nine fixation crosses. Calibration was repeated during the experiment when this threshold was lost. The task was programmed using Experiment Builder (SR Research).

3.3.2.2. Reading task. The stimuli comprised 16 experimental items (all in Appendix A) and 32 filler items of a similar structure, but with no syntactic ambiguity. The font, size, spacing and viewing angle were all similar to Experiments 2 and 3. Each experimental item was a four-line passage with the structure given in Table 3.1. The second (garden-path) and fourth (disambiguating) sentences in the passage resemble those used in Experiments 2 and 3, and Slattery et al. (2013), and items were developed from the passages used in Experiments 2 and 3. The context sentence was added as a comparison to most experiments where the garden-path sentence appears first; the intervening sentence contained a manipulation of bias towards the ambiguity.

Table 3.1.

An example experimental passage in Experiment 4. Manipulations are in bold.

Context sentence	Rebecca had lived near her neighbour for twenty years.
Garden-path sentence	While Rebecca woke up**(,)** the neighbour that was in bed shouted out loudly.
Intervening sentence	The neighbour was often disturbed by **Rebecca/sunrise** nowadays.
Disambiguating sentence	Rebecca was fully awake now and heard the shouting next door.

There were two manipulations. One was Structure: whether the garden-path sentence had no comma (*garden-path* condition) or had a comma (*comma* condition). The second was Bias: whether the intervening sentence biased towards the initial ambiguity (*biased* condition) or was neutral between the two possible interpretations (*neutral* condition). For example in Table 3.1, if the intervening sentence states that the neighbour is often disturbed by Rebecca, this perpetuates the ambiguity that Rebecca woke up the neighbour. In contrast, if the intervening sentence states that the neighbour is often disturbed by sunrise, this does not favour either Rebecca waking up or Rebecca waking

up the neighbour. The target words used were matched for length and frequency, with no significant difference between pairs on either, both $t < 1$.

The passages were made into four lists using a Latin Square design, with the second sentence either *garden-path* or *comma*, and the third sentence either *biased* or *neutral*. However, due to a coding error, both the third and fourth group of participants saw the third list of items, and no participants saw the fourth list. While it was unfortunate that the conditions were not fully crossed, every item was seen by at least some participants as a garden-path and as a comma item, and as a biased item and as a neutral item. Furthermore, the random effects structure of linear mixed effects models should account for any item-level differences. The possible effects of this error are mentioned in the Discussion of this experiment.

After drift correction, participants read a passage on the screen, before pressing a button on a controller to advance to a question. Comprehension was tested with follow-up questions, as in Chapter 2. Each of the 16 experimental items were followed by a yes-no question, for which the expected answer was always No; for example, the passage in Table 3.1 was followed by *Did Rebecca wake up the neighbour?*. Filler items were also followed by a question. Participants indicated their response to questions, again using the controller, before the next drift correction would begin.

3.4. Results

Eye-tracking records were analysed using Data Viewer (SR Research). Fixations above 1200ms were deleted; fixations below 40ms were merged with adjacent fixations or if not possible, deleted. One participant was removed due to poor calibration, leaving 63 in the final analysis. Data were analysed using linear mixed-effects models (LMEs) in

the *lme4* package (Bates et al., 2015) in R, with the same log transformation, model criticism and maximal random effects structure as discussed in Chapter 2. All models had Structure (*garden-path* vs. *comma*) and Bias (*biased* vs. *neutral*) centred and entered as fixed-effects.

3.4.1. Comprehension accuracy

Figure 3.1 shows mean accuracy across Structure and Bias. A garden-path effect is clear on biased items, but is much smaller for neutral items. There was an overall main effect of Structure, $z = 3.72$, due to significantly reduced accuracy on garden-path sentences ($M = 72.6\%$, $SE = 2.0\%$) than comma sentences ($M = 80.4\%$, $SE = 1.8\%$). There was no effect of Bias, $z = 1.09$, and no significant interaction, $z = 1.42$. Despite the lack of an interaction, the original hypothesis was that bias would specifically impact on accuracy of garden-path sentences. To test this, *post hoc* models for the two Structure conditions separately found that while Bias significantly affected accuracy in garden-path items ($M_{Bias} = 68.4\%$, $SE_{\text{Bias}} = 2.5\%$; $M_{Neut} = 76.3\%$, $SE_{Neut} = 2.7\%$), $z = 2.46$, there was no such effect in comma items ($M_{Bias} = 81.4\%$, $SE_{Bias} = 2.5\%$; $M_{Neut} = 79.3\%$, $SE_{\text{Neut}} = 2.6\%$), $z < 1$[20]. Care should be taken when interpreting these, given the lack of a significant interaction.

One additional question was whether overall accuracy was poorer than in Experiments 1 and 2. Descriptive data are presented in Table 3.2. An LME model found that accuracy was significantly poorer than for participants reading single sentences in Experiment 1, $z = 4.91$, but not poorer than for participants reading two-sentence

[20] One concern is that not all items were seen under all conditions due to the error in coding the fourth list of items. To address this, these models were re-run including only items that had been seen in both biased and neutral conditions (half of the total set of items): the pattern of results presented here were replicated, suggesting that this oversight is not driving the results. It may however have underpowered the overall Structure-by-Bias interaction, as discussed later.

passages in Experiment 2, $z < 1$. There are of course several changes between these experiments: in Experiment 4, no participants faced a concurrent memory load, the items all contained RAT (and not optionally transitive) verbs, and the items contained the added manipulation of bias. However, additional models found that the significant reduction held even if only considering no-load participants from before, $z = 3.55$, only considering RAT items from before, $z = 4.49$, and only considering the neutral items here, $z = 3.68$ (these were not run concurrently to avoid issues with power and model convergence, due to the small number of observations). There were however no significant interactions between Structure and the comparison of Experiment 1 and Experiment 4, and so the effect was across garden-path and comma items.

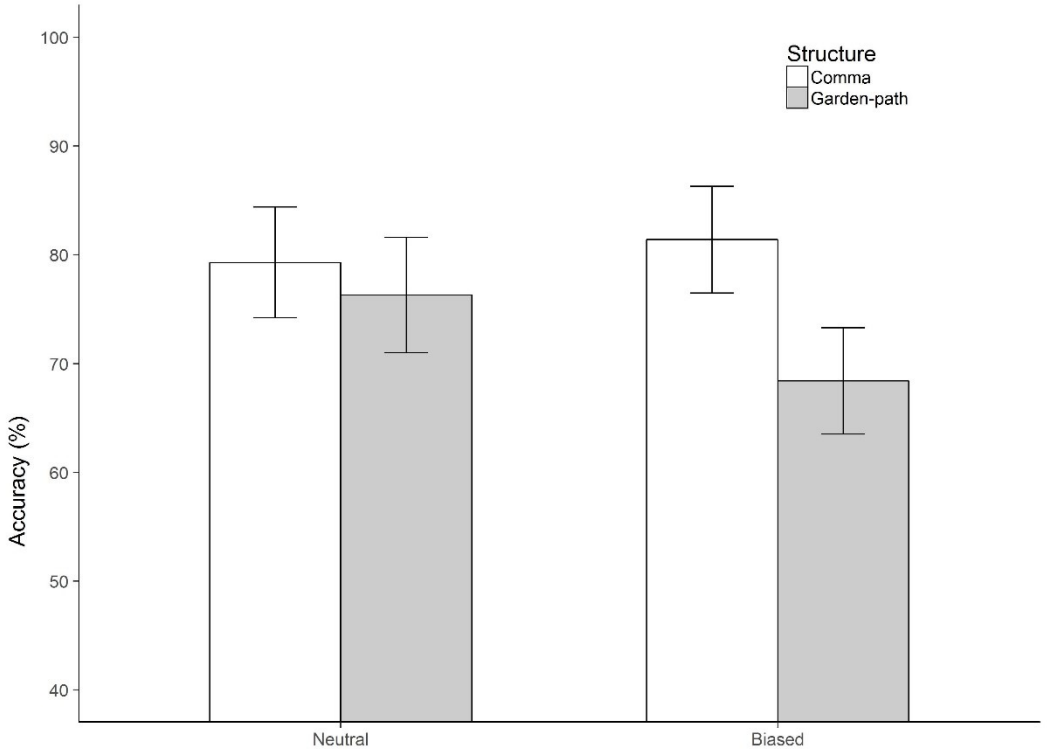

Figure 3.1. Comprehension accuracy on experimental items, split by Bias, and Structure (white bars for comma items). Error bars indicate 95% confidence intervals.

Table 3.2.

Mean percentage comprehension accuracy (standard errors in brackets) for Experiments 1, 2 and 4, by Structure.

	Comma	Garden-path
Experiment 1	85.2 *(1.2)*	80.3 *(1.4)*
Experiment 2	88.3 *(1.5)*	75.0 *(2.0)*
Experiment 4	80.4 *(1.8)*	72.6 *(2.0)*

3.4.2. Eye-tracking results

3.4.2.1. Eye-tracking measures. There were three critical regions for eye movement analyses, with four measures taken for each. For the example in Table 3.1 above, these are:

Region 1, garden-path region: (*shouted out*) – the disambiguating verb in the second sentence and the word that follows this. (This approximately resembles the *long* version of Region 1 in Experiments 2 and 3, and captured spillover effects in these longer passages).

Region 2, intervening region: (*Rebecca/sunrise*) – the biased or neutral word in the third sentence.

Region 3, reflexive region: (*fully awake*) – the disambiguating word in the fourth sentence and following word. (This approximately resembles Region 2 of Experiments 2 and 3).

The four measures were *first pass duration* (the duration of all fixations during the first visit to a region), g*o-past duration* (the total duration before exiting a word to the right, including time spent on regressions to the left), *total reading duration* (the total

amount of time spent in a region, across all visits), and *first pass regressions out* – the proportion of trials where a regression was made to the left after entering the region.

3.4.2.2. Analysis of Region 1 ["garden-path" region]. First, for comparison with Experiments 1 – 3, descriptive data for Region 1 are presented in Table 3.3. There were main effects of Structure on regressions in, $z = 2.08$, and regressions out, $z = 3.12$ of Region 1, but no effect on first fixation duration or skipping rate.

Table 3.3.

Additional descriptive statistics for Region 1 eye movements in Experiment 4, by Structure and, where relevant, Ambiguity (standard errors in brackets). All measures are defined in Table 1.1.

	Comma		Garden-path	
First fixation duration	214 *(4)*		217 *(5)*	
First pass duration	331 *(11)*		310 *(10)*	
Go-past duration	820 *(27)*		1249 *(42)*	
% skipping	10.7 *(1.4)*		12.3 *(1.5)*	
% regressions out	43.8 *(2.3)*		54.5 *(2.4)*	
	Neutral	*Amb.*	*Neutral*	*Amb*
% regressions in	54.2 *(3.3)*	61.4 *(3.2)*	65.3 *(3.3)*	64.0 *(3.2)*
Total reading duration	695 *(29)*	784 *(35)*	960 *(43)*	958 *(42)*

First-pass, go-past and total reading durations are presented in Figure 3.2, which suggests no garden-path effect in first-pass durations, but the usual effect on go-past and total reading durations. Indeed, similarly to Experiments 2 and 3, LME models[21] found

[21]There should not be an effect of Bias in Region 1 on the early measures (first-pass and go-past durations), as the participant will not have yet reached the third sentence where this manipulation occurs. The models

that Structure did not affect first-pass reading times, $t = 0.70$, but did affect both go-past, $t = 3.54$, and total reading durations, $t = 5.76$. For both later measures, the effect was in the expected direction with longer reading times in the garden-path condition [M_{GoPast} = 1249ms, SE_{GoPast} = 42ms; M_{Total} = 959ms, SE_{Total} = 32ms] than the comma condition [M_{GoPast} = 820ms, SE_{GoPast} = 27ms; M_{Total} = 741ms, SE_{Total} = 25ms]. There was also a main effect of Structure on regressions out of Region 1, $z = 3.12$, with regressions out of Region 1 on more garden-path trials ($M = .55$, $SE = .02$) than comma ones ($M = .44$, $SE = .02$). For total reading time, there was no effect of Bias, $t < 1$, but there was a trend towards an interaction, $t = 1.92$, again visible on Figure 3.2. This trend reflected significantly shorter reading durations for comma sentences in the neutral condition, compared to the biased condition, $t = 2.57$; there was no such difference for garden-path items. As this interaction was not seen in either first pass or go-past durations, the effect can be attributed to differences in the duration of revisits to Region 1.

3.4.2.3. Analysis of Region 2 ["intervening" region]. The only significant effect found on Region 2 was a Structure*Bias interaction in first pass durations, $t = 2.03$; however, this effect was driven primarily by one outlier item, and removing the item from the analysis eliminated the significant effect, $t = 1.32$. There was no main effect of Structure, $t = 1.06$, suggesting that there was no general slowdown in the garden-path condition. The absence of any other effects (all $t/z < 1$) suggests that the two Bias conditions were well-matched.

3.4.2.4. Analysis of Region 3 ["reflexive" region]. There was only one effect on the duration-based measures: a main effect of Bias on first-pass reading times, $t = 2.67$ (all others $t < 1.5$). However, this was not observed in either go-past or total reading

for these two measures therefore only included Structure as a predictor. The models were repeated with Bias included, with no change in the pattern of results.

durations, and similarly to above, appears to be driven by one item (note: a different item from the outlier discussed above); removing the item eliminated the effect, $t = 1.23$.

For regressions out of the region, neither main effect was significant, but the interaction between Structure and Bias approached significance ($z = 1.92$, $p = .055$). This marginal interaction is visible in Figure 3.2: there is an effect of Bias in comma items, $z = 2.01$, but not in garden-path items, $z < 1$. For comma items, participants made regressions out of Region 3 on a significantly smaller proportion of neutral trials ($M = .24$, $SE = .03$) than on biased items ($M = .34$, $SE = .03$).

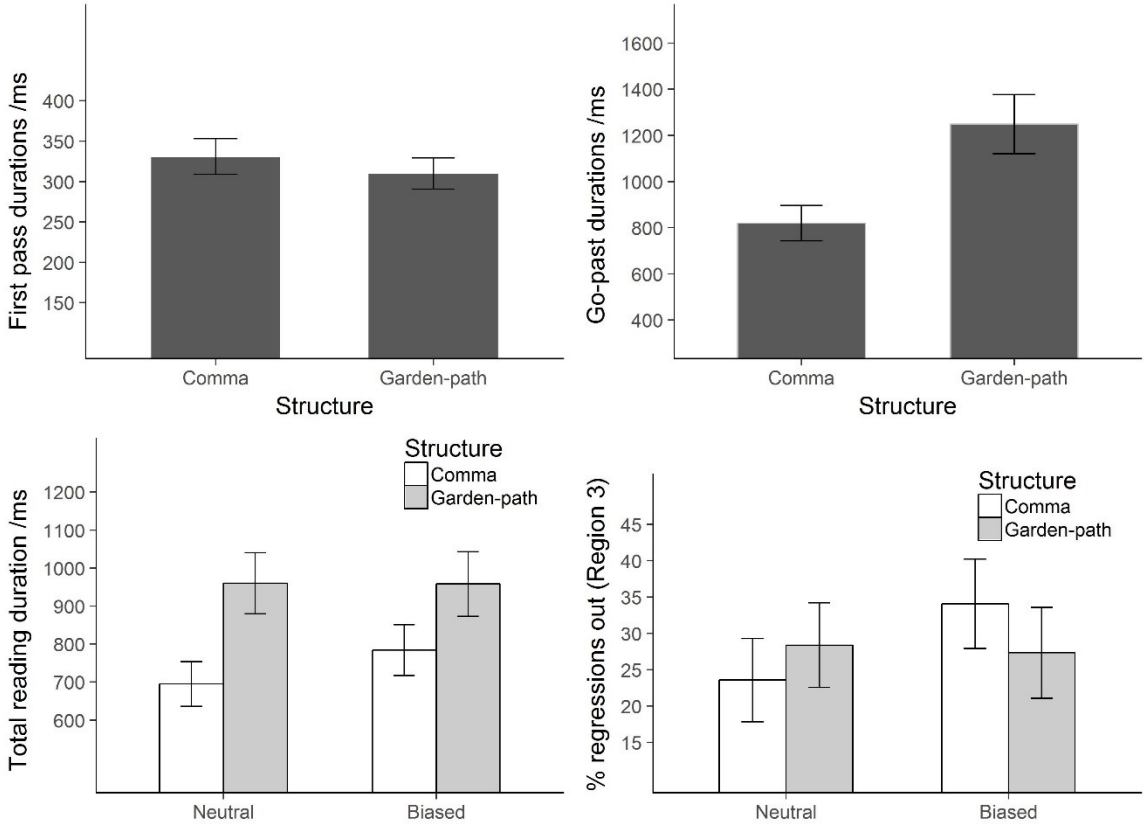

Figure 3.2. First pass (top left), go-past (top right), and total reading durations (bottom left) for Region 1, and first pass regressions out of Region 3 (bottom right). All are split by Structure; the bottom two also by Bias. Error bars indicate 95% confidence intervals.

3.5. Discussion of Experiment 4

Experiment 4 looked at how the results presented in Chapter 2 were affected by a move to four-sentence passages. In particular, how would both comprehension and eye movements be affected by adding intervening material that perpetuated the temporary syntactic ambiguity? The results were broadly similar to those in Chapter 2. Participants spent longer reading the critical region of syntactically ambiguous sentences than the identical region of unambiguous comma items, but still incorrectly answered Yes more often to questions that tapped the syntactic ambiguity. Neither embedding the item in a longer passage, nor manipulating whether the ambiguity was perpetuated in the third sentence, had any considerable effect on eye movement patterns. There was some evidence that the manipulation affected participants' representations of the items when answering questions. There was no significant Structure*Bias interaction, suggesting that caution is necessary when interpreting the result, but the effect of bias was stronger in garden-path items. This suggests that a biased third sentence may have had a small effect: the initial misinterpretation is more likely to remain active, and to interfere with comprehension when tapped by a question.

First, it is worth commenting on a limitation of this experiment. The coding error described in the Method section means that items were only seen in three out of the four possible conditions (that is, comma vs. garden-path, crossed with neutral vs. biased). This makes interpretation of the results difficult in both directions: first, significant results may be unreliable; alternatively, effects may be underpowered given the reduced number of observations in each cell during analysis. The latter issue could mean that the absence of predicted effects (such as the Structure*Bias interaction in comprehension accuracy, for which there was only a trend) may reflect a lack of power to detect these effects. To an extent, mixed effects modelling will account for this: models include estimates for

random effects due to the items used, and the effects of bias and structure on each item individually. Furthermore, the additional analyses presented on comprehension accuracy data (including analyses only exploring items seen in both conditions) allow for clearer interpretation of the results. Nevertheless, this remains a caveat in discussing the results for this experiment.

The comprehension accuracy data provide some support for the hypothesis that more comprehension errors will occur when an intervening sentence perpetuates the temporary syntactic ambiguity. To repeat the key part of the garden-path example given earlier:

(3.1a) While Rebecca woke up(,) the neighbour that was in bed *shouted out* loudly. The neighbour was often disturbed by ***Rebecca/sunrise*** nowadays. Rebecca was *fully awake* now and heard the shouting next door.

Reading the word *Rebecca* in the second sentence perpetuates the misinterpretation that Rebecca woke the neighbour; reading *sunrise* is more neutral as to whether Rebecca woke up, or woke the neighbour up. Accordingly, participants seeing *Rebecca* are expected to make more errors by incorrectly answering Yes when asked *Did Rebecca wake up the neighbour?* If Slattery et al. (2013) are correct that initial misinterpretations linger, this should result in comprehension errors. A participant may reanalyse the first sentence, attaching *the neighbour* to *shouted out loudly*, but fail to clear up the trace of *Rebecca woke up the neighbour*. The biased intervening sentence should reactivate this misinterpretation, causing further confusion when the final sentence (and ultimately, the question) is reached. This should only be seen in garden-path items, which led to the prediction of a Structure-by-Bias interaction.

This Structure-by-Bias interaction in comprehension accuracy was not significant, and there is not therefore definitive evidence that the main effect of Structure was

stronger in the biased items. There was however a trend towards this result, supported by *post hoc* tests indicating that a main effect of bias was present in garden-path items, but not in comma items, as predicted. This therefore provides tentative support for the hypothesis. The initial misinterpretation remains active and disrupts comprehension, as demonstrated by the effect of Structure on comprehension accuracy here, and in Experiments 1 to 3 and other studies (e.g., Christianson et al., 2001; Patson et al., 2009). The addition of a biased third sentence produces (numerically) more comprehension errors in the garden-path condition, demonstrating that a misinterpretation has been built, is reactivated, and then lingers more prominently when questions are subsequently asked. This is consistent with Slattery et al.'s (2013) conclusion that good-enough errors result from lingering misinterpretations that are insufficiently pruned. These results suggest that pruning may be less efficient if there is good reason to maintain the initial interpretation – for example, if new material keeps the proposition that *Rebecca woke up the neighbour* active as a viable interpretation.

A final finding was that comprehension in Experiment 4 was poorer than in Experiment 1, but was not worse than Experiment 2. There was no interaction with Structure, and so the reduction in accuracy was not limited to the garden-path items. This was not a within-item design, since this was not the main purpose of this experiment, and so some caution is required in interpreting this result. Nevertheless, the data here and in Experiment 2 do tentatively support the idea that embedding sentences in additional material (of any length) is linked to a reduction in comprehension, as seen in the sentence vs. paragraph comparison reported by Radach et al. (2008). It would be of interest to compare identical sentences embedded in various lengths of text, in order to test this theory more directly.

Beyond the comprehension data, Experiment 4 offered few additional findings to the data presented in Chapter 2 (and in previous research). Once again there was no evidence of a clear first-pass garden-path effect where sentences are presented in longer passages (cf. Experiments 2 and 3; Slattery et al., 2013). The initial sentences of items in Experiment 4 were not intended to provide any discourse context that would favour one interpretation or another, other than to introduce the characters that would feature in the passage. They should not therefore have biased towards or away from being garden-pathed (unlike the intended purpose of context sentences in studies such as Altmann et al., 1992). The effects in go-past duration and regressions out of Region 1 demonstrate that participants detected the syntactic ambiguity, but did not spend longer on their first pass of the disambiguating region to resolve this. The interpretation remains the same as in Chapter 2: participants either regress to re-read the ambiguous material, or progress to later sentences on the expectation that the ambiguity will be resolved by reading further.

The manipulation of bias in the intervening sentence had no clear effects on eye movements. It is possible that the bias was not sufficiently strong, and more detailed pre-testing of materials would be beneficial if considering future work. As in Chapter 2, there was no significant effect of Structure on this sentence containing a reflexive clause – despite this being predicted, and seen in Slattery et al. (2013)[22]. Again, Slattery et al.'s (small) observed effect may have been eliminated due to a shift in behaviour owing to the presence of questions, and differences in participants and items. There was a trend towards a Structure-by-Bias interaction in regressions out of this region. This did not reflect a larger garden-path effect in regressions on biased items; instead, there were

[22] For clarity, Slattery et al. (2013) did not find a *main effect* of Structure, but this was probably because half of their items had an implausible first sentence. If analysing just the plausible items (most comparable to those used here), Slattery et al. found an effect of Structure in first-pass durations, though not in other measures.

fewer regressions on neutral comma items. This remains consistent with the interpretation of the comprehension results. In errant cases where the comma failed to block the misinterpretation that Rebecca woke up the neighbour, the neutral context in the third sentence would deactivate this misinterpretation – leaving less need to regress out of *Rebecca was fully awake now* to attain comprehension. If this misinterpretation is instead *perpetuated* by the biased content, a regression may be required to extinguish the misinterpretation completely. This interpretation of the data is supported by the trend towards a Structure*Bias interaction in total reading durations on Region 1, suggesting that participants returned there to re-read biased comma items.

In conclusion, Experiment 4 found some evidence that the extent to which misinterpretations linger can be manipulated by a biasing intervening context. There was however little evidence of lingering eye movements (despite this being seen in Slattery et al., 2013), or of the bias manipulation affecting eye movements. Why are these effects not seen? One factor stressed so far is the relevance of asking questions to tap comprehension after each trial. It is possible that asking questions affected participants' reading (and hence eye movement) strategies: for example, it may have obscured effects in this experiment if participants re-read for longer for confirmation, despite having a reasonably intact interpretation of the passage (cf. Christianson et al., 2016). Experiment 5 therefore set out to explore how the addition or removal of comprehension questions affected eye movements.

Experiment 5

3.6. Introduction

It is clear from the experiments presented so far, and those in the wider literature (Christanson et al., 2001, 2006; Patson et al., 2009; Slattery et al., 2013; van Gompel et al., 2006), that initial misinterpretations can linger and disrupt comprehension. However, two things are less clear. First, *why* do misinterpretations linger? Slattery et al. (2013) discussed two possible explanations for lingering: incomplete syntactic pruning and memory fallibility. The former is a more specific, syntactic explanation, consistent with the good-enough approach to sentence processing; the latter is more domain-general. Investigating these explanations ideally requires both an assessment of real-time processing (ideally, via eye tracking) *and* an assessment of readers' eventual comprehension; however, few experiments that investigate syntactic ambiguity resolution have actually asked comprehension questions that specifically tap the ambiguity (but cf. Christianson et al., 2001, 2006, and Experiments 1 to 4 here). Some studies have tapped comprehension in alternative ways (for example by using syntactic priming or paraphrase production; cf. Patson et al., 2009; van Gompel et al., 2006), which may overcome issues with asking predictable questions.

In common across these studies is that comprehension is tested immediately after presentation of the text. This leads to the second question: how long do misinterpretations linger for? While these methods provide clean analyses of comprehension, it is difficult to determine whether initial misinterpretations linger beyond the short-term. Understanding this would help to distinguish competing explanations of Slattery et al.'s (2013) observation of lingering effects in the eye movement record. Experiment 5 set out to

resolve this by exploring interpretations of garden-path texts after a delay, assessing whether misinterpretations linger over a longer period of time.

By assessing comprehension after a delay rather than asking questions after each item, Experiment 5 also provided an opportunity to explore how task demands influence eye movements during syntactic ambiguity resolution. In this experiment, participants were not having their comprehension systematically and predictably tested; in fact, they were not explicitly informed during reading that their comprehension would be tested at all. If good-enough processing means that we only process text to an extent that is necessary for the current situation (Ferreira & Patson, 2007; Karimi & Ferreira, 2016; see also van den Broek et al., 2001 for a similar view), reading should be different in pattern when expecting questions, compared to not expecting questions. This would be especially true when the questions are predictable throughout the task, as in the experiments presented so far, and in Christianson et al. (2001, 2006). As discussed in Chapters 1 and 2, research on how task demands influence eye movements is fairly limited. However, several studies have explored reading under different task demands or task instructions. These found that task factors influence reading times and eye movement behaviour, including factors such as: the task instructions (whether to read for proofreading, for comprehension, or simply to read; e.g., Kaakinen & Hyönä, 2010; Kaakinen et al., 2015; Schmalhofer & Glavanov, 1986; Schotter et al., 2014; van den Broek et al., 2001); payoffs for accuracy and/or speed (Lewis et al., 2013); the presence or absence of questions, and the type of questions being asked (Kaakinen et al., 2015; McConkie et al., 1973; Radach et al., 2008; Swets et al., 2008; Wotschack & Kliegl, 2013); and features of the presented text, such as whether they are in a single sentence or a passage or paragraph (Kuperman et al., 2013; Radach et al., 2008; Whitford & Titone, 2014; Wochna & Juhasz,

2013; and in Chapter 2 of this thesis), or whether the text is task-relevant or task-irrelevant (Kaakinen & Hyönä, 2005).

This experiment therefore had two main aims. The first was to assess representations of syntactically ambiguous texts after a delay. Young adults are good at reinstantiating the exact structure of a sentence when presented with a post-item question (Christianson et al., 2006). Participants are unlikely to be able to recall items verbatim after a 10 minute delay, instead relying on a less detailed gist of the meaning of each sentence (cf. Fletcher & Chrysler, 1990; Schmalhofer & Glavanov, 1986[23]). According to the construction-integration model of comprehension (Butcher & Kintsch, 2012; Kintsch, 1998), this would be a shift from the representations of the exact words at the *surface* level to the less veridical representations of propositional content at the *textbase* level. At this stage, participants will report recognition of paraphrases of a previously-seen text (Fletcher & Chrysler, 1990), as a paraphrase will contain the same propositional content even if it differs in surface structure.

In this experiment, recollection of texts was tested by presenting participants with a series of sentences purporting to be paraphrases of items they had previously seen. Half of the paraphrases were accurate, while half tapped the garden-path misinterpretation. If syntactic ambiguities linger due to insufficient pruning, participants should tend to incorrectly accept paraphrases that tap initial misinterpretations of garden-path items (for instance, *Emma was washing her baby*). I therefore predicted a poorer ability to discriminate paraphrases for garden-path items than comma items, consistent with a good-enough approach. (This prediction is of course also likely to be consistent with other accounts, most notably the noisy-channel approach, cf. Gibson et al., 2013; Levy et

[23] Christianson et al. (2006) reported good recall (c. 80%) on a 2-alternative forced-choice sentence memory task (e.g., if the sentence had said …*the house was small and blue*…, the foil was: *the house was small and white*…). This was based on unambiguous filler items though, so recall should have been easier than the task here.

al., 2009; this is discussed further in Chapter 6). This finding would also question an account based solely on memory fallibility: if this were the case, there should be significant interference to the verbatim representations of *all* items, and there would be no reason to predict a difference based on whether the item contained a syntactic ambiguity. Accordingly, there is no reason to expect effects to be specific to garden-path items. Linked to this aim, there was also the opportunity to investigate links between online and offline performance by seeing whether certain eye movement patterns were associated with improved accuracy on the paraphrase verification test; that said, no strong links were predicted, given the delay between tasks and the scarcity of strong links evidenced in the literature.

The second aim of Experiment 5 was to explore how task demands affect eye movements. By using identical stimuli to Experiments 2 and 3, this allowed a direct comparison between the eye movements of participants experiencing questions, and participants here who were not. I predicted that reading in this experiment would be more superficial, with shorter reading durations and fewer regressions. It was less clear whether this would interact with Structure (as in previous experiments, this term refers to whether a comma was present or absent). On a good-enough account, there should be a reduced garden-path effect on re-reading measures in this experiment: participants will be less inclined to engage in costly re-reading, as the demands of the task do not require this. This should therefore mean a smaller effect in go-past and total reading durations (recall that there was no first-pass garden-path effect in Experiments 2 and 3, and this was expected to be replicated).

Finally I explored individual differences in working memory, using a reading span task (Daneman & Carpenter, 1980). A weak link, if any, was expected between working memory span and reading durations. Traxler et al. (2012) found that a link between

working memory and eye movement measures (cf. Traxler et al., 2005) disappeared when overall reading speed was added to the model, with significant shared variance between these two predictors. A stronger link was predicted between memory span and accuracy on the paraphrase verification task (cf. Daneman & Merikle, 1996 for correlations between reading span and reading comprehension). Participants with high working memory spans have been shown to be more effective at learning information from text (e.g., Kaakinen & Hyönä, 2007; Kaakinen, Hyönä, & Keenan, 2003; von der Malsburg & Vasishth, 2013) and at discriminating genuine paraphrases from foils at the textbase level (Radvansky & Copeland, 2004). High-span participants should therefore show better recollection of the propositional content of experimental items, discriminating paraphrases more effectively by accepting genuine paraphrases, and rejecting foils. The reading span task also acted as a delay between the initial reading task and the paraphrase verification test, presumably preventing verbatim recollection of items.

3.7. Method

3.7.1. Participants

Eighty students at the University of Oxford (n = 74) or Oxford Brookes University (n = 6) took part in this experiment (55 female; mean age = 21.4, range = 18 – 30 years), again representing subjects across humanities (n = 19), social sciences (n = 8) and other sciences (psychology, n = 37; other n = 16). All received £5 or course credits as compensation. The sample size was higher than for previous experiments for two reasons. First, I wanted a similar sample size to the total number across Experiments 2 and 3, to explore the second aim of assessing the impact of removing questions. Second, I expected that the effect of Structure on the verification task may not be as strong as on

comprehension questions, and that a larger sample may provide the power necessary to detect an effect if it existed.

3.7.2. Eye-tracking methodology

An Eyelink 1000 eye-tracker (SR Research) recorded right eye movements at 1000 Hz during the reading task. Participants sat 55cm from the screen. Pupil and corneal reflection thresholds were adjusted according to the eye-tracker instructions. The eye-tracker was then calibrated and validated to <.75° using nine fixation crosses. Calibration was repeated during the experiment when this threshold was lost. The task was programmed using Experiment Builder (SR Research).

3.7.3. Design

After calibration, participants completed three tasks. First, they were asked to read passages on a screen. Participants were asked to read the passages carefully but were not explicitly informed about any future tasks relating to the passages. After this, they completed a reading span task, resembling that used by Daneman and Carpenter (1980). This provided a measure of memory span, and induced a gap of approximately 10 minutes between the end of the reading task and the paraphrase verification task. This final task tested recollection of the passages from the initial reading task.

3.7.4. Reading task

The reading task resembled the one used in Experiments 2 and 3, except that no questions were asked. Participants read a 2-sentence passage on the screen, before pressing a button on a controller to advance to the next screen. The stimuli comprised the same 16 experimental items as Experiments 2 and 3, presented in the same way. As a reminder, all used reflexive absolute transitive verbs, and took the form in (3.2):

(3.2) While Emma undressed(,) the child that was small and happy played on the bed.
Emma finished undressing herself and walked towards the wardrobe.

Participants saw eight as garden-path items (without the comma), and eight as comma items (the variable "Structure"); whether participants saw a given item with or without a comma was counterbalanced across participants. Experimental items were interspersed with 32 filler passages that resembled experimental items but without any ambiguity.

3.7.5. Reading span task

This task was based on the reading span task first used in Daneman and Carpenter (1980), although the sentences used here were designed for this experiment. Participants were presented with a sentence on the screen and were asked to read the sentence out loud, and then to recall the final word of that sentence. The sentences were all between 13 and 15 words long. The to-be-recalled words were high-frequency six-letter nouns, and were chosen from the Subtlex-UK corpus (van Heuven, Mandera, Keuleers, & Brysbaert, 2014); frequencies are given in the corpus in Zipf values (where a Zipf value = $\log_{10}$(frequency/billion words); M_{Zipf} = 5.2, range = 5.00 – 5.68). A full list of sentences can be found in Appendix A. Participants were first asked to read two sentences in a row, and were then asked to recall the two words they had seen. They did this three times. They were then asked to read three sets of three sentences, followed by three sets of four sentences, and finally three sets of five sentences. The dependent measure was the total number of words recalled, with a maximum possible score of 42 – scoring in this way, rather than an all-or-nothing score for each set of words, has been demonstrated to be a more valid scoring system (cf. Conway, Kane, Bunting, Hambrick, Wilhelm, & Engle, 2005; Friedman & Miyake, 2005; see Kaakinen & Hyönä, 2007, and Kim & Christianson, 2013 for examples of studies using this scoring method).

3.7.6. Paraphrase verification task

This task was designed to assess recollection of the sentences that had been read in the reading task. Participants saw a series of sentences. They were told that the sentences would either be a paraphrase of one of the passages seen earlier, or an incorrect foil designed to resemble the material from an earlier passage (for instance, they would feature the same names) but with the meaning changed. They were instructed to respond Yes (if it was a correct paraphrase) or No (if it was a foil) using the controller. Participants were advised that they had 10 seconds to respond to each item, and that if they weren't sure, they should make a best guess from what they could remember.

In total, participants saw 32 paraphrase sentences. Sixteen of these referred to the filler passages with half requiring a Yes response, half requiring a No. The other sixteen referred to the eight garden-path experimental items, and the eight comma experimental items. For half of each of these ("Correct Acceptance" paraphrases), the correct response was Yes; for example in (3.3). The other half tapped the temporary ambiguity and required a No response ("Correct Rejection" paraphrases), such as in (3.4):

(3.3) Emma was undressing herself while her baby played.
(3.4) Emma was undressing her baby.

This variable (Correct Acceptance vs. Correct Rejection) is referred to from here as *Expected Response*. Once again, it was counterbalanced across participants whether, for a given experimental item, they saw a Correct Acceptance or a Correct Rejection paraphrase. This counterbalancing was crossed with the counterbalancing of Structure. Put simply, a quarter of participants saw a given item as a garden-path passage and with a Correct Acceptance paraphrase; a quarter as a comma passage with a Correct Acceptance paraphrase, and so on.

3.8. Results and Discussion

Eye-tracking records were analysed using Data Viewer (SR Research). Fixations above 1200ms were deleted; fixations below 40ms were merged with adjacent fixations or if not possible, deleted. Seven participants were removed due to poor calibration, leaving 73 in the final analysis. Data were analysed using linear mixed-effects models (LMEs) in the *lme4* package (Bates et al., 2015) in R, with the same log transformation, model criticism and maximal random effects structure as discussed in Experiment 1. Details of fixed effects are discussed for each section of the results separately, but all models had Structure (garden-path vs. comma) centred and entered as a fixed-effect; any other fixed effects were also centred.

For clarity, results and discussion are presented separately for this experiment's two aims: first, to explore verification task accuracy, and whether verification task performance was linked to eye movements and/or to individual differences in reading span; and second, to look at how removing questions affected eye movements more generally, compared to results from Chapter 2. Both aims are then reviewed in a General Discussion.

3.8.1. Aim 1: Comprehension after a delay

3.8.1.1. Verification task performance. Participants performed well on questions referring to filler items (83% for Correct Acceptance items; 76% for Correct Rejection items), indicating that general recall ability was good. Figure 3.3 shows mean accuracy for experimental paraphrases, split by Structure and by Expected Response (whether the paraphrase was a Correct Acceptance item where a Yes response is correct, or a Correct Rejection item, requiring a No response). Participants achieved comparable levels of

accuracy (approximately 70%) in three of the four conditions. There was however a reduction in performance for Correct Rejection foils that related to garden-path passages, with accuracy falling at about chance level. The specific drop in accuracy for garden-path items requiring *rejection* shows that participants had an increased tendency to answer Yes to these items.

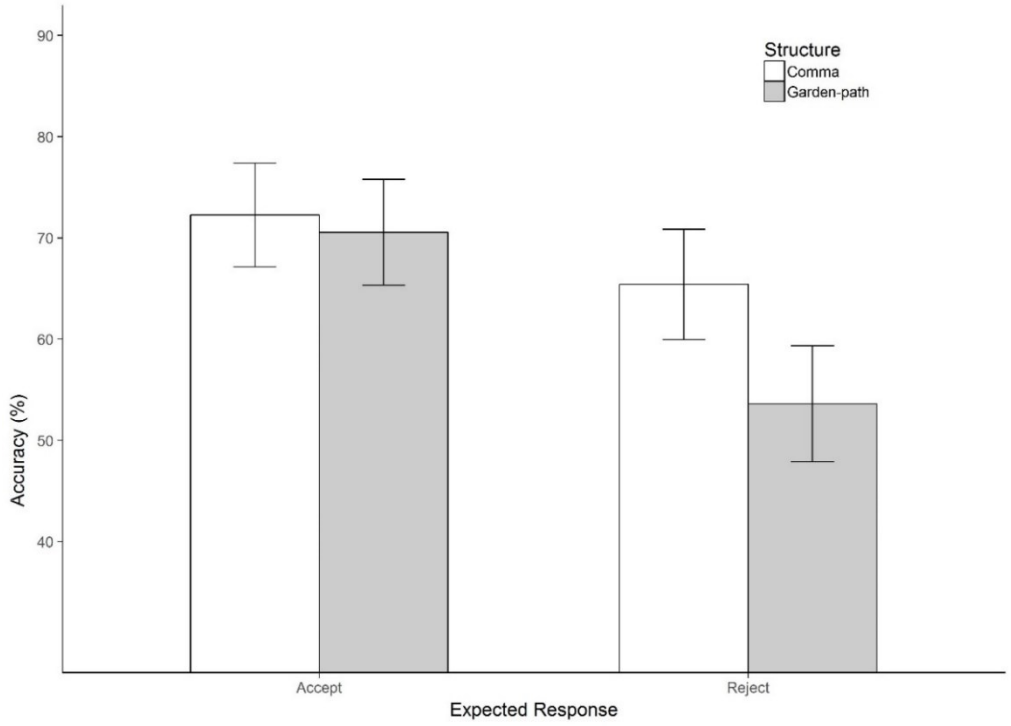

Figure 3.3. Accuracy on verification task experimental items, split by Expected Response (Accept or Reject), and Structure of the experimental item to which the verification item pertained (white bars for comma items). Error bars indicate 95% confidence intervals.

As there are both target paraphrases and foil paraphrases, these data are most appropriately analysed in terms of signal detection theory. In signal detection theory, measures such as d' indicate a participant's ability to distinguish a signal from noise. In this task, participant who answers Yes to all items would achieve 50% accuracy, but would show very poor discrimination between signal and noise. Two participants can have similar overall accuracy scores but different d' scores; consider, for example, Person

A who has 70% accuracy at correctly accepting Yes items, and 70% at correctly rejecting No items, vs. Person B, with 90% accuracy at accepting Yes items, but only 50% at rejecting No items. In this experiment, I was interested in whether participants are more disrupted by noise in the garden-path items, with noise representing their initial misinterpretations. A lower d' score would support this assertion (see Radvansky & Copeland (2004) for use of a similar metric).

D' scores were therefore calculated for garden-path and comma items separately, across all participants and all items. These were calculated as the z-score for the proportion of hits (correctly answering yes to any Correct Acceptance item), minus the z-score for the proportion of false alarms (incorrectly answering Yes to a Correct Rejection item), cf. Macmillan and Creelman (2005). This produced a d' of 0.631 for garden-path items [z(0.705)-z(0.464)] and 0.987 for comma items [z(0.723)-z(0.346)]. The d' is higher for comma items, indicating increased sensitivity to the signal compared to the garden-path items. This reflects the improved accuracy at correctly rejecting foils.

There is no straightforward way to directly compare d' scores while simultaneously accounting for random effects of participants and items: d' would have to be calculated on a by-participant basis, collapsing across items (or by-items collapsed across participants, but this would have limited power with only sixteen items, and would also disregard relevant inter-participant differences such as reading span). Presented below is: first, an analysis of d' scores that does not account for random effects of items, and second, a more indirect analysis that does account for random effects.

To directly compare each participant's d' scores[24] for garden-path and comma items, the simplest solution is a paired-sample *t* test or Wilcoxon test. Visual inspection of

[24] Some participants had hit rates or false alarm rates of 0% or 100%, which would produce infinite *z* scores. These were therefore replaced with 1/16 and 15/16 respectively, with 16 being double the number of comma or garden-path items (cf. Macmillan & Creelman, 2005). Mean d' scores therefore differ slightly from those presented above due to collapsing across items, making these corrections, and recalculating.

d' scores showed them to be negatively skewed; this was confirmed by a significant Shapiro-Wilk normality test on the comma d' scores, $W = .94$, $p < .01$, and a marginally significant test on the garden-path d' scores, $W = .97$, $p = .05$. A pairwise Wilcoxon test was therefore used as it does not assume normality. Participants had significantly better discriminability for comma items [$M_{d'} = 1.11$, $SE_{d'} = 0.15$], compared to garden-path items [$M_{d'} = 0.70$, $SE_{d'} = 0.15$], $V = 1168$, $p = .02$. Notwithstanding the violation of normality, this was checked against a *t* test: this found a marginal effect of Structure, $t(144) = 1.92$, $p = .06$, in support of the Wilcoxon test result.

The analyses presented above suggest that d' scores were significantly higher for comma items than garden-path items; however, these methods do not account for the random effects of items, and may therefore be less reliable. The d' scores can also be analysed in an indirect way by fitting a binomial LME model with fixed effects of Structure and Expected Response, and the dependent variable of the button pressed by participants when answering the question (Yes button or No button). A significant interaction between Structure and Expected Response here would imply that the d' scores were significantly different from each other, as it would indicate that Structure was moderating the extent to which the Expected Response predicted the button actually pressed (i.e., that Structure influenced sensitivity to the correct answer: d'). While not testing d' scores directly, this method is more robust in accounting for random effects of participants and items. The model found a main effect of Expected Response, $z = 4.72$, indicating that participants tended to press the Yes button when the correct answer was Yes, unsurprisingly. The main effect of Structure failed to reach significance, $z = 1.71$, meaning that Structure alone did not create a particular tendency to respond with the Yes button or No button. However, there was a significant interaction, $z = 2.37$, confirming

that participants were more sensitive to distinguishing the truth or falsity of the comma items than the garden-path items.

Finally, for consistency with previous experiments, a binomial LME model was fitted with Accuracy as the dependent variable (i.e., whether the paraphrase was correctly accepted or rejected, as appropriate), and with fixed effects of Structure and Expected Response. There was a main effect of Structure, $z = 2.42$, and of Expected Response, $z = 2.55$. Both were in the expected direction: participants had a better recollection for comma items than garden-path items, and were better at Correct Acceptance items than Correct Rejection items. There was only a trend towards an interaction between Structure and Correct Answer, $z = 1.75$, $p = .08$, suggesting that the analysis using d' was a more sensitive measure. It is worth adding that for the quarter of items that were garden-path Correct Rejection paraphrases (i.e., that required rejection of garden-path foils), performance was at chance on a one-sample *t* test, $t(290) = 1.23$, CI [47.8%, 59.4%]. It was above chance in all other conditions, all $t > 5$.

In sum, the verification test data demonstrated that lingering misinterpretations specifically affect recollection of garden-path items. Participants' d' scores, a test of their ability to discriminate valid paraphrases from foils, were significantly higher for comma items than the syntactically ambiguous garden-path ones. This was demonstrated by the interaction between Expected Response and Structure as predictors of the button pressed by participants, consistent with the less robust (but more intuitive) analysis using a Wilcoxon test. This means that, as predicted, true paraphrases of comma items could be distinguished from false ones with reasonable accuracy; in contrast, participants were willing to accept foil paraphrases that tapped initial misinterpretations of garden-path items.

3.8.1.2. Eye-movement patterns. Eye movements were analysed similarly to Experiments 2 to 4. These were for two different regions, shown in (3.5) in bold.

(3.5) While Emma undressed(,) the child that was small and happy **played *on the*** bed. Emma finished **undressing herself** and walked towards the wardrobe.

Region 1 captured initial recognition of the temporary ambiguity. Similarly to Chapter 2, this was analysed for the disambiguating verb alone (*played*), but also for a longer Region 1 including a spillover region (the word after the disambiguating verb, or if this word was short, the two words; see Chapter 2 for full details). Region 2 was chosen to test for lingering online effects. The definitions of measures (first pass durations, go-past durations, total reading duration, and first pass regressions out) remain as before.

Descriptive data for the one-word Region 1 are presented in Table 3.4. For the measures not described below, linear mixed effects models only found one significant effect of Structure, on regressions into Region 1, $z = 5.17$. The effect on first fixation durations was not significant, $t = 1.62$, and there was no effect on skipping rate.

In Region 1, there was no significant effect of Structure on first pass durations, $t = 1.78$, despite a small numerical difference between comma [$M = 276$ms, $SE = 7$ms] and garden-path [$M = 299$ms, $SE = 8$ms] items. There were effects of Structure on go-past, $t = 2.48$, and total reading durations, $t = 6.95$. The effect of Structure on the proportion of trials with a first-pass regression did not reach significance, $z = 1.51$. As expected, reading durations were longer and there were more regressions for garden-path passages [$M_{GoPast} = 594$ms, $SE_{GoPast} = 31$ms; $M_{Total} = 588$ms, $SE_{Total} = 18$ms; $M_{Regressions} = .31$, $SE_{Regressions} = .02$] than for comma passages [$M_{GoPast} = 475$ms, $SE_{GoPast} = 21$ms; $M_{Total} = 412$ms, $SE_{Total} = 12$ms; $M_{Regressions} = .27$, $SE_{Regressions} = .02$]. These findings replicate observations from Experiments 2 to 4. The results were similar in pattern for the longer

Region 1, except that like in Chapter 2, there were significantly more regressions out of the longer region for garden-path passages, $z = 5.21$. This suggests that regressions were also likely to be made out of the spillover region.

In Region 2, there was no significant effect of Structure on first-pass or go-past durations, $t < 1.4$, but it did significantly affect total reading durations, $t = 2.37$. This result is discussed further in the next section.

Table 3.4.

Additional descriptive statistics for Region 1 eye movements in Experiment 5, by Structure and Verification Accuracy (standard errors in brackets). Measures are defined in Table 1.1.

	Accurate		Inaccurate		*All trials*	
	Comma	G-path	Comma	G-path	*Comma*	*G-path*
First fixation duration	236 (6)	269 (6)	238 (7)	240 (9)	*237 (4)*	*250 (5)*
First pass duration	299 (7)	307 (10)	267 (13)	295 (12)	*276 (7)*	*299 (8)*
Go-past duration	489 (26)	530 (43)	468 (35)	629 (38)	*475 (21)*	*594 (31)*
Total reading duration	428 (15)	608 (22)	404 (21)	576 (32)	*412 (12)*	*588 (18)*
% regressions out	27.4 (2.6)	32.0 (2.8)	27.4 (3.8)	31.2 (3.8)	*27.4 (2.2)*	*31.5 (2.2)*
% regressions in	30.4 (2.8)	53.6 (3.0)	33.6 (4.0)	50.4 (4.0)	*32.6 (2.3)*	*51.6 (2.4)*
% skipping	25.8 (2.2)	30.8 (2.2)	27.4 (3.2)	23.8 (3.1)	*26.9 (1.8)*	*26.4 (1.8)*
% skipping (*longer R1*)	10.4 (1.4)	12.7 (1.5)	8.5 (2.3)	9.4 (2.2)	*9.1 (1.2)*	*10.6 (1.3)*

3.8.1.3. Links between verification task and eye movements. Of interest was whether there was any item-level link between online and offline performance. The eye movement data were therefore split by whether participants ultimately answered correctly on the verification task *for that item* (*Verification Accuracy*, Accurate vs. Inaccurate; note

that “Accurate” could be a correct acceptance *or* a correct rejection, based on the type of paraphrase presented; conversely, “Inaccurate” could mean a false alarm or an incorrect rejection). LME models were re-run, with fixed effects of Structure and Verification Accuracy, and the interaction between these.

The results for Region 1 can be seen in Figure 3.4. As is evident from Figure 3.4 (Panel A), the small garden-path effect on first pass reading durations is limited to items linked to paraphrases that were later answered accurately; other than this, there is little sign of any links to accuracy. LME models supported this, with no significant effects of Verification Accuracy, nor interactions with Structure – leaving little evidence of a link between online and offline measures. (The results for the longer Region 1 were broadly similar in pattern).

Verification Accuracy was added as a fixed effect to models for Region 2, to explore links to offline performance. There were again no significant main effects or interactions for either first-pass or go-past durations, all $t < 1.4$. The main effect of Structure on total reading durations, $t = 2.37$, was qualified by an interaction with Accuracy, $t = 2.01$; the main effect of Accuracy was not significant, $t = 1.74$. The interaction, displayed in Figure 3.4 (Panel D), reflects the fact that a garden-path effect was present only for items that were ultimately answered accurately (be it a correct acceptance or a correct rejection, as appropriate) in the verification task, $t = 3.06$; for items that were ultimately answered inaccurately (a false alarm or an incorrect rejection, as appropriate), there was no effect of Structure, $t < 1$. For items answered accurately, participants showed a small but significant increase in reading durations on Region 2 for garden-path items [$M = 532$ms, $SE = 19$ms] compared to comma items [$M = 472$ms, $SE = 14$ms].

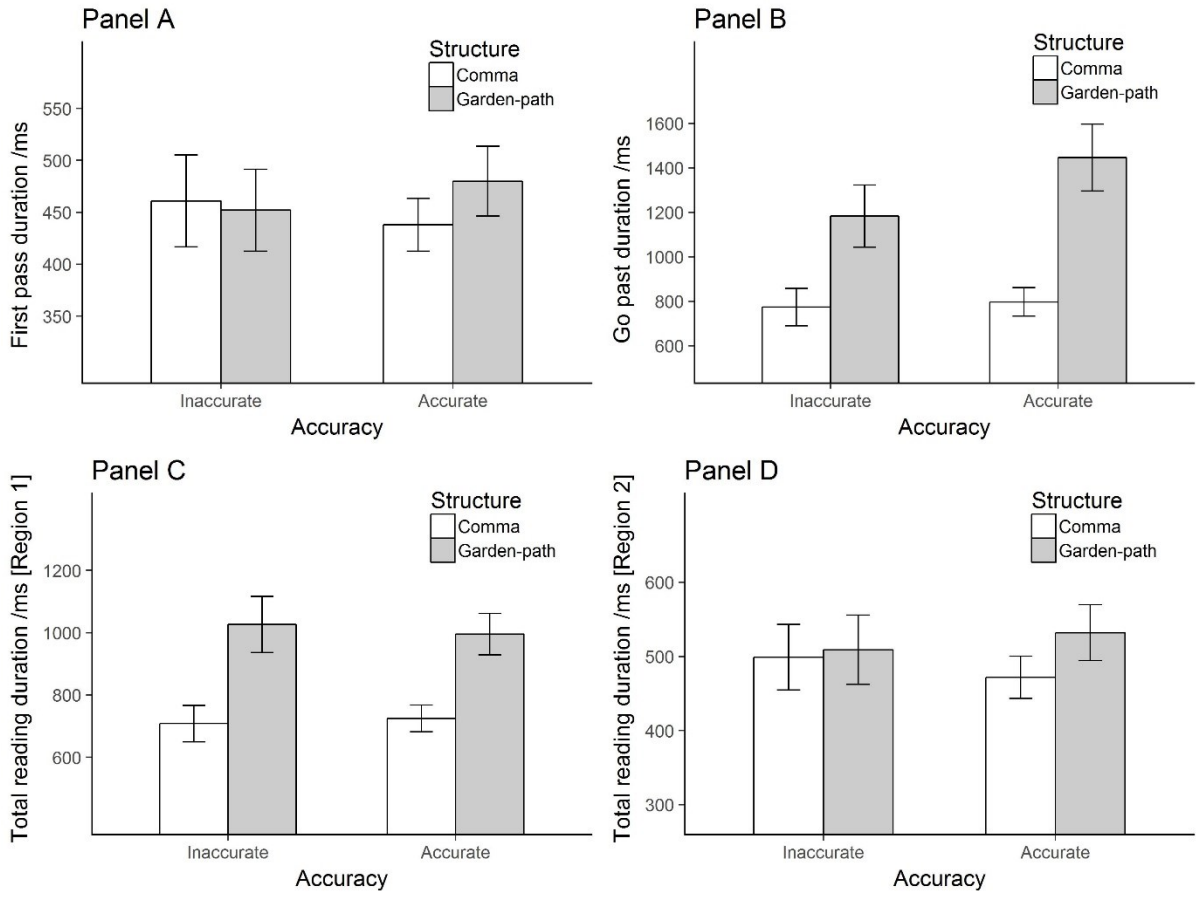

Figure 3.4. First pass (Panel A), go-past (Panel B), and total reading durations (Panel C) for (long) Region 1, and total reading durations for Region 2 (Panel D), by Verification Accuracy (on the verification task for that item; inaccurate on left, accurate on right), and Structure (white bars for comma items). Error bars indicate 95% confidence intervals.

The eye movement results can be summarised as follows. As seen in Experiments 2 to 4, a garden-path effect was observed in go-past durations and total reading durations (but not significantly on first pass durations) on the disambiguating region, Region 1. There was an effect on Region 2: this is similar to the effect that was seen in Slattery et al. (2013) but that has not been seen in the experiments presented in this thesis so far. This was seen in total reading durations, but interestingly, further analyses demonstrated that this was limited to items where participants went on to accurately accept or reject a paraphrase (which one depending on the paraphrase that was presented) relating to that item. This suggests that additional re-reading of this region was linked to better

recollection of the propositional content of these items. Other than this, there were no other item-level links between eye movements and verification task performance.

3.8.1.4. Links to working memory span. I examined whether reading span was linked both to performance on the verification task, and to eye movement measures. Participants generally performed well on the span task ($M = 33.5$, $SD = 4.0$; the maximum score was 42). There were significant Pearson's correlations between a participant's reading span, and both their overall accuracy on the verification task, $r = .24$, $p = .04$, and their d' scores across all items[25], $r = .24$, $p = .04$. Memory span was also added as a covariate to the model for overall verification task accuracy, reported above. This did not affect the pattern of results, but span was a significant covariate, $z = 2.09$.

To explore the link between verification task performance and memory span further, I created a new variable of *span group* by splitting participants into three groups based on their memory span ("low", $n = 27$; "medium", $n = 24$, and "high", $n = 22$), similarly to Kaakinen and Hyönä (2007). D' scores were re-examined by adding span group as a discrete variable to the LME model of d' scores (modelling if participants pressed the left [Yes] or right [No] button, with fixed effects of Structure*Expected Response). Contrasts were set to first compare the low and medium groups, and then to compare both of these together to the high group. Low ($d'_{COM} = 0.78$, $d'_{GP} = 0.49$) and medium ($d'_{COM} = 0.69$, $d'_{GP} = 0.38$) groups were not significantly different from each other, $z < 1$, showing fairly poor discrimination. In contrast, the high group were significantly better for both Structures ($d'_{COM} = 1.70$, $d'_{GP} = 1.12$), $z = 3.43$. There was no

[25] Three notes on these analyses: a) to avoid infinite *z* scores for d', I again replaced hit/false alarm rates of 0% or 100% by $^1/_{32}$ or $^{31}/_{32}$ respectively; b) noting the earlier point about non-normality of data, these results were re-run using Spearman's test, with equivalent results; c) the identical *r* and *p* values are not an error, and reflect the high but trivial correlation between participants' overall accuracy and overall d' scores, $r = .99$.

interaction with Structure, and so there was no evidence that this increased ability was specific to garden-path items.

Correlations were also calculated between each participant's span and that participant's mean log-transformed reading durations on each region, presented in Table 3.5 (the longer Region 1, containing the spillover region, was used). Memory span had a weak but significant correlation with total reading duration on Region 1, but not for first-pass durations[26] (go-past durations was not significant, $p = .11$, although the correlation was also not significantly different from that of total reading durations using a Fisher's r to z transformation, $z < 1$, making it difficult to draw clear conclusions from the difference between these).

Table 3.5.

Correlations between memory spans, eye movement measures and overall verification task performance (proportion accuracy and d').

	Region 1 first pass	Region 1 go-past	Region 1 total durations	Region 2 first pass	Region 2 go-past	Region 2 total durations	Comp. accuracy	Overall d' score
r	.01	.19	.24*	-.23*	.10	<.01	.24*	.24*

* $p < .05$

Memory span was added as a covariate to the eye-movement models reported earlier: it was a significant covariate for total reading duration on Region 1, $t = 2.02$, but not for first pass duration, $t < 1$, or go-past duration, $t = 1.58$. I also looked at how the

[26] First pass durations may not be sensitive on long multiple-word regions, as they sum over several fixations. When tested on first pass durations for just the disambiguating verb, there was a trend towards a negative correlation, $r = -.20$, $p = .09$ – suggesting that high-span readers quickly make a saccade away from the verb.

three span groups compared to one another, with the same contrasts as used to compare d' scores. Again, there was no difference between the low and medium span groups. An interaction between Structure and Span Group, $t = 2.50$, indicated that the high span group showed significantly longer total reading durations on Region 1 on garden-path items specifically (M = 1123ms, SE = 50ms) than either the medium (M = 1016ms, SE = 49ms) or low span (M = 902ms, SE = 40ms) groups; there was no such difference between span groups for comma items.

There were no significant correlations for later reading durations on Region 2. There was a weak but significant *negative* correlation between reading span and first-pass reading durations on Region 2. Further analysis found this to be driven primarily by comma items, indicating less surprise at the reflexive term in Region 2 among higher-span readers.

3.8.1.5. Summary of Aim 1 results: Comprehension after a delay. In sum, there were three main findings. First, participants showed better discriminability for comma items than garden-path items in the paraphrase verification task. Further analysis demonstrated that participants performed at chance when asked to reject foil garden-path paraphrases, and were therefore willing to accept foils based on their initial misinterpretations as if they were accurate. In the eye movement record, the results were broadly similar in pattern to those seen in Experiments 2 to 4. It was interesting that garden-path effects were still seen despite the absence of questions, although the size of these effects is discussed further in the next section. There was the first sign in this thesis of a garden-path effect in the second sentence, as reported by Slattery et al. (2013). This was only clear in items that were subsequently answered correctly, but nevertheless, the parallel with Slattery et al. (who also did not ask questions tapping the misinterpretation) is of interest. Finally, correlations with reading span demonstrated that individual

differences in working memory moderated performance on the reading task. High-span participants read garden-path sentences for longer, and in turn, were better at discriminating target paraphrases from foils in the verification task (cf. von der Malsburg & Vasishth, 2013).

3.8.2. Aim 2: How does the presence/absence of questions affect eye movements?

The stimuli used in this experiment were the same as in Experiments 2 and 3. This allowed for a direct comparison of eye movement patterns based on whether the task included questions after each item or not. Data from this experiment were therefore combined with data from Experiments 2 and 3. Models were run on reading durations with fixed effects of Structure (as before) and Task (*questions* vs. *no questions*). To partial out effects of the concurrent task in Experiments 2 and 3, I also included Load as a covariate[27] with three levels (*no-load*, for no-load participants from Experiment 2 and all participants here; *2-back* for 2-back participants from Experiment 2; and *4-back* for participants from Experiment 3).

The first analyses were for reading durations on Region 1. As there were few differences between the shorter and longer versions of Region 1, these were done on the longer Region 1, including the spillover region. The data, split by Structure and Task, are displayed on Figure 3.5. Figure 3.5 demonstrates that other than in first pass durations, no-questions participants show shorter reading durations. However there did not seem to be any great difference in the size of garden-path effects for the two groups. In line with these findings, there was a main effect of Task on go-past durations, $t = 2.06$, total

[27] Load was added as a covariate and not a fixed effect because it was only used in the *questions* participants, and so is not fully crossed with Task (leaving empty cells if an interaction term was added).

reading durations, $t = 3.64$, and the proportion of trials where a first-pass regression was made, $t = 3.10$, but not on first-pass durations, $t = 1.37$. Reading durations were longer and there were more regressions for participants who were asked questions. There were however no significant interactions between Structure and Task, all $t < 1.4$. For Region 2 eye movement measures, there was a main effect of Task on both first-pass, $t = 2.01$, and total reading durations, $t = 3.25$; the effect on go-past durations did not reach significance, $t = 1.69$. There were again no significant interactions.

One consideration is that eye-movement data from Experiment 3 may be anomalous, due to the low accuracy of participants who had a concurrent 4-back task. The above analyses were repeated with the data from Experiment 5 compared only to the data from Experiment 2 (and not Experiment 3). The pattern of results was broadly similar with two exceptions. First, the interaction between Structure and Task for Region 1 total reading durations approached significance, $t = 1.73$. Second, the main effect of Task on Region 2 go-past durations did reach significance, $t = 2.04$. While these results provide more support for the predictions, they should be treated with caution, to account for both the number of analyses run here (risking a Type I error), and their marginal nature.

In conclusion, as predicted, participants who were not expecting targeted comprehension questions after each item had shorter reading durations and fewer regressions, compared to participants from Experiments 2 and 3 who did expect these questions. There were no clear interactions with Structure, indicating a general effect rather than one that was specific to syntactically complex sentences.

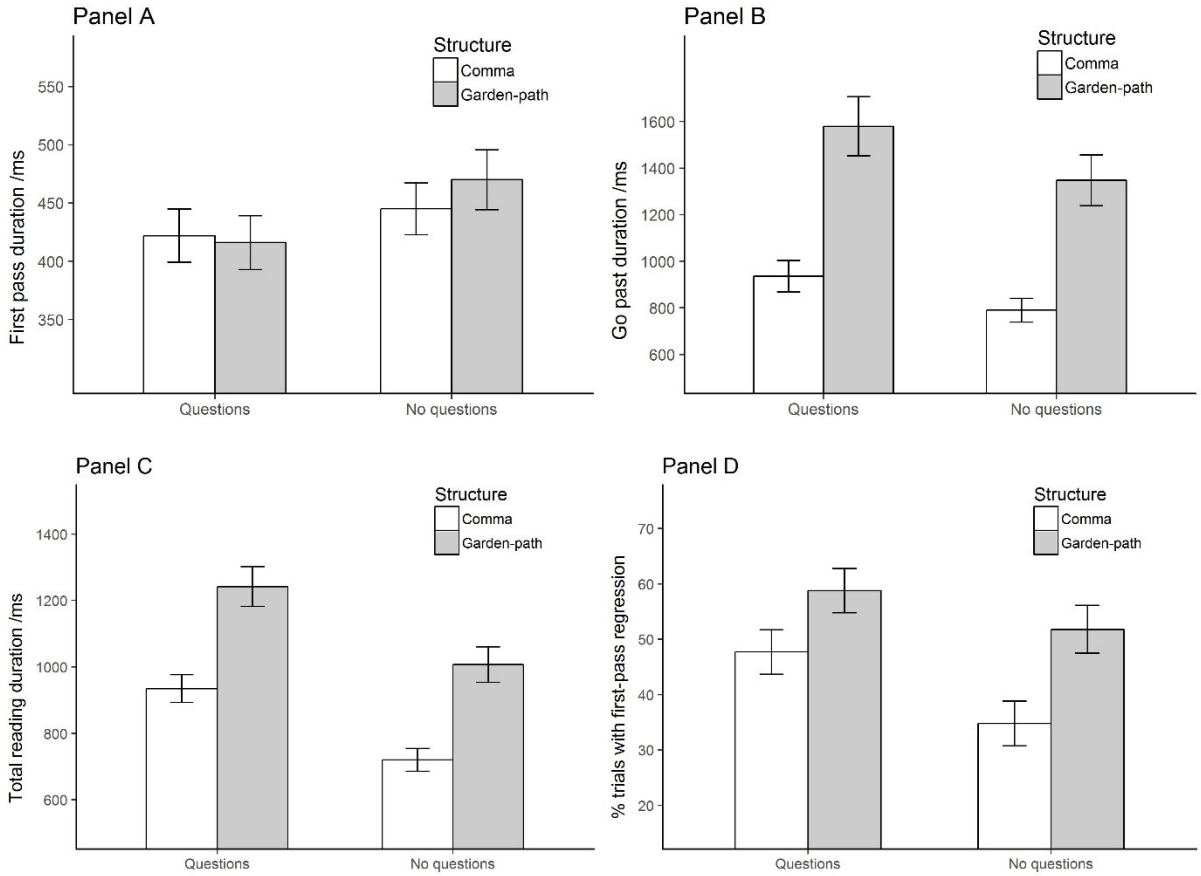

Figure 3.5. First pass (Panel A), go-past (Panel B), total reading durations (Panel C), and the proportion of trials with a first-pass regression (Panel D) for **longer** Region 1, split by whether participants saw questions (on left; from Experiments 2 and 3) or not (on right), and Structure (white bars for comma items). Error bars indicate 95% confidence intervals.

3.9. General Discussion of Experiment 5

The two original aims of this experiment were to explore memory for syntactically ambiguous texts following a delay, and to see whether removing comprehension questions affected eye movement behaviour while reading. On the first aim, participants showed reasonable recollection of comma items, discriminating effectively between incorrect and correct paraphrases. With garden-path items, however, participants were less effective at distinguishing between correct and incorrect paraphrases. When required to reject incorrect paraphrases that tapped the garden-path misinterpretation, performance

was at chance. Interestingly, participants with high memory spans were better at discriminating the paraphrases. On the second aim, there were still similar garden-path effects in go-past and total reading durations (although like in Experiments 2-4 when using stimuli in passages, not in first-pass durations). However, participants not expecting questions showed shorter reading durations across stimuli, offering evidence of task demands affecting reading behaviour.

In the earlier experiments, it was noted that participants were high-performing, being university students with comparatively high comprehension of syntactically complex texts (compared to, say, Christianson et al., 2001, 2006). The memory span results here support this conclusion: participants performed very well on the task, averaging 33 out of 42. This suggests that most participants achieved a span of 4 at least once; indeed, many participants achieved a span of 5 on at least one occasion. In comparison, the "medium span" group (i.e., middle of three groups) in Kaakinen and Hyönä (2007) only had a mean score of 22.82 (using a similar method and scoring system to that used here) and often failed to achieve a span of 4. Similarly, Friedman and Miyake (2005) gave participants double the number of trials as used here and reported a mean score of 52/84, equivalent to 26 out of 42 using the methodology here (the *maximum* score from any participant on Friedman and Miyake's first session was 74/84, equivalent to 37 out of 42 here). This may question the representativeness of the sample used in this experiment. It may also mean that the measure lacked sensitivity for this group (instead favouring the use of 6-word or even 7-word spans, cf. Kaakinen & Hyönä, 2007). Notwithstanding these concerns, performance on the verification task (for experimental items) was not at ceiling, and no participant achieved a span of 5 on all three trials of the span task. Furthermore, using a lower-performing sample may have led to floor effects and made it more difficult to ascertain differences between performance on comma and

garden-path items. Finally, this point does not negate a comparison *between* participants in this experiment, even if results cannot necessarily be compared to other studies.

Looking first at verification task performance, the results demonstrated that initial misinterpretations of syntactically ambiguous texts lingered beyond the immediate post-item comprehension test used in Experiments 1-4. While previous research has identified lingering of misinterpretations in eye movements on a second sentence (Slattery et al., 2013; but see also Experiments 2 and 3 of this thesis), and of course lingering in responses to immediate comprehension questions (Christianson et al., 2001, 2006; and Experiments 1 to 3 here), Experiment 5 finds lingering effects after a lengthy delay. Christianson et al. (2006) suggested that younger participants are able to reinstantiate representations of the verbatim content of sentences, even if this process is not always without error, especially for garden-path sentences. However, this related to questions asked immediately after sentence presentation. It is highly unlikely that participants were relying on verbatim representations after such a delay in this experiment (due to the time between presentation and recall, and due to the intervening span task). Participants must instead have been working from less detailed representations of the propositions represented by the sentences (Schmalhofer & Glavanov, 1986; van Dijk & Kintsch, 1983). The results from this experiment demonstrate that these representations contain traces of garden-path misinterpretations, consistent with the good-enough framework.

The results here also gave an insight into how individual differences affect participants' ability to recall the content of text faithfully. Performance on the reading span task predicted accuracy on the verification task: high-scoring participants showed better abilities to discriminate between true and false paraphrases. They also spent longer reading the disambiguating part of garden-path items, consistent with previous research (von der Malsburg & Vasishth, 2013). This supports the view that working memory is

associated with accurate retrieval of information during sentence processing, especially where processing load is high (cf. Caplan & Waters, 2013; Christianson et al., 2006; Daneman & Merikle, 1996; King & Just, 1991; Nicenboim et al., 2016; von der Malsburg & Vasishth, 2013). The link between reading span and eye movement measures was however weaker, providing some support for a distinction between memory processes underpinning online syntactic processing, and those linked to offline text recollection and comprehension (cf. Caplan & Waters, 2013; Traxler et al., 2012).

The eye movement patterns from this experiment were broadly of a similar pattern to the results presented in Chapter 2, and in previous research (e.g., Slattery et al., 2013). Replicating Experiments 2 and 3, there was no first-pass garden-path effect on the disambiguating Region 1 for sentences presented as the first sentence in a passage (this was also seen in Experiment 4, where the garden-path sentence was the second sentence in a passage). As discussed in Chapter 2, this suggests that participants prefer to make a regressive or progressive saccade to gather more information, rather than spending additional time reading the disambiguating region as happens in single sentences. The replication of this finding strengthens the comment that future research should consider this critical difference when designing stimuli. Nevertheless, garden-path effects were again found in go-past and total reading durations, consistent with participants finding these sentences more difficult to parse. Interestingly, there was also evidence of a garden-path effect in the reflexive term of Region 2, similar to Slattery et al. (2013), but limited here to items where participants ultimately answered the paraphrase correctly. No effects were seen in Experiments 2 and 3 (or, using slightly different stimuli, in Experiment 4). The finding of an effect in Experiment 5 offered a tentative conclusion that the Region 2 effect may be affected by task demands (neither Slattery et al. nor this experiment contained comprehension questions that specifically tap the misinterpretations, but

Experiments 2 and 3 did), and by attention to the task; however, further experiments would be required to explore this.

This conclusion has some support from the more general task effects seen when comparing eye movement patterns in this study to those seen in Experiments 2 and 3, where participants received questions after each item. The overall pattern was for shorter reading durations and fewer regressions in participants in Experiment 5 (who faced no questions), consistent with a more superficial reading style. This was not limited though to syntactically ambiguous texts: there was no interaction between Structure and Task for any measure. Participants in Experiments 2 and 3 spent considerable time re-reading if they expected to be asked comprehension questions about the text being read. Participants in Experiment 5, in contrast, may have had some idea that their comprehension would be tested (they were advised that this was a study about reading and memory, and that there would be additional tasks), but had no explicit expectation of their comprehension being tested, and in particular had no reason to expect what any later test would be tapping. Accordingly, they terminated re-reading more quickly. This is consistent with a good-enough approach positing that we process texts only to the extent necessary for the current task demands (Ferreira & Patson, 2007). Furthermore, in Experiments 2 and 3, and also in studies such as Christianson et al. (2001, 2006), participants learned fairly quickly what the question will be: questions all followed the same *Did X y the Z?* pattern. Accordingly, they may re-read both comma and garden-path sentences for longer in order to achieve the goal of answering these specific questions correctly. This is consistent with Christianson et al.'s (2016) evidence that re-reading and regressions are more for *confirmation* of what has been read, rather than always being to *revise* initial misinterpretations. If the presence of questions encourages re-reading simply to confirm what has been read the first time, there is no reason to expect any clear difference

between comma sentences and garden-path sentences (as the desire for accuracy will be equally true for all sentences). This would explain the absence of a Structure-by-Task interaction in the comparison of Experiments 2 and 3 to this experiment.

The results of this experiment highlight the importance of task demands and instructions in psycholinguistic research. This is an understudied area (see Hyönä & Kaakinen, 2011; Radach et al., 2008). Further comparisons between different task demands on different sentence structures may elucidate the precise factors that influence decisions about whether to re-read, and if so, how long to do this for. This should lead into incorporating task-dependence in models of eye movements, and of reading generally (see Logačev & Vasishth, 2016a, for one such implementation). Greater recognition of task demands seems a critical part of understanding what it means for reading to be "good enough".

The paraphrase verification data also offer a paradigm for assessing participants' comprehension of syntactically ambiguous texts following a delay. While paraphrases have been used previously (e.g., Kim & Christianson, 2013 in a similar verification task; Patson et al., 2009 in a production task), these have usually been presented immediately after each item is presented. Other studies (e.g., Fletcher & Chrysler, 1990; Schmalhofer & Glavanov, 1986) have looked at whether participants accept paraphrases after a delay, but these have not been specifically testing syntactic ambiguity. Experiment 5 found that syntactic ambiguities can linger, and that representations of those ambiguities can be tapped after a delay using this paradigm. This supports Patson et al. (2009) who asserted that Christianson et al.'s (2001) original findings were not simply an artefact of their research design (namely, asking questions after each item that force reanalysis): good-enough effects can be seen across offline tasks (in Patson et al.'s case, in paraphrase production). The paradigm used here also offers potential to explore just how long

misinterpretations linger for, by extending or reducing the period of the delay between items and testing. It is unknown whether the reading span task interfered with participants' representations of the items they had read. The sentences that were read were unrelated to the experimental items, but it was nevertheless a reading task. Substituting this for other tasks, possibly non-linguistic ones, would assess whether these results are a product of the reading span task acting as an interference, as well as a delay. However, the relatively high performance on paraphrase verification for fillers and comma items supports the conclusion that this tendency to answer Yes on garden-path foil paraphrases (which should be rejected) is specific to lingering misinterpretations of temporarily ambiguous text.

In summary, this experiment found additional support for the good-enough framework, demonstrating that initial misinterpretations of syntactically ambiguous sentences are held on to for even longer than has been seen previously. Even after a 10 minute delay, participants showed improved sensitivity to comma items over garden-path items, consistent with an incomplete syntactic analysis interfering with comprehension. Furthermore, altering the task demands by removing questions produces more superficial eye movements, again consistent with an account of participants only re-reading for longer if this is necessary given their current demands.

3.10. Conclusion

In the introduction to this chapter, three main questions were posed. The first was whether the two-sentence passages used in Experiments 2 and 3 were insufficiently complex, and whether stronger garden-path effects may be visible in longer 4-sentence passages. Experiment 4 offered little evidence to support this, and eye movement patterns

were broadly similar to those presented in Experiments 2 and 3. Comprehension was comparable to the results of reading two-sentence passages in Experiment 2, but was again down on the single-sentence condition of Experiment 1. There was though a notable decline in comprehension accuracy when the intervening sentence perpetuated the misinterpretation, suggesting that reactivating the initial misinterpretation may strengthen its interference.

This links to the second question of how long lingering would be seen for. Experiment 4 demonstrated that misinterpretations had a greater effect on comprehension question accuracy if additional material pointed towards them. In Experiment 5, it became clear that these misinterpretations lingered in participants' representations for even longer. After a 10 minute delay and the completion of the intervening reading span task, participants were significantly poorer at discriminating between paraphrases of garden-path items than of comma items. Instead, participants were willing to accept paraphrases that tapped the initial misinterpretation as a veridical gist of what they had read earlier. This is consistent with Slattery et al's (2013) explanation that initial misinterpretations are insufficiently pruned, and linger to interfere with later processing. Here, it is clear that this lingering can be seen for substantially longer than seen before.

The final question was how task demands would affect eye movements. In a comparison to data presented in Chapter 2, Experiment 5 found that when participants were not expecting to be asked specific questions, they spent less time reading critical regions of passages. This was however a general effect and was not limited to syntactically ambiguous sentences, suggesting an overall strategy shift towards less detailed sentence processing. What was particularly interesting was how individual differences in memory span affected both online processing and later comprehension: high-span readers read sentences for longer, and were subsequently better at

discriminating the garden-path paraphrases. This suggests an important role for these individual differences in moderating good-enough processing. That said, the participants in these experiments were still from a fairly homogeneous pool – all were young university students, with reasonably high working memory spans (even in those deemed "low"). An interesting question is how the patterns of eye movements and comprehension seen so far in this thesis would translate into a different population. Chapter 4 therefore moves to look at syntactic ambiguity processing in older adults.

Chapter 4

Good-enough effects in older adults

In Experiment 5, differences in working memory were associated with different patterns of eye movements and comprehension of syntactically ambiguous sentences (see also Christianson et al., 2006; von der Malsburg & Vasishth, 2013). This association, and especially how comprehension relates to individual differences in working memory, has also been investigated by considering differences associated with ageing. Reported in this chapter, Experiment 6 investigated reading in older adults in order to explore further individual differences in both eye movements and comprehension during syntactic processing, and in particular, to see how older adults would be affected by the dual-task paradigm introduced in Experiment 1.

4.1. Effects of ageing on reading and cognition

4.1.1. Reading in older adults

Older readers (usually defined as those aged 65 and above) show a different pattern of eye movements to younger readers (e.g., Kemper et al., 2004; Kemper & Liu, 2007; Kemper & McDowd, 2006; Kliegl et al., 2004; Payne et al., 2014; Rayner et al., 2006, 2009; Stine-Morrow et al., 2006a, 2010; von der Malsburg et al., 2015). To briefly recap research discussed in Chapter 1, studies have commonly found a general slowing of reading, due to both more fixations and longer fixation durations. Older adults also show greater "wrap-up" effects – pausing at syntactic boundaries (Payne & Stine-Morrow, 2012; Stine-Morrow et al., 2010). These effects are complemented by a "riskier" (Rayner et al., 2006, 2009) pattern of eye movements: more initial skipping of words, but for difficult sentences, more regressions and more time spent re-reading to compensate for

the earlier skipping. Rayner et al.'s original discussion of risky strategies was specifically about (unconscious) strategies arising from low-level lexical processing differences. The idea that older adults' reading patterns are more superficial has however been extended to higher-level processing (DeDe, 2014, 2015; von der Malsburg et al., 2015). For example, von der Malsburg et al. (2015) found that scanpaths in older adults were less consistent (i.e., there was more variability between trials), but were also less affected by the syntactic complexity[28] of a sentence than scanpaths of younger adults. Their eye movements are therefore riskier in a slightly different meaning from Rayner et al.'s lexical processing: older adults make greater use of discourse information and wider knowledge, but less use of explicit syntactic structure.

Furthermore, DeDe (2014) found that older adults were more likely to use lexical cues to predict upcoming syntactic structure, but were also more disrupted when these cues produced an incorrect interpretation. For example, take sentences (4.1) and (4.2):

(4.1) While the waiter served the woman *the food* ***was still*** too hot.
(4.2) While the waiter kissed the woman *the food* ***was still*** too hot.

Younger adults showed more disruption at *the food* in (4.2) than (4.1), suggesting that they had difficulty in reinterpreting *the food* as the start of a new clause. In contrast, older adults showed no such disruption. DeDe concluded that older adults can efficiently retrieve the properties of *kissed*, namely that *kissed* can only accept one noun as an object; in contrast, *served* is a ditransitive verb that can take two nouns (e.g., *the waiter served the woman the food*). Younger adults do go on to show disruption at *was still* in (4.1), having interpreted *served* as ditransitive, and the sentence as meaning that the waiter was

[28] In von der Malsburg et al., syntactic complexity was indexed by: a) surprisal (the probability of seeing a new word, or that word's part of speech, given the existing structure e.g., of seeing a verb like *played* in *While the father changed the baby played with its toys*), and b) retrieval costs (how difficult it is to retrieve earlier material so the new word can be incorporated into the existing structure); see Boston, Hale, Kliegl, Patil, and Vasishth (2008); Boston, Hale, Vasishth, and Kliegl (2011) for details.

serving *the food* to the woman. Older adults however showed even greater disruption at *was still* in (4.1). This suggests that they use lexical and syntactic cues to make predictions (see DeDe, 2015, for evidence that older adults also make more use of one such cue, animacy), but are more disrupted if their predictions are incorrect, consistent with a higher-level "risky" strategy.

For simple sentences, the approach to reading taken by older readers is generally successful (cf. Bicknell & Levy, 2010, for evidence that risky strategies are generally beneficial). For more complex sentences, such as those containing a syntactic ambiguity, a need for additional reading time and more regressions makes reprocessing costly. A similar need for increased re-reading was seen in Chapter 2 when younger adults read complex sentences while completing a concurrent task. Unlike the pattern of results reported in Chapter 2, however, older readers tend to show poorer comprehension – especially when answering questions related to syntactically complex or ambiguous sentences. This has been observed in reduced accuracy for temporarily ambiguous reduced relative clauses (Kemtes & Kemper, 1997), "dense" sentences that contain multiple concepts (Stine-Morrow et al., 2006a), and of most relevant here, temporarily ambiguous early-closure sentences of the type used in this thesis (Christianson et al., 2006).

Christianson et al. (2006) found that older adults showed decreased accuracy compared to younger adults on optionally-transitive (OPT) items, such as (4.3) and (4.3a), although not on reflexive absolute transitive (RAT) items, such as (4.4)/(4.4a).

(4.3) While the man hunted(,) the deer ran into the woods.
(4.3a) Did the man hunt the deer?
(4.4) While the father changed(,) the baby played with its toys.
(4.4a) Did the father change the baby?

The measure of comprehension was the proportion of "Yes" responses to questions (4.3a) and (4.4a), since answering Yes is not explicitly licensed by the sentence content. There is however a difference between the two question types. In (4.4a), a Yes answer is definitively incorrect: if the sentence has been correctly parsed, it is clear that the father is getting changed himself and not changing the baby. In contrast, the fact that the man is "hunting" in (4.3) (without a direct object) and the deer running into the woods, does not preclude that the man is hunting the deer. Christianson et al. argued that older readers are more likely to answer Yes to (4.3a) because of a reliance on an incomplete representation acquired during sentence processing, and an inability to recall the sentence structure verbatim upon reading the question.

This view is supported by the fact that older adults perform worse on recollection of surface (verbatim) and textbase (gist) representations of text[29] (Radvansky et al., 2001, 2003; Shake et al., 2009). In contrast, younger adults (with better overall working memory) are able to reinstantiate the sentence structure more effectively when presented with the question. This explanation is consistent with the idea that older adults process text less deeply (Craik & Simon, 1980), and demonstrate deficits for memory of texts accordingly (cf. Johnson, 2003; Radvansky et al., 2001, 2003; Verhaeghen, Marcoen, & Goossens, 1993). Older readers use heuristics (in (4.3), that *hunting the deer* is compatible with *hunted [SOMETHING]*) to choose the most likely answer. In contrast, younger readers are less likely to answer Yes, following what is explicitly stated in the sentence; in (4.3), it is not stated that the man is hunting the deer). This heuristic is not necessarily linked to older readers' increased world knowledge and experience:

[29] Radvansky et al. (2001) argued that older adults may use textbase representations as a scaffold to achieve a consistent, situation model-level comprehension of the text. The textbase representations, which are more grounded in the text, then rapidly decay. This may reflect older adults' reading strategies, as discussed later.

Christianson et al. (2006) found no evidence that older readers were more likely to use real-world knowledge (via a plausibility manipulation).

4.1.2. Explanations of ageing effects

These age differences in reading have been attributed to a number of factors. To some extent, differences result from biological consequences of ageing. These can be low-level, such as differences in vision and oculomotor processing (Burke & Osborne, 2007; Jordan et al., 2014; McGowan et al., 2015; Mund, Bell, & Buchner, 2010; Owsley, 2011; Paterson et al., 2013; Tun et al., 2009). Vision changes in many ways with age – most importantly in areas such as visual acuity, contrast sensitivity and visual processing speed (for a review, see Owsley, 2011). It has therefore been suggested that older adults adjust their processing strategies due to the added effort required, in particular when task demands are high (DeDe & Flax, 2016; Gao et al., 2012; Tun et al., 2009). This may result in some of the differences in eye movements, and in turn, in comprehension.

Above and beyond perceptual changes, older adults also show declines in working memory (e.g., Bopp & Verhaeghen, 2005; DeDe et al., 2004; Gordon et al., 2016; Stine-Morrow & Miller, 2009; Waters & Caplan, 2003) and in inhibitory abilities (Connelly et al., 1991; Hasher, Zacks, & May, 1999; Zacks & Hasher, 1997; but cf. Kemper & McDowd, 2006; Kemper et al., 2008). Even having accounted for these, older adults also show a general decline in crystallised intelligence (e.g., Miller, Stine-Morrow, Kirkorian, & Conroy, 2004), and in the efficiency of encoding (*processing speed*; cf. Craik, 1983; Malyutina & den Ouden, 2016; Salthouse, 1992, 1996, 2004).

There is substantial evidence to support a decline of working memory with age, but this is underpinned by similarly substantial disagreements in how this decline is theoretically conceptualised. Age-related differences in reading have been linked to

general declines in working memory (e.g., Christianson et al., 2006; DeDe et al., 2004; Kemper & Liu, 2007; Stine-Morrow et al., 2006a; but for an alternative view, cf. Caplan & Waters, 1999; Caplan et al., 2011; Waters & Caplan, 2001, 2003). For example, Kemper and Liu (2007) found that individual differences with age in later eye movement measures (such as total reading duration and the proportion of regressions from critical regions) were mediated by differences in working memory capacity. That said, other research has questioned a link between general working memory resources and online sentence processing, finding instead that working memory capacity mediated age effects only on comprehension (DeDe et al., 2004; see also Caplan & Waters, 1999; Caplan et al., 2011; Waters & Caplan, 2001, 2003). These disagreements may in part depend on differences in what is considered an "online process" (or in some cases, what is deemed "syntactic processing"). Are regressions and total reading durations reflective of online syntactic processing or of post-syntactic revision? Furthermore, are results from an auditory moving-window paradigm (where sentences are presented auditorily, on a participant-controlled word-by-word basis; cf. DeDe et al., 2004) sufficient to rule out a role of working memory on online processing? Nevertheless, there is reasonable evidence of some link between working memory and differences in reading with age.

Another well-researched topic concerns age-related differences in inhibition, most notably work by Hasher, Zacks and colleagues (e.g., Carlson et al., 1995; Hamm & Hasher, 1992; Kim, Hasher, & Zacks, 2007; Zacks & Hasher, 1997). Hasher, Zacks and colleagues argue that ageing deficits in reading are indicative of a broader reduction in the efficiency of inhibition skills, linked to poorer executive control. As such, differences in reading can be attributed to more irrelevant information being maintained and disrupting comprehension. In support of this general idea, older readers find it more difficult to inhibit distractors (Carlson et al., 1995; Connelly et al., 1991; Zacks & Hasher, 1997).

This suggests that the problem with ageing may not be having too little in working memory, but actually having too *much* (Hamm & Hasher, 1992; Healey et al., 2013; Radvansky et al., 2005; Zacks & Hasher, 1997). Seeing working memory in terms of the efficiency of cue-based retrieval, rather than a capacity-based system (cf. Gordon et al., 2012; van Dyke & Johns, 2012) may provide a way of combining these two approaches: poor working memory is akin to reduced ability to inhibit distractors, and hence a difficulty in successfully retrieving earlier chunks of text upon reaching a later cue. This interpretation is consistent with Kemper and McDowd's (2006) study where older and younger adults read sentences containing distractor words. Older adults were slower overall, but showed no reduction in comprehension or other deficits in text processing, suggesting that while they faced more difficulty with identifying distractors and targets in working memory, they had no overall inhibition deficit.

It is also possible that differences in working memory and inhibition may reflect a more domain-general decline in processing speed. This has been shown most effectively in work by Salthouse (e.g., Salthouse, 1996), whose regression models partialled out effects of other cognitive deficits to identify a more general role for processing speed. More broadly, a role for processing speed has been seen in younger adults: Traxler et al. (2012) found that working memory had no predictive value on how long participants spent reading syntactically complex sentences once overall reading speed (an index of processing speed) was added to the model. This may explain individual differences in reading more generally, and potentially, the general decline in older readers (cf. Macdonald & Christiansen, 2002; but see also Just & Varma, 2002). Salthouse (e.g., 1996) has suggested a general decline in processing speed links to a degradation in cognitive abilities such as those engaged during reading and comprehension. In sum, the

decline of domain-general cognitive abilities with age appears to impact on older adults' comprehension and perhaps online sentence processing.

Explanations based on age-related sensory or cognitive deficits suggest an immutable decline in abilities. There may however be explanations that don't appeal to consequences of ageing in a simple sense: older readers, aware of their more limited cognitive resources and slower oculomotor functions, may adopt strategies to counteract these deficits (Salthouse, 1996). For instance, older readers will have acquired more world knowledge and reading experience (Salthouse, 2003; Stine-Morrow et al., 2006b) and vocabulary (Verhaeghen, 2003). They may therefore rely on greater use of predictions, and accordingly read text more superficially on their first pass, explaining the increased skipping rate (Macdonald & Christiansen, 2002; Rayner et al., 2006, 2009; von der Malsburg et al., 2015). Predictions might speed up reading of simple sentences, but may have a counterintuitive effect when faced with syntactically complex sentences: an expectation of one interpretation may make errors hard to overcome. Most sentences will therefore be processed with ease, hence why age effects are often elusive (cf. Salthouse, 2004). The difficulty arises with ambiguous structures – not least because these structures are relative infrequent in everyday language (DeDe, 2014; Salthouse, 2004).

The use of strategies means that under easy conditions, effects of ageing are small if not absent. Under more difficult conditions (e.g., when reading syntactically complex sentences, or when faced with a high cognitive load), these strategies might at least give the impression of an age-related impairment, due to a tendency to re-read. Older adults may prioritise top-down predictions over bottom-up input, or conversely, may allocate additional time for re-reading (or checking generally, in non-reading tasks) to maintain accuracy, while preserving time and resources. Such strategies are usually theorised to be unconscious (see e.g., DeDe, 2015; Rayner et al., 2009), but nevertheless may be the root

of differences in reading patterns when compared to younger adults. Nevertheless, if effects of ageing are strategic, a change in task instruction may produce a rebalancing of strategies and priorities, and in turn, similar results in older adults to younger ones. Consistent with this view, studies that manipulate response deadlines (e.g., Laver, 2000), or encourage participants to prioritise speed (e.g., Mitzner et al., 2010) have found that older adults can perform similarly to younger adults if task demands do not allow additional time for checking initial responses.

One theory that encapsulates these differences in strategy with age (and with task demands) is Stine-Morrow et al.'s self-regulated language processing model (SRLP; e.g., Stine-Morrow et al., 2006b). The SRLP model links reading behaviour to underlying goals, suggesting that resources[30] are allocated in order to best meet those goals. At each level of comprehension (word, textbase and discourse; Butcher & Kintsch, 2012; Kintsch, 1998), text is processed to the extent that is necessary for the given task[31], with negative feedback loops to identify if each standard has been achieved, or if more processing is required. Resources may be allocated more to the textbase and discourse levels if reading for overall meaning, or the word level if focused on surface structure. This process will be influenced by features of the text (for instance, based on its complexity), but also depends on task demands and individual differences (Smiler et al., 2003; Stine-Morrow & Miller, 1999; Stine-Morrow et al., 2004, 2006a, 2008). Resource allocation is operationalised in this research by regressing reading times on text properties; e.g., a substantial change in reading times as word-level properties such as lexical frequency are manipulated implies a high allocation to word-based processing.

---

[30] Note: the term "resources" is not intended to mean the same thing as working memory resources; for example, Smiler et al. (2003) found that adding load via a concurrent task did not reduce the amount of overall resources available for allocation, even if load can influence how those resources are *allocated*.

[31] There is a similarity to van den Broek's *standards of coherence* (e.g., van den Broek et al., 2001), except that standards of coherence are thought of as being an overall goal for the text, while Stine-Morrow's model conceptualises multiple "standards" for a given text, at the various levels of comprehension.

Stine-Morrow et al. (e.g., 2006a, 2006b) have applied the SRLP to ageing. They argue that reduced efficiency with age makes goal-directed resource allocation more effortful, and so this allocation mechanism does not sufficiently accommodate for other effects of ageing. This could explain the increased reliance seen in older adults on discourse-level processing (i.e., integrating the content of text with existing knowledge) over surface- and textbase-level processing. On this view, allocation is not shifted towards these lower levels of processing, even if they are required for the task. Stine-Morrow et al.'s research has supported this account, finding that unlike younger adults, older adults are less adaptive in resource allocation when changing from an accuracy goal to a speed goal (Stine-Morrow et al., 2006a), or when reading difficult parts of complex sentences (Stine-Morrow, Ryan, & Leonard, 2000), suggesting a decline in self-regulation. Instead, older adults engage in more discourse-level processing on their first pass, with more lexical processing than younger adults in re-reading (Raney, 2003; Stine-Morrow et al., 2004).

Other research from Stine-Morrow and colleagues points towards individual differences even within ageing: older adults with high working memory capacity seem to show less difficulty in resource allocation (e.g., Payne et al., 2014; Stine-Morrow et al., 2002). This again suggests that the effects of ageing, especially in high span older adults, may not be immutable, and may be capable of (unconscious) strategic manipulation. Stine-Morrow and Miller (2009) link the SRLP model to Ferreira and colleagues' good-enough framework, suggesting that good-enough effects in comprehension may be conceptualised as insufficient resource allocation to textbase processing (or indeed, vice versa).

4.1.3. Dual-task performance in older adults

One way of separating cognitive limitations from strategic shifts of behaviour is by manipulating task demands, such as by using the dual-task paradigm used in Chapter 2. Dual-task studies are useful as they require executive control for participants to assign their attentional and memory resources (Baddeley, 1996); the demands of the second task may alter resource allocation during reading (Smiler et al., 2003). They therefore allow exploration of strategic differences (however unconscious) in how different people respond to cognitive demand. If age differences in reading ability were mediated by limitations in working memory, inhibition or general processing speed, older participants should perform worse on dual-task studies containing a reading task.

The evidence for a decline in dual-task performance with age is variable. There is some evidence for poorer dual-task performance in older adults (Goethe et al., 2007; Hartley, 1992; Kemper & Herman, 2006; Riby et al., 2004; Salthouse, 1994; Verhaeghen, Steitz, Sliwinski, & Cerella, 2003). Especially relevant is Kemper and Herman (2006), who found that older readers slowed down their reading when asked to simultaneously recall words, regardless of what those words were (in contrast, younger readers were only affected if to-be-recalled words matched the content of sentences, such as recalling men's names while reading a sentence containing other men's names; see Connelly et al., 1991 for a similar finding). Kemper and Herman found no link to performance on a separate inhibition task, and attributed the result to working memory limitations (but cf. Smiler et al., 2003, who found no link between dual-task performance and working memory measures). Other studies have failed to find an age difference in dual-task studies (Baddeley et al., 2001; Brebion, 2003; Hartley, 2001). Intriguingly, several studies have found that older participants can actually perform *better* on these tasks (Brebion, 2001; Daneman et al., 2006; Kemper et al, 2003, 2004, 2009; Stine-Morrow et al., 2001). These

differences in results undoubtedly stem in part from differences in methodology (in particular the nature of the two tasks, and the degree to which they rely on executive control as opposed to more automatic processes, such as with implicit memory tasks) and in participants (for reviews, see Goethe et al., 2007; Riby et al., 2004; Verhaeghen et al., 2003). The evidence of improved performance suggests that older readers do not always perform poorly on dual-task studies as an inevitable consequence of age-related limitations. Rather, these results may reflect older adults' use of different, more conservative strategies to approach the dual-task situation (Glass et al., 2000; Kemper et al., 2009). Alternatively, it has been suggested that older readers have a "cognitive reserve" that is only invoked when faced with high cognitive load (Brebion, 2001).

The above results referred to dual-task performance specifically; however, differences in strategic approaches may also explain reading differences with ageing, as measured by both eye movements and comprehension. It has been suggested that older readers are more conservative, favouring accuracy over speed (e.g., Daneman et al., 2006; Rabbitt, 1979; Salthouse, 1996; Smith & Brewer, 1995; Starns & Ratcliff, 2010). This assertion is supported by behavioural (Glass et al., 2000; Goethe et al., 2007; Kemper et al., 2009; Oztekin, Gungor, & Badre, 2012; Payne et al., 2014) and neuroimaging research (e.g., Forstmann, Tittgemeyer, Wagenmakers, Derrfuss, Imperati, & Brown, 2011). For example, using a speed-accuracy trade-off paradigm, Starns and Ratcliff (2010) found that while young participants speeded up responses on demand, older adults continued to respond slowly in order to minimise their error rate. A shift towards a conservative accuracy-driven strategy would explain the added re-reading seen in older readers when they detect error or ambiguity: the increased time spent reprocessing complex sentences is converted into a better understanding of what has been read.

If older adults seek accuracy over speed, why do they show riskier eye movement patterns (DeDe, 2014; von der Malsburg et al., 2015) and impaired comprehension on questions tapping syntactic ambiguities (Christianson et al., 2006)? It seems counterintuitive to read more quickly by relying on predictions, while simultaneously prioritising accuracy over speed. This can however be reconciled by reference to strategic use of heuristics in order to compensate for declines in cognitive abilities such as working memory. Although use of heuristics at first seems inconsistent with a conservative, accuracy-driven strategy, the two are compatible. On this view, heuristics will rely on older adults' increased knowledge and experience and will generally result in accurate responses (cf. risky lexical processing in Rayner et al., 2006). In everyday life, the strategy is effective; it is only in the difficult conditions of garden-path experiments that this strategy might be non-optimal. By using heuristics, time and resources remain for more detailed processing on those occasions where heuristics fail, especially for older adults with good working memory (Stine-Morrow et al., 2006b).

Experiment 6

4.2. Background to this experiment

Reading in older adults has been studied using eye movements and answers to comprehension questions, but there has been far less research comparing both simultaneously (but cf. Daneman et al., 2006; Kemper et al., 2004, 2008; Kemper & Liu, 2007; Kemper & McDowd, 2006; Khan & Daneman, 2011). Furthermore, dual-task studies of reading in older participants rarely use eye-tracking to explore real-time online processing (but see Kemper & McDowd, 2006; Kemper et al., 2008), preferring to use self-paced reading times or to only consider comprehension without online measures. To remedy this, Experiment 6 explored comprehension and eye movements concurrently, to look at how the imposition of a concurrent task influenced reading in older readers. Using the same stimuli and design as Experiment 1 provided an opportunity to compare the younger readers under load in Experiment 1 to older readers in Experiment 6. In Chapter 2, I asserted that the eye movement patterns of younger adults under load resembled those seen in older readers in previous research[32] (with more regressions and longer total reading durations; Kemper et al., 2004); Experiment 6 allowed this assertion to be tested.

Like Experiment 1, Experiment 6 compared garden-path sentences presented in isolation, to unambiguous sentences containing a disambiguating comma. Questions after each sentence tapped the initial misanalysis. This reading task was combined with a between-participants manipulation of load: a concurrent 2-back task provided insight into the effects of load on eye movements and comprehension in older readers, and more generally, into how older readers respond to a dual-task. Would they prioritise the reading

[32] This is similar to findings that older participants with high working memory spans resemble low-span younger participants (Kemper et al., 2004), although the comparison between effects of age and of individual differences in working memory is not always valid (e.g., Jost, Bryck, Vogel, & Mayr, 2011; Payne et al., 2014).

task in pursuance of greater accuracy, the n-back task for accuracy on that instead, or show a difference in performance on both tasks?

The first predictions were for comparisons with younger readers in Experiment 1. I predicted that the no-load older readers would show longer reading durations than younger readers (Kliegl et al., 2004) – especially in re-reading measures. An effect in re-reading would resemble the patterns of 2-back younger participants in Experiment 1. As for comprehension accuracy, the results were expected to replicate Christianson et al. (2006): poorer comprehension in older adults relative to younger adults, but only on items containing OPT verbs.

The mixed results from previous dual-task studies with older adults made it difficult to predict effects of load. A reasonable expectation was a decline in comprehension accuracy as a result of the dual-task. The only reason this might not happen would be if the older readers disregarded and performed poorly on the n-back task, prioritising comprehension accuracy instead (cf. Smiler et al., 2003). For eye movements, differences associated with ageing were expected to become even more exaggerated with increased load, manifesting in longer re-reading times. This was likely to be moderated by verb type (OPT vs. RAT), based on Christianson et al's (2006) findings that older adults only showed a difference from younger adults on question answering for OPT verbs (and not for RAT items). Verb type was of more theoretical importance in older adults, and results comparing the two verb types are therefore discussed in more detail here than in Experiment 1.

4.3. Method

4.3.1. Participants

Thirty-two older adults (mean age = 74 years, range = 64 – 86 years, 16 male) took part in this experiment. The original intention was to find a similar number of participants to Experiment 1; difficulties in recruitment of suitable older adults prevented this, but the sample size remained larger than several of the experiments in Christianson et al. (2006). All reported corrected-to-normal vision, and no diagnoses of sensory, reading or other cognitive impairments. Participants were university-educated (from a range of subjects, but primarily humanities and social sciences), and were contacted via alumni offices at colleges of the University of Oxford (n = 24), or as part of a control participant pool at the University of Oxford's Cognitive Neuropsychology Centre (n = 8). All received £10/hour as compensation. As in Experiment 1, participants were allocated into two groups. One half completed the experiment with a concurrent memory task (*2-back* group) and the other half complete the same reading task, but without the additional memory task (*no-load* group). All participants gave informed consent in accordance with ethical approval from the University of Oxford's Medical Sciences Interdivisional Research Ethics Committee.

4.3.2. Materials and Procedure

4.3.2.1. Eye-tracking methodology. Eye-tracking methodology was the same as in Experiment 1. To reiterate, an Eyelink 1000 eye-tracker (SR Research) recorded right eye movements at 1000 Hz. The eye tracker was set up and calibrated in accordance with the instructions; calibration was to <.5° using three fixation crosses. The task was programmed using EyeTrack (http://www.blogs.umass.edu/eyelab/).

4.3.2.2. Reading task. The task was identical to the one used in Experiment 1. Following drift correction, participants read a sentence on the screen, before pressing a button on a controller to advance to the next screen. The sentence disappeared after 10s. There were again 32 experimental items (see Appendix A), all taking the form of: *While the father changed(,) the baby played with its toys*. Participants saw 16 of these as a *garden-path* sentence (with no comma), and 16 as a *comma* sentence (with the disambiguating comma), this variable being Structure. Whether a given item was seen with or without a comma was counterbalanced across participants. Experimental items were interspersed with 40 filler sentences that resembled the experimental sentences but without any syntactic ambiguity. Half of the stimuli contained OPT verbs, and half contained RAT verbs. Christianson et al. (2006) found significant differences in how older readers interpreted sentences with these two verb types, and the effect of Verb (*OPT* vs. *RAT*) was therefore also included. The sentences varied in length; for similar reasons to Experiment 1, this is not discussed in detail here. Including length into the statistical models did not affect the pattern of results or produce any additional effects.

Comprehension was assessed with a yes-no question that followed each sentence on a new screen. The 32 experimental items were followed by questions such as *Did the father change the baby?* for which the correct answer was always no. Sixteen of the 40 filler sentences were also followed by a question for which the correct answer was always yes. Participants indicated their response to questions using the controller. The next sentence then appeared.

4.3.2.3. Working memory load task. Participants in the 2-back condition completed a concurrent task designed to tap verbal working memory resources. The task was the same as in Experiment 1, using twenty-four words in capital letters embedded in filler sentences, which participants were instructed to recall. On eight unpredictable

occasions during the experiment, a two-alternative forced-choice question ("What was the second-to-last word you saw in capitals?") was presented, with the foil alternative either being a similar word, or one seen in elsewhere in the experiment.

4.3.2.4. Digit span. At the end of the reading task, participants were tested on a forward digit span task, and then a backward digit span task. (Four participants did not complete these tasks due to time constraints). On each trial, the experimenter read a series of digits to the participant, with one digit read every second. At the end of the series, the participant was asked to recite the series of digits in the correct order, either forwards or backwards as required. Participants were tested firstly with three trials at series length 3, and then three trials at series length 4, and then three at length 5 etc., until they incorrectly recalled the series on at least two of the three trials at a certain level.

Overall performance was good on both the forward (M = 7.2, SD = 0.9; range = 6 – 9) and backward (M = 5.5, SD = 1.0; range = 4 – 8) span tasks. These results demonstrate that participants had comparable working memory with older adults used in other cognitive psychology studies (in Bopp & Verhaeghen's (2005) meta-analysis, the means for older adults were 7.06 and 5.34, respectively). The digit span tasks are not discussed further: digit span, especially forward digit span, has a weaker connection to reading ability and individual differences with ageing than, for example, the reading span task used in Experiment 5 (cf. Bopp & Verhaeghen, 2005; Daneman & Merikle, 1996), and the dependent variable of maximum span length is not a sufficiently sensitive measure to explore individual differences with this sample size.

4.4. Results

Eyelink software (http://www.blogs.umass.edu/eyelab/) was used to analyse eye-tracking records. Two participants were removed due to excessive tracker loss (both in the no-load condition), leaving 30 participants in the final analysis. Trials were removed when participants blinked or otherwise did not fixate on the disambiguating verb (9.2 % of trials). Fixations above 800ms or below 80ms were also deleted. As discussed in Chapter 2, these exclusion criteria differ slightly from those used in Experiments 2 to 5 – but are identical to the ones used in Experiment 1, allowing a direct comparison between these datasets.

Data were analysed using linear mixed-effects models (LMEs) in the *lme4* package (Bates et al., 2015) in R, with the same log transformation, model criticism and maximal random effects structure as in Chapter 2. Results are discussed first for the older participants only. All models had Structure (garden-path vs. comma), Verb (*OPT* vs. *RAT*) and Load (no load vs. 2-back) centred and entered as fixed-effects. These data were then compared to the data from the younger participants in Experiment 1. Models combining data from the two experiments contained fixed effects of Structure, Load, Verb and Age (younger vs. older), and all interactions. These models were complex (with 4-way interactions and, initially, 3-way random slope interactions on Item); accordingly, most models in this experiment required simplification of the random effects structures to reach convergence.

4.4.1. N-back task performance.

Performance on recalling words in the n-back task was notably lower than for younger participants in Experiment 1 (M = 62.5%, SE = 4.3%; for younger participants,

M = 84.6%, *SE* = 2.5%); this was confirmed by a two-sample *t* test, *t* = 4.45. Overall performance in older participants was significantly above chance according to a one sample *t* test, *t* = 2.91, although six out of the 16 2-back participants performed at or below chance. This demonstrates that while some attention was paid to the concurrent task, the older adults did not achieve the high levels of accuracy seen in younger readers.

4.4.2. Performance on comprehension questions

Figure 4.1 shows mean accuracy across Structure, Load and Verb. There is little difference between the two Load groups. Collapsing across Load, there is a comparable garden-path effect for both verb types, but accuracy is much poorer for OPT items than for RAT items. Confirming these findings, an LME model found that accuracy was significantly poorer for questions after garden-path sentences (*M* = 74.8%, *SE* = 2.1%) than comma sentences (*M* = 81.3%, *SE* = 1.9%), *z* = 2.29. Accuracy was poorer on OPT items (*M* = 65.0%, *SE* = 2.3%) than RAT items (*M* = 90.5%, *SE* = 1.4%), *z* = 4.57; the Structure by Verb interaction was not significant, *z* = 1.45. There was no difference between the two Load groups, and no significant interactions with Load, all $z < 1$.[33]

[33] As Load had little effect, models were re-run without Load, to ensure they had not been overfitted for the available data. The simpler model produced the same pattern of results, with near-identical estimates.

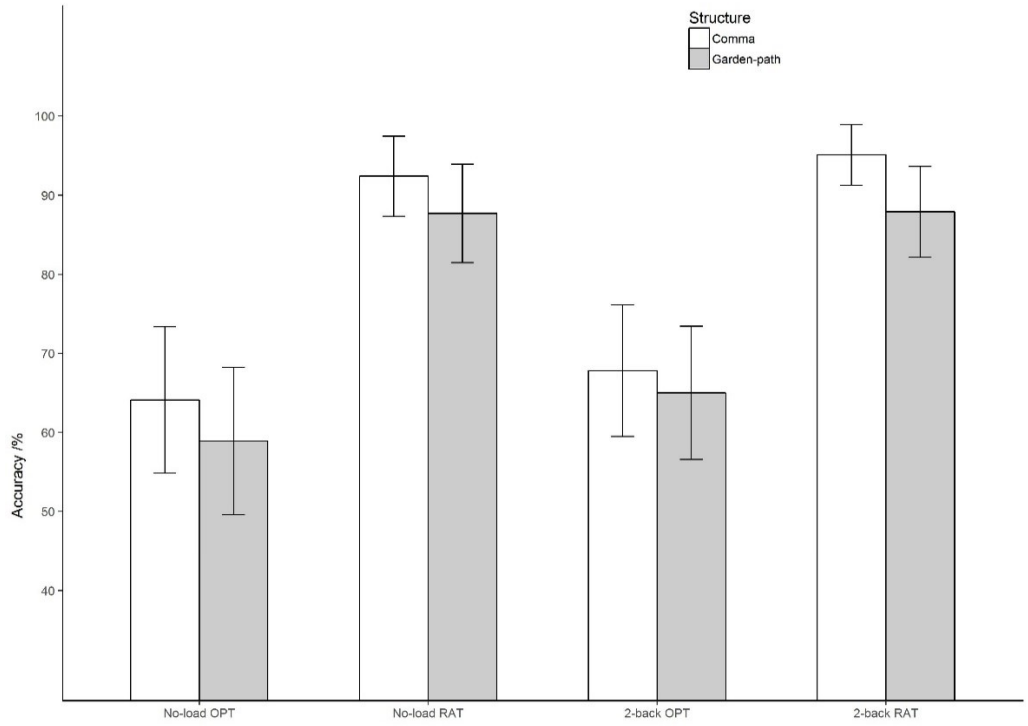

Figure 4.1. Mean accuracy for comprehension questions by older participants, by Load (no-load on left two sets of bars; 2-back on right two sets of bars), Verb (on first and third sets of bars) and Structure (white bars representing comma items; grey bars garden-path items). Error bars indicate 95% confidence intervals.

4.4.3. Effects on eye movements

For comparison with Experiment 1, the same four dependent measures were analysed: *first pass duration*, *go-past duration*, *total reading duration*, *and second-pass duration*. For all measures, fixation durations are for the critical disambiguating verb (*changed* in *While the father changed(,) the baby played with its toys.*) Once more though, additional measures are presented in Table 4.1, by Structure and Load. Like in Experiment 1, these additional measures were analysed using linear mixed effects models. There was a main effect of Structure on first fixation durations, $t = 2.28$, regressions in, $z = 2.44$, and regressions out, $z = 3.35$. Other than a marginal effect of Load on regressions in, $z = 2.01$, no other effects were seen on these measures.

Table 4.1.

Additional descriptive statistics for Experiment 6, by Structure and Load (standard errors in brackets). All measures are as defined in Table 1.1.

	No-load		2-back	
	Comma	G-path	Comma	G-path
First fixation duration	243 (5)	273 (8)	237 (6)	263 (8)
First pass duration	268 (8)	309 (10)	264 (9)	293 (9)
Go-past duration	350 (25)	660 (59)	344 (23)	517 (46)
Second pass duration	358 (21)	363 (22)	349 (25)	388 (33)
Total reading duration	413 (19)	514 (22)	442 (17)	559 (22)
% regressions out	13.9 (2.7)	30.8 (3.6)	14.1 (2.5)	24.6 (3.1)
% regressions in	32.7 (3.6)	40.7 (3.8)	47.6 (3.6)	60.2 (3.6)
% skipping	20.7 (2.8)	19.2 (2.7)	23.9 (2.7)	22.7 (2.7)

First pass duration is shown in Figure 4.2 (Panel A). While an overall garden-path effect can be seen, the only sizeable effect was for no-load participants on RAT verbs. An LME model confirmed that first pass durations were significantly longer for garden-path items [M = 301ms, SE = 7ms] than comma items [M = 266ms, SE = 6ms], t = 3.27. There were no further significant main effects or two-way interactions, all $t < 1$, but there was a three-way interaction between the variables, t = 2.45.

The three-way interaction can best be explained as follows (figures in parentheses refer to the percentage increase between the mean reading durations for comma items and for garden-path items; i.e., the size of the garden-path effect). For no-load participants, Figure 4.2 (Panel A) shows a clear difference between the two verb types: OPT items showed no real garden-path effect (5% increase), while for RAT items, the difference is clear (25% increase). In contrast, older adults under load showed a trend towards a small

garden-path effect in both verb types (14% for OPT items, 8% for RAT items). Put another way, the two Load groups showed similar patterns on OPT items, but for RAT items, the no-load group showed a clear garden-path effect while the 2-back group did not. Planned comparisons supported this interpretation: in no-load participants, a main effect of Structure, $t = 2.68$ was qualified by an interaction with Verb, $t = 2.29$; in 2-back participants, the same main effect of Structure, $t = 2.48$ (reflecting an overall garden-path effect), was not qualified by an interaction, $t = 1.16$. There was therefore only evidence for verb type moderating the garden-path effect in *no-load* participants.

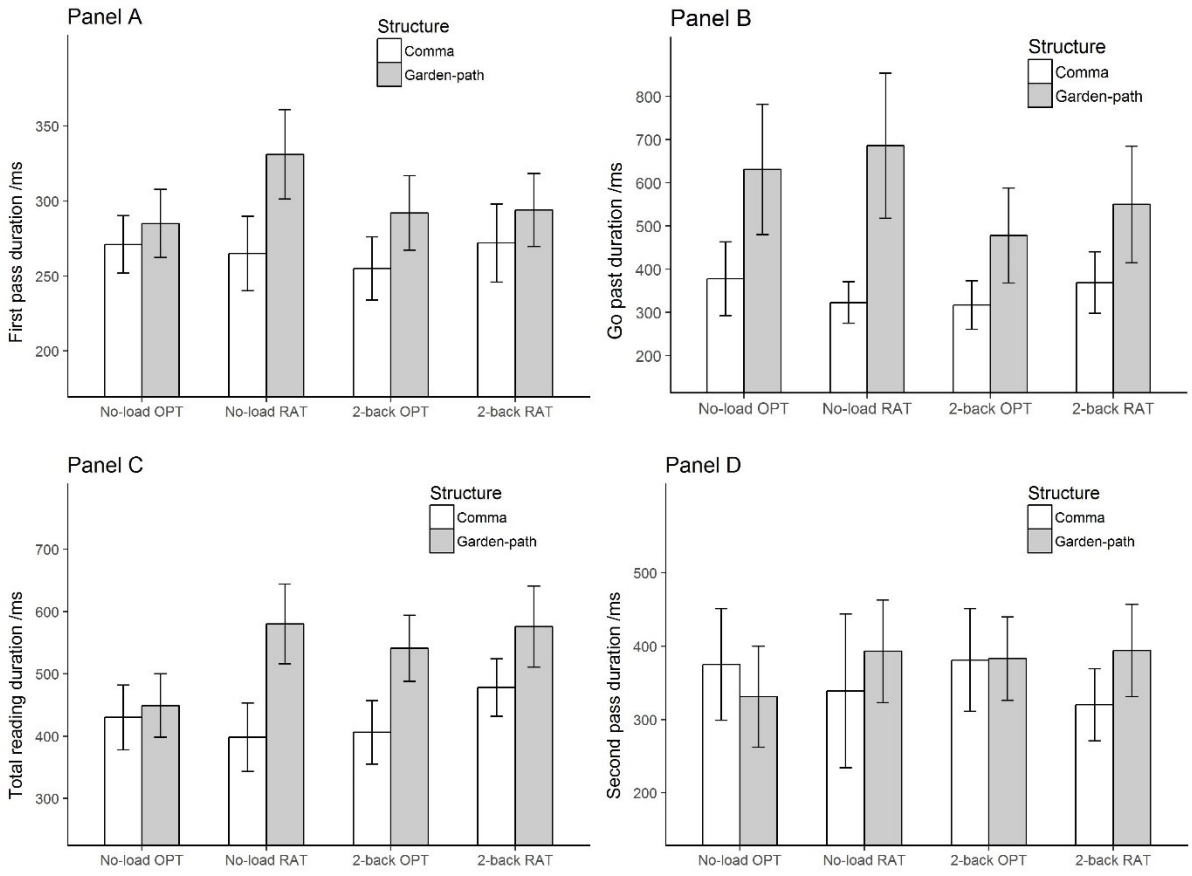

Figure 4.2. First pass (Panel A), go-past (Panel B), total reading (Panel C) and second pass (Panel D) durations, by Load (left two sets of bars for no-load; right two, 2-back), Verb (first and third sets of bars represent OPT items; second and fourth RAT), and Structure (white bars for comma items). Error bars indicate 95% confidence intervals.

For go-past durations, a garden-path effect is visible in Figure 4.2 (Panel B) across conditions. This was supported by a main effect of Structure, $t = 4.46$. There was no Load effect, $t = 1.37$, and no other significant effects or interactions, all $t < 1$. The effect of Structure again reflected longer go-past durations for garden-path sentences [M = 585ms, SE = 37ms] than comma sentences [M = 347ms, SE = 17ms].

Total reading duration on the critical verb is shown in Figure 4.2 (Panel C). A garden-path effect was present in all conditions except for no-load participants on OPT verbs. The LME model supported the effect of Structure, $t = 4.96$, in the expected direction between garden-path [M = 538ms, SE = 15ms] and comma items [M = 429ms, SE = 13ms]. Other than a non-significant trend towards a main effect of Verb, $t = 1.50$, the only other significant effect was a three-way interaction between Structure, Verb and Group, $t = 3.10$.

To explain the interaction (see Figure 4.2, Panel C), I took a similar approach to first pass durations, with similar findings. There was an interaction between Structure and Verb for no-load participants, $t = 2.56$, but not for 2-back participants, $t = 1.48$. No-load participants showed a similar pattern to first-pass durations, with no evidence of a garden-path effect for OPT items (4% increase in reading durations), but a clear effect for RAT items (46% increase). Participants in the load condition instead showed a garden-path effect both for OPT items (33% increase) and for RAT items (20% increase). This suggests that the two Load groups approached each verb type differently.

Finally, second pass durations were examined. Figure 4.2 (Panel D) shows no clear pattern of results, with reasonable variation between trials. The only discernible sign of a garden-path effect is for RAT verbs. Older adults made a second pass on a similar proportion of trials to younger adults in Experiment 1 (42% of comma trials; 54% of

garden-path trials). The small numerical difference between garden-path [M = 379ms, SE = 16ms] and comma items [M = 352ms, SE = 18ms] was not reliable, t = 1.26. A Structure by Verb interaction, t = 2.47, reflected a significant garden-path effect for RAT verbs [M_{COM} = 326ms, SE_{COM} = 24ms; M_{GP} = 394ms, SE_{GP} = 24ms], t = 2.89, but not for OPT verbs, $t < 1$. No other effects or interactions were significant, all $t < 1$.

To summarise, the imposition of load via a concurrent task had no effect on comprehension accuracy, although performance on the 2-back task was reasonably poor. As predicted, and as in Christianson et al. (2006), performance on the comprehension questions was affected by both Structure and Verb. Load did affect eye movements, with three-way interactions of Structure, Verb and Load in first-pass and total reading durations. As total reading duration includes first pass durations, this suggests that the effect was reasonably early in processing, especially as the three-way interaction was not seen in go-past or second pass durations. No-load participants showed a clear difference between reading patterns for OPT verbs and RAT verbs, with garden-path RAT items causing a disruption not seen in OPT items. In contrast, the 2-back participants showed similar garden-path effects in both verb types.

4.4.4. How do older readers compare to younger readers?

Results from older adults were compared to those of the younger adults in Experiment 1. First, Table 4.2 presents a comparison of younger and older adults on a number of sentence-level measures. Linear mixed effects models were run on these statistics to look for effects of age: there was a main effect of age on sentence reading time, t = 4.28, average fixation duration, t = 2.80, number of fixations, t = 2.93, forward saccade length, t = 2.41, and average saccade duration, t = 7.22. All showed longer

reading times, longer saccades and more fixations in older adults, as seen in previous research (e.g. Kliegl et al., 2004; Rayner et al., 2006).

Table 4.2.

Sentence-level and selective region-level statistics for younger adults in Experiment 1 compared to older adults in Experiment 6. Standard errors in brackets.

	Younger readers	Older readers
Sentence reading time /ms	3573 *(42)*	4562 *(59)*
Average fixation duration /ms	201 *(1)*	221 *(1)*
Number of fixations /trial	15.6 *(0.2)*	18.2 *(0.2)*
Forward saccade length /characters	8.7 *(0.1)*	9.1 *(0.1)*
Average saccade duration /ms	27.8 *(0.1)*	34.0 *(0.1)*
% regressions out of region	18.6 *(1.1)*	20.9 *(1.5)*
% regressions in to region	47.1 *(1.4)*	45.9 *(1.9)*
% skipping of critical region	17.0 *(0.9)*	21.7 *(1.4)*
First fixation duration in region /ms	219 *(2)*	254 *(3)*

Figure 4.3 shows comprehension accuracy for both age groups, across Structure and Verb (results are collapsed across Load for ease of presentation, given the absence of Load effects on comprehension accuracy in both age groups). There are again visible effects of Structure and Verb in both age groups, with the effect of Verb especially apparent in older participants; consistent with this, an LME model found main effects of Structure, $z = 3.28$ and Verb, $z = 4.09$, and a trend towards an Age-by-Verb interaction, $z = 1.90$. This trend indicated that on OPT items, older readers [$M = 64.1\%$, $SE = 2.3\%$] were less likely to answer No than younger readers [$M = 73.8\%$, $SE = 1.5\%$]. In contrast,

the two age groups were near-identical on RAT accuracy, with just 0.5% separating them. There was no overall effect of Age, $z = 1.32$, and there was no evidence that omitting commas affected overall comprehension in the older adults more than younger ones, with no Age-by-Structure (or Age-by-Structure-by-Verb) interaction. Other than a trend towards a Structure-by-Verb interaction, $z = 1.60$ (representing a trend towards a greater garden-path effect in RAT items), no other effects reached significance.

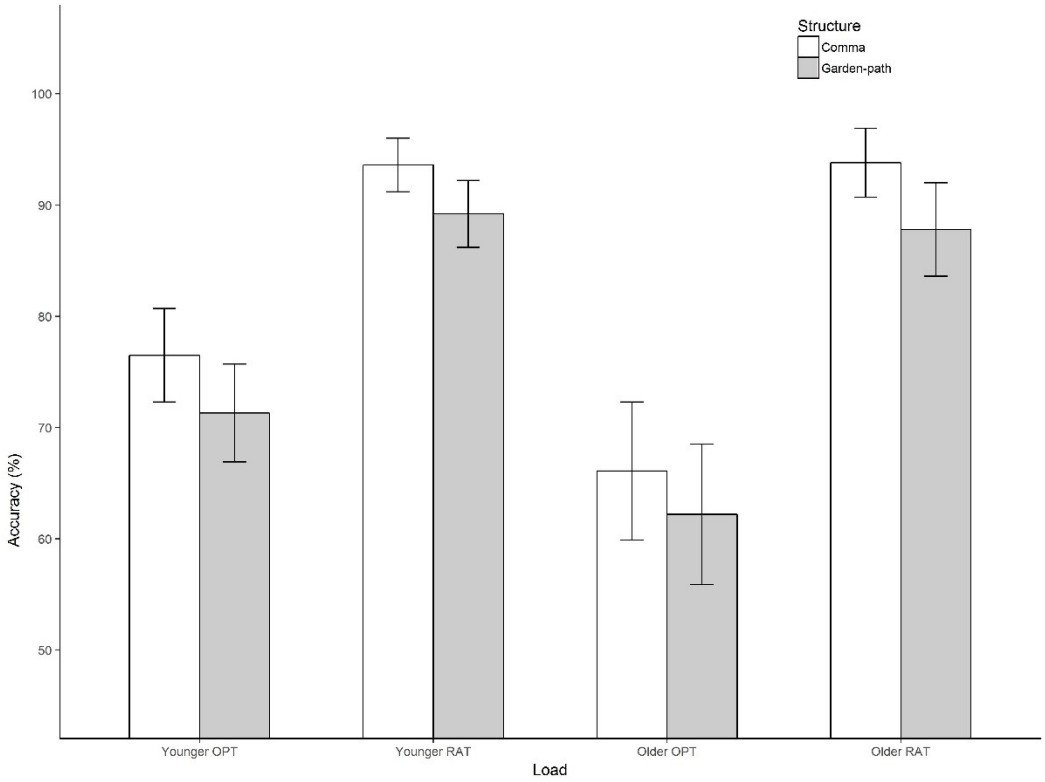

Figure 4.3. Comprehension accuracy by Age (left two sets of bars for younger; right two, older), Verb (first and third sets of bars for OPT verbs; second and fourth RAT verbs) and Structure (white bars for comma items). Error bars indicate 95% confidence intervals.

Turning to reading durations, the model for first-pass durations produced main effects of Structure, $t = 4.17$, and Age, $t = 3.55$, demonstrating longer reading durations for garden-path items than comma items, and for older adults than younger adults. The Structure-by-Age interaction, $t = 1.69$, was not significant, but pointed to a greater garden-path effect in older readers[34]. The only significant interaction was Structure-by-Load-by-Verb, $t = 2.06$, as seen in older adults; there were also a trend towards a 4-way interaction, $t = 1.85$ (perhaps reflecting the three-way interaction seen in older adults, but not in younger ones). Most of these results therefore reflect those presented for each group separately.

For go-past durations, the same effects of Structure, $t = 5.05$ and Age, $t = 2.81$ were found, alongside a Load-by-Verb interaction, $t = 2.19$: 2-back participants took longer to go-past RAT verbs than OPT verbs, while no-load participants show no difference (note: this interaction was not significant in either age group separately).

For second-pass durations, there were similar effects of Structure, $t = 2.78$, and of Age, $t = 2.02$. There was also the Structure-by-Load interaction reported in younger adults in Experiment 1 (but not seen in older adults here[35]), $t = 2.06$, and marginal interactions of Age-by-Verb, $t = 1.92$, and Structure-by-Age-by-Verb, $t = 2.01$. This supports the view that a garden-path effect was seen in both age groups for RAT verbs, with only younger participants showing an effect in OPT verbs.

Finally for total reading durations, there were again main effects of Structure, $t = 7.45$, and of Age, $t = 2.74$, with a trend towards a main effect of Verb, $t = 1.83$. There was also a significant Structure-by-Load interaction, $t = 2.53$, qualified by a Structure-by-Load-by-Verb interaction, $t = 2.35$; both were qualified by an interaction between all four

[34] The size of the garden-path effect (i.e., mean garden-path first pass durations minus mean comma first pass durations) was 35ms in older adults (266ms vs. 301ms), but only 13ms (240ms vs. 253ms) for younger adults.

[35] A trend towards a Structure*Load*Age interaction, $t = 1.68$ may reflect this distinction.

variables, $t = 2.65$. The clearest explanation of these interactions can be seen by looking back at the results presented earlier and displayed on Figure 4.4, which shows the results for total reading duration for OPT (Panel A) and RAT (Panel B) verbs, each split by Age, Group and Structure. As seen in Figure 4.4, the Structure-by-Load interaction is seen regardless of verb type in younger readers, but is stronger for OPT verbs in older readers. The results in Figure 4.4 also support the original hypothesis that younger readers under load would resemble older readers without load: this is far more apparent in RAT verbs, where the pattern of data from 2-back young and no-load old participants are almost identical.

In summary, there were three main differences between the age groups. First, there was no significant age difference in comprehension overall (or for RAT items alone), but older adults did show reduced accuracy on questions relating to OPT items, with a marginal Verb by Age interaction. Second, older adults had longer reading durations overall. There was some evidence that garden-path effects on eye movements were greater in older adults, with a trend towards a Structure by Age interaction in first pass durations, and a four-way interaction (with Load and Verb) in total reading durations. The final and most clear effect of age was in how the two verb types were differentially read, especially as measured by total reading durations. For RAT sentences, no-load older adults resembled 2-back younger adults, as predicted in Experiment 1. For OPT items, younger adults showed similar eye movements to those seen on RAT items. In contrast, no-load older adults showed no garden-path effect when reading OPT items. Older adults under load did show a garden-path effect, producing a four-way interaction.

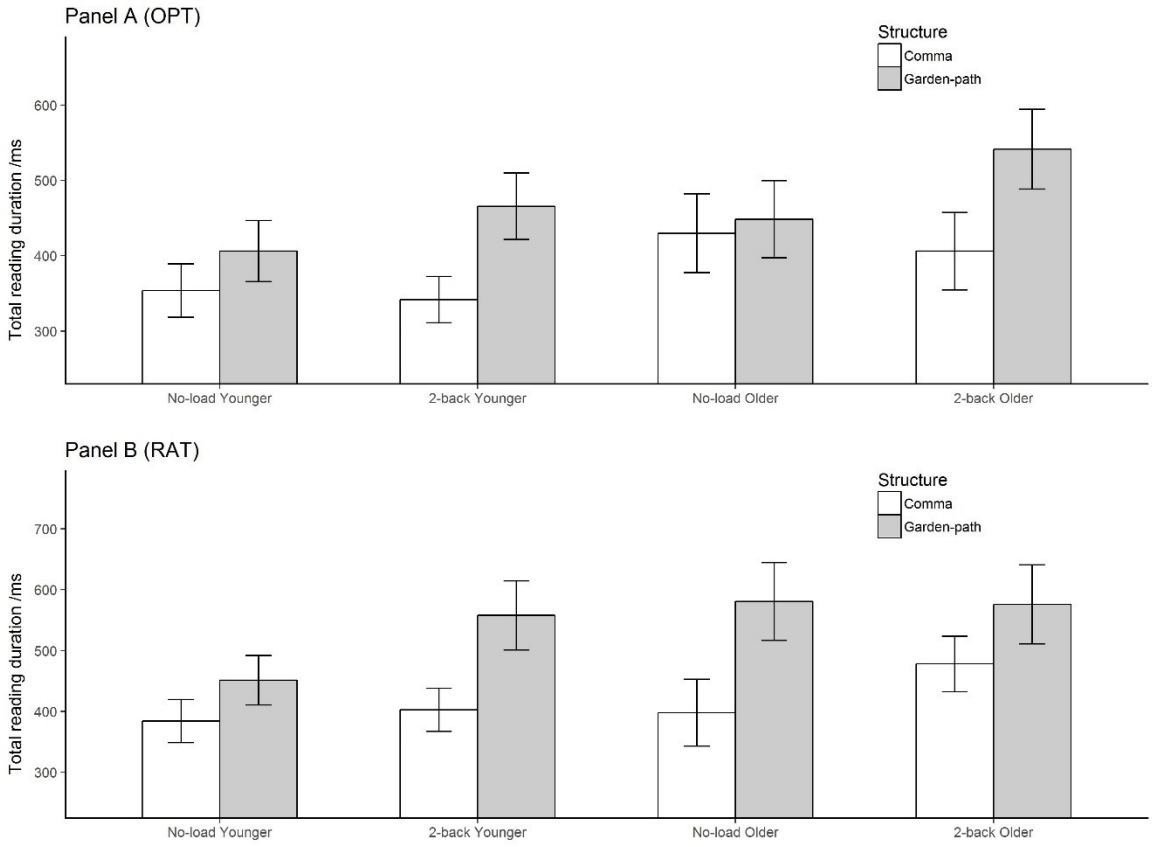

Figure 4.4. Total reading durations for OPT items (Panel A) and RAT items (Panel B), split by Age (left two sets of bars for young; right two, old), Load (first and third sets of bars indicate no-load; second and fourth, 2-back) and Structure (white bars for comma items). Error bars indicate 95% confidence intervals.

4.5. Discussion

This experiment had two aims: first, to look at similarities and differences in older adults' eye movements and comprehension compared to younger adults in Experiment 1, and second, to consider how a concurrent task affected this. The concurrent task had little effect on comprehension accuracy in the older adults, an absence also seen in younger adults in Experiment 1. In the eye movement record, the garden-path effect was notably stronger for RAT verbs than OPT verbs. Furthermore, there was a difference based on Load: no-load participants showed a significant difference between the two verb types (with only a small, non-significant garden-path effect for OPT items), while 2-back participants showed a moderate garden-path effect on both verb types. Comparing the two age groups, older adults performed less well than younger adults on the comprehension questions for sentences containing optionally-transitive (OPT) verbs (but regardless of Structure). There was no such age effect for comprehension of sentences containing reflexive absolute transitive (RAT) verbs, with a marginal Age-by-Verb interaction. Verb type was also an important factor in older adults' eye movements, moderating the extent of the Structure*Load interaction in total reading durations (the interaction was clearer for OPT items); this was not seen in the younger adults. Finally, the original prediction that younger adults under load would resemble older adults reading without concurrent load was borne out: 2-back younger participants showed similar eye movements to no-load older participants, but interestingly, only on RAT items.

An initial point to reflect on is the comparatively poor performance by older participants on the concurrent n-back task. It was noteworthy in Experiments 1 and 2 how well younger participants did on the 2-back task while simultaneously performing well on comprehension accuracy. In contrast, older readers struggled with the n-back task: the average participant only answered 5 out of the 8 memory items correct. This suggests that

older readers prioritised sentence processing over word recall – a finding that has been seen before in at least two dual-task ageing studies (Brebion, 2003; Smiler et al., 2003). Importantly however, this does not mean that the concurrent task had no impact on sentence processing: the three-way interactions between Structure, Verb and Load in first-pass and total reading durations negate this conclusion for a start. Instead, the results demonstrate that older readers adopt a different strategy when faced with a dual-task, a strategy that prioritises what they perhaps see as the more important task, namely sentence comprehension.

Another point to note is that this study does not draw any conclusions about individual differences *within* older adults; only on differences between the two age groups. The digit span demonstrated that these participants were broadly comparable to those participating in previous research (cf. Bopp & Verhaeghen, 2005), but digit span is not a strong predictor of comprehension, or of individual differences with ageing (unlike, for instance, the reading span task used in Experiment 5). It is not sufficiently sensitive to act here as a valid measure of individual differences. Nevertheless, its inclusion helps to demonstrate that the sample used here were not extraordinary compared to other studies. While discussing individual differences, one limitation of this study is that participants' vision was not tested. This is recommended, particularly for research using older adults (McGowan, Paterson, & Jordan, 2013). Participants were however asked to bring and wear corrective glasses or lenses that were suitable for viewing a screen at the required distance. It was also checked that they were comfortable with reading during practice trials, and the eye tracker was recalibrated as necessary, mitigating against this concern to an extent.

Returning to the results, there was no effect of load on comprehension accuracy, mirroring Experiments 1 and 2. In Chapter 2, I argued that the 2-back task may not have

been sufficiently demanding. Here, this argument is less persuasive, given that performance on the n-back task was much worse than in younger adults. A more plausible interpretation is that the older readers simply shifted their priority towards performing well on the comprehension questions, even if this served to limit performance on the n-back task. If the load were increased (such as in the 4-back task used in Experiment 3), comprehension in the older readers may begin to be adversely affected; there is though the possibility that older adults will still strategically prioritise the reading task. The older participants here were chosen from a university-educated population to provide a valid comparison for the younger participants in Experiment 1. While this is important for comparisons discussed later, these participants may have been able to use more efficient strategies to overcome the dual-task than older adults in the wider population (cf. Payne et al., 2014), prioritising the task they consider more relevant in order to attain high accuracy. Other participants of a similar age may have struggled with *both* tasks, as seen in dual-task studies previously (Riby et al., 2004; Verhaeghen et al., 2003).

The comprehension data are strikingly similar in pattern to Christianson et al.'s (2006), although like the younger readers in Experiment 1, comprehension overall in this experiment was considerably higher. Nevertheless, older adults' specific increase in Yes responses on OPT items is identical in pattern (note: the Age-by-Verb interaction was only marginal). This pattern of poorer performance on OPT items was similar for comma items (there was no Age-by-Verb-by-Structure interaction), suggesting that the punctuation cue does not extinguish this interpretation. These results provide further support for Christianson et al.'s (2006) conclusion that older readers are more likely to rely on good-enough representations built on propositional content, rather than on verbatim recall of the sentence's syntactic structure. Accordingly, they are less likely to trigger extensive reinterpretation of items such as *While the man hunted the deer ran into*

the woods, since their interpretation that the man hunted *[SOMETHING]* is consistent with the question asking if the man hunted *the deer*. Christianson et al. concluded that older readers use a semantic-based heuristic upon reaching the question: they see the proposition *the man hunted the deer* as consistent with the two propositions from the sentence (*the man hunted [SOMETHING]; the deer ran into the woods*) and answer Yes. Younger readers, with better working memory, are more able to reinstantiate the content of the sentence verbatim when faced with a question, and are more likely to answer No.

For RAT verbs, the story is different. Christianson et al. (2006) presented three main findings about older adults' comprehension. First, older readers had intact representations of the main clause of sentences, with equal performance to younger readers on questions such as *Did the deer run into the woods?*[36]. Second, older readers answered No as readily as younger readers on RAT questions such as *Did the father change the baby?*. This indicates that the noun phrase *the baby* is not available to fill the gap in *The father changed [SOMETHING]*. Finally however, unlike younger readers, older adults did not frequently answer Yes to *Did the father change himself?* (older adults only answered Yes 50% of the time, compared to over 70% in younger adults). This final result suggests that they often had not fully reinterpreted the sentence and cannot fill the gap in *The father changed [SOMETHING]* with the reflexive (i.e., that he got changed *himself*). Taken together, these findings suggest that older readers attach *the baby* to *played with its toys*, leaving the object of *The father changed [SOMETHING]* empty. When presented with the question *Did the father change the baby?*, they answer No as readily as younger readers (as *the baby* is attached to *played with its toys*, and so cannot

[36] Christianson et al. only tested this on OPT items, but as they note (p. 217): "The questions here concern the main clause of the sentences rather than the subordinate clause that contained the RAT or OPT verb. Thus, there is little reason to anticipate that the results would be different had we tested the RAT verb sentences in this manner."

fill the gap in *The father changed [SOMETHING]*). In contrast, when presented with *Did the father change himself?*, older adults (in particular, low-span older adults) cannot reinstantiate the original, accurate syntactic structure and often also answer No, unlike younger readers who instigate reanalysis and answer Yes.

I did not ask questions such as *Did the father change himself?* in this experiment, and so cannot conclude whether the older adults were able to reinstantiate the correct interpretation of RAT items or not. However, Experiment 6 did replicate Christianson et al.'s findings that older readers were as able as younger readers to dismiss the misinterpretation of the father changing the baby. The comprehension accuracy results here are therefore consistent with Christianson et al.'s explanation: older readers sometimes rely on propositional content and do not reinstantiate the exact syntactic structures when presented with a question, while younger readers are more able to fully recall and therefore reanalyse sentences.

Christianson et al. (2006) did not measure eye movements. They did measure reading durations in their first experiment (Experiment 1A), finding a main effect of age, and an interaction between age and what is called here Structure. However, they also found that when participants only had a fixed time to read sentences, the same pattern of comprehension was observed, suggesting that their results (replicated in pattern here) were not a product of older adults re-reading for considerably longer. Nevertheless, there is an opportunity here to connect Christianson et al.'s interpretation of the comprehension data, to differences in real-time online processing. The eye movement results provide further support for Christianson et al.'s account. Considering only the no-load participants for now (for the most appropriate comparison), there is little evidence of a strong garden-path effect for OPT items in either first-pass or total reading durations. This demonstrates that participants neither spent significantly longer incorporating the disambiguating verb

into their existing syntactic interpretation, nor did they spend any longer on returns to the disambiguating verb in order to reinterpret the sentence. In contrast, a clear garden-path effect was observed for RAT items (Figure 4.2). This therefore suggests that unlike younger readers (who showed a garden-path effect in first pass and total reading durations, regardless of verb type), older readers do not tend to reanalyse OPT items, consistent with their subsequent increased rate of answering Yes to questions.

An effect is visible, however, in go-past durations for OPT verbs. An alternative explanation could be that the older participants may detect the ambiguity, but choose to regress out of the disambiguating verb rather than spend additional time processing it (removing any effect in first pass durations). On this account, they would use the time spent regressing to re-read and reinterpret the sentence, avoiding any need to return to the disambiguating verb (removing any effect in total reading duration). If this account were correct, there should be improved accuracy on the garden-path OPT trials where no-load participants made a regression out of the disambiguating verb, compared to trials where they did not. I analysed this *post hoc* to see if such an effect existed. On the 107 no-load garden-path OPT trials, older adults made a first-pass fixation on the disambiguating verb on 81 trials; of these, they made a first-pass regression on 29 trials, and did not on 52. If the second account is correct, there should be higher accuracy on the 29 trials than on the 52. In fact, the results are the opposite: accuracy is far better on the trials where participants *don't* regress (M = 63.5%, SE = 6.7%) than the trials where they do (M = 48.3%, SE = 9.3%). For comparison, I ran the same analysis for RAT items, and no such difference existed (83% accuracy if a regression was made; 88% if it wasn't).

The number of trials is too low to say definitively that the with-regression trials are anomalous; that is, that these trials alone represent the cases where participants face significant processing difficulty. However, there is certainly no evidence to suggest that

participants are using longer go-past durations to attain accuracy. This therefore supports the original conclusion: older readers do not show a notable garden-path effect (and hence, show little sign of disruption) in the eye movement record for OPT items. This absence of reprocessing is consistent with their poor performance on comprehension questions following sentences containing OPT verbs.

For no-load participants, RAT items were clearly detected as ambiguous and reanalysed (as frequently as younger readers, at least), while for OPT items, syntactic analysis is terminated upon reaching a consistent interpretation. How does the addition of load affect this? There was a dampening of this verb effect, with 2-back participants showing a moderate garden-path effect in both verb types, on first-pass, go-past and total reading durations. This suggests that participants reading under load were less sensitive to cues in the text on first-pass. For total reading durations, 2-back participants showed a garden-path effect, even in OPT verbs. One interpretation is that participants traded speed (of reading) for increased question accuracy, especially seeing as participants performed relatively poorly on the n-back task. Thus, 2-back participants read all sentences more superficially at first, with text-based cues being of less relevance to first-pass reading durations. Being aware of their memory limitations, they re-read sentences (of both verb types) to verify their understanding. In contrast, no-load participants reach a consistent interpretation of OPT verbs (for instance, that the man is hunting something, and that there is a deer) and then proceed without any need to re-read.

This account of increased superficiality followed by increased re-reading resembles the results and discussion of younger participants reading under load in Experiment 1. The results here remain consistent with the good-enough approach to sentence processing: older readers under load read more superficially at first, and unlike their no-load counterparts, are less responsive to cues such as the verb type. However, the older

readers later trade additional re-reading time to achieve increased accuracy. The main difference between older and younger 2-back participants is that the younger participants can do this while simultaneously performing well on the 2-back task; it is only at a 4-back level that they are affected (as in Experiment 3). There is some parallel in the results of Gao et al (2012), who found that older adults' reading was affected by moderate visual noise, whereas it took a higher degree of noise to alter the reading behaviour of younger adults. Older participants seemed unable to maintain performance on both tasks, and prioritised one task over the other. The older participants showed no impairment in comprehension, but showed a noticeable decline in n-back task performance. This is consistent with previous literature suggesting age-related limitations in dual-task performance (Goethe et al., 2007; Kemper & Herman, 2006; Riby et al., 2004; Verhaeghen et al., 2003).

A final point to take from the results is the similarity between younger 2-back participants and older no-load participants on RAT verbs. It is well recognised that older readers make more fixations of longer duration (e.g., Kemper et al., 2004; Kliegl et al., 2004; Rayner et al., 2006). The results of Experiment 1 demonstrate that similar effects can be induced in younger readers under load. This suggests that the longer reading durations may be linked to attempts by older readers to ensure high accuracy in the face of limitations in memory and other cognitive abilities (such as inhibition and a decline in general processing speed). The similarity is less clear for OPT verbs. Younger readers showed a clear garden-path effect on OPT items, regardless of load; no-load older readers showed little sign of an online garden-path effect on these items. This supports accounts that older readers are more likely to opt for semantic-based heuristics (Christianson et al., 2006), while younger readers (even under load) use added time for increased *syntactic* processing.

This study has three important implications. First, it provided eye-tracking data that support Christianson et al.'s (2006) account of comprehension differences in older readers. Items containing OPT verbs did not trigger reanalysis in older readers. This was a critical age difference: OPT verbs did not show the effect seen in RAT verbs, or in both verb types in younger readers. The results therefore support Christianson et al.'s suggestion that older readers rely on heuristics more readily: since the two propositions in OPT items [(a) the man is hunting SOMETHING; (b) the deer ran into the woods] are consistent, there is no need to trigger reanalysis, either online, or afterwards when the question is asked. In RAT verbs, online reanalysis is required. This reanalysis process may of course not be completed, as evidenced by Christianson et al.'s (2006) finding that older adults only perform at chance on questions such as "Did the father change himself?"

More theoretically, this study gives an insight into the role of domain-general cognitive processes (such as working memory) in syntactic ambiguity processing. This design does not explain *why* older readers show no garden-path effect for OPT verbs, or why they are considerably less likely to answer No correctly to questions tapping this misinterpretation. That said, Christianson et al. (2006) found that the plausibility of a garden-path sentence (for example, *While the man hunted the deer ran into the woods* vs. *While the man hunted the deer paced in the zoo*) did not affect comprehension performance in older readers; they concluded that older adults were not using world knowledge to aid syntactic analysis (but cf. DeDe, 2015). Experiment 6 provides further support for Christianson et al.'s conclusion that older readers terminate analysis upon reaching a plausible interpretation, most probably due to reduced cognitive resources. This experiment also demonstrated how a dual-task condition affects this performance: participants adopt a more superficial reading style, supplemented by re-reading to attain increased comprehension accuracy. This suggests that the early termination of reanalysis

is strategic: older adults devote additional resources to reading if the task conditions require this, supporting the idea of a "cognitive reserve" (Brebion, 2001). In practice, this means that older readers may read even more superficially at first, but upon reaching an ambiguity that has to be resolved, their goal to attain accurate comprehension leads them to re-read for longer to resolve that ambiguity. This is consistent with Stine-Morrow et al.'s SRLP model: older participants here (whose working memories, based on the digit span task, were reasonably good) allocated additional resources to processing the meaning of the texts during re-reading, to attain the goal of comprehension (cf. Smiler et al., 2003; Stine-Morrow et al., 2004, 2006b).

While a working memory based explanation is most relevant, the results may also be broadly consistent with an inhibition-based theory (e.g., Zacks & Hasher, 1997). On this account, misinterpretations linger because they are not sufficiently inhibited, with age effects reflecting poorer inhibitory abilities. This would be consistent with evidence that older adults are poorer at inhibiting competitors and cascading associations (Hamm & Hasher, 1992; Hasher et al., 2007; Healey et al., 2013; Radvansky et al., 2005); in this case, the "competitors" are the initial misinterpretations, leading to an increased tendency to answer 'yes' when reinstantiated by the question. It is not entirely clear how to distinguish these two accounts, and as mentioned in the introduction to this chapter, the distinction may be less relevant if adopting a cue-based retrieval account of working memory (see Gordon et al., 2012). But if the effects observed here are due to inhibition, it is not clear why there would be a difference between OPT and RAT items, in terms of eye movements and comprehension. This distinction is more appropriately explained with reference to greater use of semantic-based heuristics (Christianson et al., 2006). Representations of the text are processed in accordance with prior expectations: for OPT items, older adults adopt an unwarranted inference that, for example, the man is probably

hunting the deer; for RAT items, this inference is not available, and reanalysis is more veridical to what the text says.

The final implication of this experiment's results is their support for previous findings that older adults can perform well on dual-task experiments, despite their reduced working memory capacity. Although the dual-task affected reading behaviour as indexed by eye movements, it did not impair subsequent comprehension. However, this came at the cost of relatively poor performance on the concurrent n-back task, suggesting that older adults are more likely to prioritise one task.

It would be of interest to see whether these results hold in three further settings. First, how would older readers cope with a more difficult concurrent task, such as the 4-back task used in Experiment 3? The most likely prediction is that comprehension would be significantly impaired, as in Experiment 3; it is also possible that older adults would find the concurrent task too difficult, and simply ignore it in pursuance of maintaining comprehension accuracy. A second question is: how would older adults change their reading behaviour if task instructions were manipulated to encourage prioritisation of the concurrent task rather than the sentence reading task? With a similar manipulation of task instructions, Smiler et al. (2003) found that reading behaviour was unimpaired, although performance on the concurrent task did improve. The third question is how older adults would be affected by removal of the comprehension questions after each sentence, as in Experiment 5. In Experiment 5, there was no manipulation of load – but if load were introduced, older adults may shift their priority in pursuance of "accuracy" on the concurrent task. This may however depend on individual differences between older adults, with high-span adults more able to regulate their processing (Payne et al., 2014).

In conclusion, this experiment provided evidence from eye-tracking to support differences in how older readers interpret different types of early closure garden-path sentences, based on verb type. The results also demonstrated that adding cognitive load via a concurrent task led to more superficial eye movements, although older adults were also likely to prioritise the reading task, re-reading for longer to attain comparable comprehension to younger readers. The question of individual differences in eye movements is taken forward in Chapter 5.

Chapter 5

Scanpaths during syntactic ambiguity resolution

5.1. Introduction

The research presented in Chapters 2 to 4 demonstrated that when reading syntactically ambiguous sentences, both eye movements and eventual comprehension can be affected by load (via a concurrent task), task demands (such as the length of stimuli, and the presence/absence of questions) and individual differences (ageing). With eye movements, two main effects were observed. Re-reading was more common and continued for longer if participants expected to receive questions (the comparison of Experiments 2/3 and 5), or faced a concurrent 2-back load (Experiment 1). There was also some evidence for increased re-reading with age in Experiment 6, although this was limited to sentences containing reflexive absolute transitive verbs. Additionally, the early first pass effect of syntactic ambiguity, which is common in experiments such as these (Clifton et al., 2007; Frazier & Rayner, 1982), was less apparent when reading stimuli embedded in longer passages (Experiments 2 to 5) relative to isolated sentences. Taken together, these results support an adaptive reading process, with more superficial reading where comprehension can still be maintained, but an increase in re-reading where this approach leads to poorer comprehension.

Importantly however, all of these conclusions are inferred from analyses that focused solely on small, critical areas of the text, such as the disambiguating verb. An alternative methodology is to consider patterns of eye movements while reading the entire text. This chapter uses scanpath analyses (von der Malsburg & Vasishth, 2011, 2013) to carry out exploratory analyses of eye movements across the entire texts. This is intended to give a

brief insight into the similarities and differences between traditional eye movement measures and scanpath analyses – and to look at individual differences that may not have been detected by the usual dependent variables.

The standard eye movement measures used in reading research (see Rayner, 1998, 2009) focus either on the duration of time spent on critical words or regions (e.g., first pass duration, total reading duration), or on the probability or nature of saccades from a critical region (e.g., first pass regressions out, or analyses of saccade launch or landing sites). There is good reason for this. First, it has a strong theoretical motivation: if a model of syntactic analysis predicts disruption at a certain point in a sentence, focusing analyses on that region allows for clear testing of this hypothesis across tightly controlled sentences. Running analyses across many regions may also increase the false positive rate (von der Malsburg & Angele, 2017). There is also the practical point that analysing small regions is more tractable, compared with the difficulty of condensing thousands of fixation locations and durations across entire sentences or texts into a valid and usable measure.

For these reasons, few sentence processing studies have looked at scanpaths more broadly. Frazier and Rayner's (1982) seminal paper did consider this, with the aim of determining how the parser reanalysed a sentence after detecting a syntactic ambiguity. By examining both scanpaths (qualitatively) and transitional probabilities of regression landing sites, they concluded that after detecting an ambiguity, the parser returns to the site where the ambiguity arose (*selective reanalysis*). Meseguer et al. (2002) also provided support for selective reanalysis: they found that landing sites of regressions out of the post-disambiguation region tended to fall around the region where the ambiguity first arose.

Frazier and Rayner (1982), and Meseguer et al. (2002) were interested in explaining regressive eye movements during syntactic ambiguity resolution. As discussed in previous chapters, there is evidence that regressions occur for several different reasons: to revise errors, to confirm what has been read before, or perhaps simply to stall while processing of a word is ongoing (Christianson et al., 2016; Meseguer et al., 2002; Mitchell et al., 2008; von der Malsburg & Vasishth, 2011, 2013). A crude measure of the proportion of trials with a regression out of a critical region demonstrates if readers are more likely to make a regression on some trials than others, but does not determine what happens after the regressive saccade. This can be analysed by looking at landing sites of regressions, but this in turn is less informative about the temporal properties of these eye movements: how long do readers regress for, before continuing? It is therefore useful to analyse entire scanpaths when considering regressions, not least because it makes it easier to separate regressions that occur for distinct reasons.

von der Malsburg and Vasishth (2011) introduced a novel analysis method that considered entire scanpaths, and compared these scanpaths on both their spatial and temporal properties. Their analyses used *Scasim*, a measure that captures the similarity between each possible pair of scanpaths (i.e., between each pair of trials in an eye-movement study). Similarity is defined in terms of what is called *edit-distance* – essentially, the number of changes that are necessary to change fixations in one scanpath, into fixations in another (for full details, see von der Malsburg & Vasishth, 2011). The measure takes into account both the spatial properties of fixations, in terms of their x- and y- coordinates (and not just regions of interest), and their temporal properties (the fixation duration). As a simple example, take sentence (5.1) from Chapter 2:

(5.1) While the father changed the baby that was cuddly played with its toys. The father had to finish changing his clothes to pick up and look after the baby.

Participant 1 may reach the word *played*, fixate on it for 300ms, and then make a progressive eye movement to the word *with*, staying there for 200ms. Participant 2 may reach *played*, fixate for 400ms, and then move to *with* for 200ms. Participant 3 may reach *played*, fixate for 300ms, and regress to *baby*, remaining there for 200ms. Participants 1 and 2 have a reasonable similarity: the spatial properties of their scanpaths are identical (i.e., they have fixated in the same place), and the only edit needed to turn Participant 1's scanpath into Participant 2's is the increase in duration of 100ms on *played*. In contrast, the edit-distance needed to change to Participant 3's scanpath is greater: the coordinates of the second fixation are very different, being several words back. As such, there would be high similarity between the scanpaths of Participants 1 and 2, but less between Participants 1 and 3; there would be even less between Participants 2 and 3, since both a temporal (400ms instead of 300ms) and a spatial edit are required.

There are several ways that these similarity scores can be analysed. Two examples can be found in von der Malsburg and Vasishth (2013), and von der Malsburg et al. (2015). In the first example, the authors looked at individual differences in regression patterns, and in particular how this was affected by working memory capacity. They analysed scanpaths collected from reading Spanish sentences containing attachment ambiguities, focusing on regression paths: if a reader regresses out of the disambiguating region, where do they go and for how long? Using cluster analysis, von der Malsburg and Vasishth found three main types of pattern: rereading of the entire earlier part of the text, a rapid regression back by one region, and "checking" regressions, out of the spillover region (following the disambiguating verb) back to the disambiguating verb. For more syntactically complex sentences, regression paths were more likely to follow the re-reading pattern, and this was even more common in high-span readers. The rapid regressions pattern was more common in unambiguous sentences. von der Malsburg and

Vasishth concluded firstly that it is unlikely that the parser makes a fixed decision about reanalysis (contrary to what is suggested by the garden-path model; Frazier & Rayner, 1982). They also concluded that readers with larger working memory capacities are more likely to make attachment decisions, and to spend longer re-reading after being garden-pathed. It is worth reiterating the discussion in Chapters 1 and 2 about how attachment ambiguities differ from the subject-object ambiguities used in this thesis: it is not obligatory to reanalyse attachment ambiguities in order to reach a consistent parse, making an analysis of scanpaths in subject-object ambiguities of interest.

Similarly, von der Malsburg et al. (2015) looked at individual differences in scanpaths; in this study, predictor variables were syntactic complexity and participant age. They analysed scanpaths from the Potsdam sentence corpus, with younger and older adults reading sentences that ranged in complexity (indeed, many sentences were fairly easy, unlike the ambiguous sentences used by von der Malsburg & Vasishth, 2013). They were looking for differences in the *regularity* of scanpaths – how similar scanpaths were to one another. The results showed that scanpaths were less regular as sentences increased in complexity, but also less regular with age – meaning that there was more variation in scanpaths for complex sentences, and for older readers. As discussed in Chapter 4, the interaction of sentence complexity and age was interesting: older adults showed a smaller effect of syntactic complexity, suggesting that their scanpaths were less driven by syntactic properties of the text.

The work by von der Malsburg and colleagues gives insight into how scanpath analyses can be used to investigate sentence processing, and individual differences in sentence processing. They also demonstrate that scanpath analysis can be more informative than conventional eye movement measures, being able to distinguish between different patterns that would normally be collapsed into a single measure. In this final

empirical chapter, von der Malsburg and Vasishth's (2011, 2013) scanpath analysis was used to reanalyse the data from several experiments reported earlier in this thesis. The analyses were fundamentally exploratory, and were intended to investigate effects that may be of interest for future controlled experiments.

The analyses were conducted on the results from Experiments 2, 3 and 5. These are of particular interest because they contain three manipulations that may affect scanpaths more broadly than is captured by conventional eye movement measures. The first manipulation, across all three experiments, was Structure – whether sentences contained a disambiguating comma or not. Garden-path sentences should be associated with a greater proportion of scanpaths containing one or more regressions, in order to reanalyse the sentence and attain comprehension. The second is load, manipulated in Experiments 2 and 3. The conventional eye movement measures found little effect of load on eye movements, although there was a significant decline in comprehension with the 4-back task used in Experiment 3. As traditional eye movement measures may be masking differences in reading behaviour, reanalysing the data using scanpaths offers an additional opportunity to test the effects of load. This might reveal, for example, evidence of more superficial reading in the difficult, 4-back load condition. The third manipulation was the task; namely, the presence (Experiments 2 and 3) or absence (Experiment 5) of questions. In Experiment 5, it was demonstrated that there was significantly less re-reading of sentences when questions were not expected and participants had no reason to believe they were reading for comprehension. The prediction would therefore be for shorter scanpaths with less regressive eye movements in the Experiment 5 data. Beyond these manipulations, Experiment 5 contained a measurement of reading span, with differences found for high-span readers compared to low- and medium- span readers. This variable was therefore also considered, with the prediction that high-span readers would show

scanpath patterns with more re-reading, demonstrating more careful (and less superficial) checking of the text.

Due to methodological issues, data from Experiments 1, 4 and 6 are not analysed here. Unfortunately, Experiments 1 and 6 were programmed using software from which extraction of the fixation-by-fixation data needed for scanpath analysis is more difficult. Experiment 4 used different stimuli from the other experiments, making a comparison more challenging. Furthermore, the four-sentence passages in Experiment 4 will have considerably more fixations, and hence far more variation in those fixations. Drawing clear conclusions from these data may therefore be more difficult.

The analyses took a similar approach to von der Malsburg and Vasishth (2013), using cluster analyses to distinguish eye movement patterns – and assessing whether the three manipulations of structure, load and task influenced the likelihood of scanpaths containing the patterns identified by different clusters. These analyses were exploratory: to make them tractable, they were restricted to two specific subsets of the full dataset. The first subset was reasonably broad: all scanpaths from the point where participants read the disambiguating verb, until reaching the end of the passage. This analysis was intended to explore both regressive and progressive eye movements out of this region, and in particular to explore further the issue of differences between first-pass and go-past reading durations, and whether this was influenced by the three manipulations. This was followed by a more targeted analysis, based on von der Malsburg and Vasishth (2013)'s work: analysis was conducted only on scanpaths where a regressive eye movement was made out of the disambiguating (or spillover) region, until going past the spillover region. This was to distinguish different types of regressive movement, and to see whether these varied based on the key manipulations of structure, load and task that varied within and between experiments.

5.2 Method

5.2.1. Data and design

For these analyses, data were taken from participants in Experiments 2, 3 and 5. There were three variables in those experiments: Structure (half of all trials comma, half of all trials garden-path), Task (all participants in Experiments 2 and 3 answered questions after items; all in Experiment 5 did not), and Load (half of participants in Experiment 2 had no n-back task [no-load], half in Experiment 2 had a 2-back task, all of Experiment 3 had a 4-back task; all in Experiment 5 could also be described as no-load). The influence of these three variables was investigated using scanpath analyses. I also looked at memory span in scanpaths from Experiment 5, with participants split again into "low", "medium" or "high" span.

5.2.2. Data extraction and cleaning

Data were extracted using Data Viewer (SR Research) to produce a dataset with one row per fixation. Data were then processed using R. First, fixations were deleted if they fell outside the passage of text (12% of fixations). The y- coordinates of fixations were standardised to the midpoint of each line (i.e., fixations on the top line were standardised to the midpoint of that top line, and similarly fixations on the second line of text were standardised to the midpoint of line two). Finally, as sentences differed from one another in length (due to variation in word length across sentences) the x- coordinates were also standardised. These transformations are important as the scanpath analysis only take x- and y- coordinates of fixations into account. Without standardisation, a fixation at a given point of one sentence may correspond to the disambiguating verb, while a fixation at the same point on the screen may correspond to the spillover region in a different

sentence. Therefore, following von der Malsburg and Vasishth (2013), the size of all regions of interest for all sentences were calculated, and the x- coordinates of fixations were standardised to the layout of the first experimental item[37]. So, for example, if a region was longer in a given sentence than the standardised layout, fixations were adjusted slightly to move them to the left. As such, the location of fixations analysed here are marginally different from the locations in the raw data.

5.2.3. Scanpath extraction and analysis

Analysis was carried out using v.1.05 of the *scanpath* package for R (von der Malsburg & Vasishth, 2011, 2013). To assist with understanding the extraction process, sentence (5.1) is repeated here, with breaks between each region of interest:

(5.1) While the father/ changed(,)/ the baby that was cuddly/ played/ with its/ toys. The/ father had to finish/ changing his clothes/ to pick up and look after the baby.

For instance, the disambiguating verb is in region 4, and the spillover region is region 5. Relevant scanpath patterns were extracted using the *match.scanpath* function in the *scanpath* package. This function uses regular expressions, a sequence of characters that determine search criteria: for instance, (4+[1-3]) would return all fixations on region 4 that are followed by a fixation on any of regions 1 to 3. Regular expressions were used to extract scanpaths that fitted two patterns for the two analyses. The first pattern is referred to as *post-disambiguation scanpaths*: extracted scanpaths were those where a fixation was made on either the disambiguating verb OR the spillover region, with the endpoint of the scanpath being the first fixation on the final region (*to pick up…*). 79% of trials contained this pattern and were analysed. The second pattern will be referred to as *regression paths*:

[37] This item was chosen partly for simplicity, but also as its regions were of about average length.

extracted scanpaths were those where a fixation was made either on the disambiguating verb or the spillover region, followed by a regressive eye movement out of that region, with the endpoint of the scanpath being the first fixation after going past the spillover region (i.e., on *toys* or later). 49% of trials were found to have this pattern (incidentally, a similar proportion to the 52% found in von der Malsburg & Vasishth (2013), using similar inclusion criteria, but based on sentences with a different syntactic construction). I will now outline the additional steps carried out in the analysis. These additional analyses were carried out firstly on the *post-disambiguation scanpaths* data, and then on the *regression paths* data.

After extracting scanpaths, the *scasim* function was used to evaluate similarities between each pair of scanpaths that were to be analysed. As discussed above, *scasim* takes into account both the duration and location of fixations, and assesses similarity based on how much one fixation would have to be edited to translate it into the other (for full details, see von der Malsburg & Vasishth, 2011). This analysis used the standardised x- and y- coordinates of fixations, as described above. Similarity scores were then normalised to take into account scanpath length, by dividing scores by the total number of fixations (to eliminate trivial effects that may arise if, say, one scanpath has more fixations than another, even if those fixations follow the same pattern).

Similarity scores were then fitted to a multidimensional map of scanpaths with seven dimensions, using the isoMDS function in the R package *MASS* (Kruskal, 1964). Seven dimensions were chosen, similarly to von der Malsburg and Vasishth (2013). This map was then used as the basis for a cluster analysis, in order to identify qualitatively distinct patterns of scanpaths. Initially, cluster models were fitted for between 2 and 20 clusters, with a Bayesian information criterion used to decide the best fit for data. As discussed in the Results section, this produced a high number of clusters, with little difference (of

interest to this research) between several combinations of these clusters. The model was then re-fit to identify fewer, larger clusters – a smaller number of broader clusters provided an effective balance between separating qualitatively different patterns, and identifying a usable distinction between these patterns. For each cluster, a "prototype" scanpath was found: the scanpath closest to the centre of that cluster.

5.3. Results and Discussion

Results and a brief discussion for the two subsets of data will be presented separately, before bringing findings together in the Discussion section. Several analytic features were identical for the two subsets of data. After identifying the four larger clusters, binomial linear mixed effects models (Jaeger, 2008) were fitted for each cluster (this was the same procedure as used in von der Malsburg & Vasishth, 2013). For each, the dependent variable was whether a scanpath belonged to that cluster or not, and predictor variables were Structure, Task, Load, and the interactions of Structure-by-Task, and Structure-by-Load. Span Group was later added for the *post-disambiguation* scanpaths data, as outlined below. However, Load and Structure-by-Load did not have an effect on any of the models; to avoid overfitting models, they were therefore re-fit using only the other variables. It is these refitted models reported below, but the absence of Load effects is discussed in the Discussion.

5.3.1. Analysis of post-disambiguation scanpaths

The first analyses were on scanpaths starting at a fixation on either the disambiguating verb or the spillover region, and terminating at a fixation on the final region. The cluster analysis initially produced nine clusters, with prototype scanpaths displayed on Figure 5.1. Several of the scanpaths appeared to be reasonably similar, and so these were combined to make four larger clusters, displayed on Figure 5.2. These four scanpaths can be described (and defined) as follows. The first prototype, in the top left hand corner, shows a pattern of progressive eye movements, but with a regression out of the reflexive region (**checking**). The top right scanpath shows efficient, progressive eye movements that quickly move to the end of the passage (**rapid progression**). The bottom left pattern is very different: there is a regression out of the spillover region, to re-read the region containing the initial noun phrase (in (5.1), this is *the baby*), before a considerable amount of time is spent reaching the end of the passage (**selective reanalysis**). Furthermore, there is a regression out of the reflexive region (*changing his clothes*). The bottom right is a slower version of the *rapid progression* pattern, again without regressive eye movements (**standard progression**).

The proportion of scanpaths belonging to each of the four clusters were as follows: *checking* (18%); *rapid progression* (10%)*, selective reanalysis* (46%), and *standard progression* (26%). The *rapid progression* and *checking* patterns were therefore less common, and given the sparseness of data points, more caution should be taken in interpreting the results of these patterns.

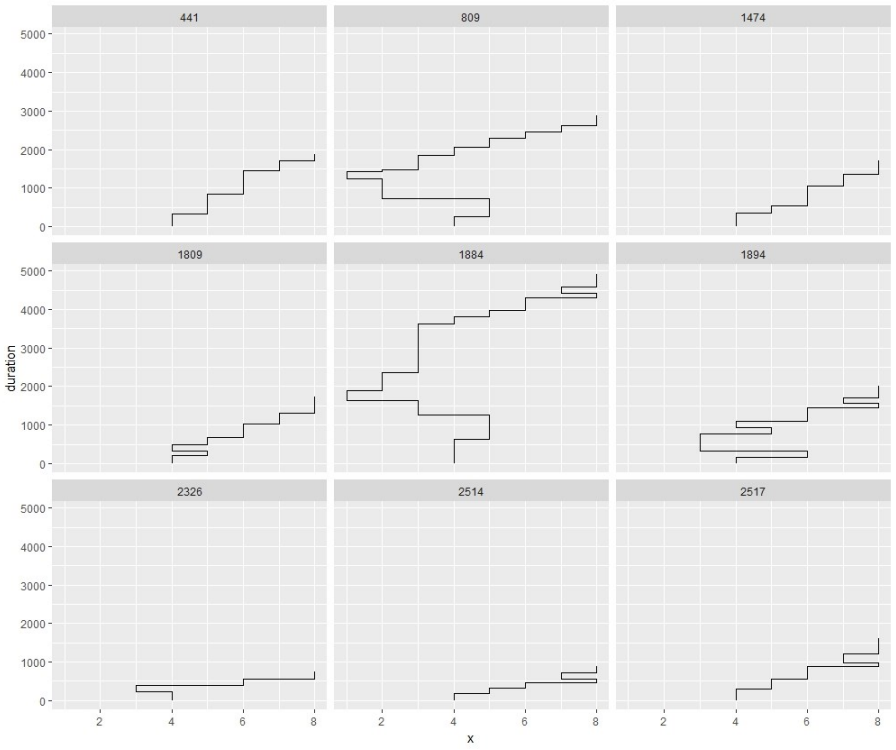

Figure 5.1. Prototype scanpaths for the original nine post-disambiguation clusters.

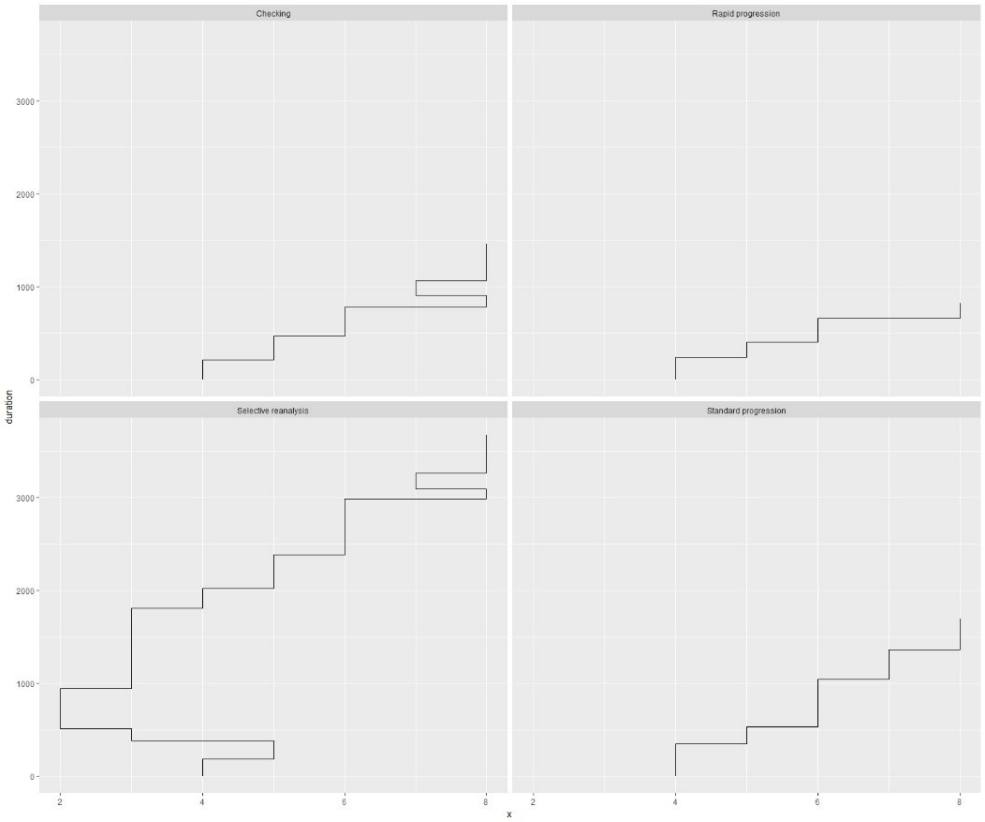

Figure 5.2. Prototype scanpaths for the four larger post-disambiguation scanpath clusters (top left, *checking*; top right, *rapid progression*; bottom left, *selective reanalysis*; bottom right, *standard progression*).

For *rapid progression*, the model found no effects of Structure or Task, nor any interaction, all $z < 1.4$. For *checking*, there were main effects of both Structure, $z = 5.68$, and of Task, $z = 4.10$; the interaction did not reach significance, $z = 1.78$. This pattern was more common for comma items than garden-path items, and for no-questions than for participants asked questions. The trend towards an interaction demonstrated a wider gap between the Structure conditions in the no-questions participants. For the *standard progression* pattern, there was a main effect of Structure, $z = 6.64$, and of Task, $z = 2.99$, qualified by an interaction between the two, $z = 2.53$. For participants asked questions, this pattern was more common for comma items than garden-path items; for participants without questions, there was a much smaller effect of Structure, and this pattern occurred more frequently for both garden-path and comma items. Finally, for the *selective reanalysis* pattern, there was a strong main effect of Structure, $z = 10.90$, and an effect of Task, $z = 4.29$, although the interaction between the two was not significant, $z < 1$. The pattern occurred more frequently on garden-path items, and in participants asked questions.

Since both the *standard progression* and *selective reanalysis* patterns elicited main effects of Task, one interesting question was whether there was an effect of memory span on the no-questions participants' scanpaths. In Experiment 5, high-span participants read passages for longer in critical areas than low- and medium-span participants. The scanpaths were therefore analysed again, but only for participants in the no-questions condition, and with main effects of Structure and Span Group ("low" vs. "medium" vs. "high"; see Chapter 3 for full details), and their interaction. Span Group was entered with contrasts firstly of "low" vs. "medium", and then "low" and medium" vs. "high". For the *standard progression* pattern, there were no main effects of span, $z < 1.5$, but Structure interacted with the difference between low/medium, and high-span readers, $z = 2.40$ (the

difference between low and medium span readers was not significant, $z = 1.67$). There was little effect of Structure on the proportion of trials showing this pattern in low- and medium-span participants; the pattern was on 30% and 31%, respectively, of garden-path item trials (cf. 40% and 28%, respectively, for comma items). In contrast, the *standard progression* pattern only appeared on 18% of high-span readers' scanpaths on garden-path items, even though it was on 35% of their comma item scanpaths.

For the *selective reanalysis* pattern, there were main effects of Span, both in the difference between low- and medium-span readers, $z = 2.19$, and between these two and high-span readers, $z = 2.23$. The interaction between the low/medium and high-span readers with Structure did not reach significance, $z = 1.79$. Looking at the means, the pattern was more apparent on garden-path trials than on comma trials. However, the proportion of garden-path trials with this pattern increased with memory span, from the low (38%), to medium (51%), to high (62%); there is less of a difference for comma items (18% vs. 30% vs. 28%).

To summarise, there were three main findings from these analyses. First, the omission of a comma not only affected reading durations on critical regions, but produced differences in the likelihood of making qualitatively distinct scanpaths. The selective reanalysis pattern was the most common overall, but it was especially prevalent when reading garden-path items compared to comma ones. To illustrate why this might be, the prototypical scanpath for the *selective reanalysis* cluster showed a return to the point where the ambiguity arose (e.g., *the baby*), and this regressive movement appeared to originate from the spillover region after the disambiguating verb. This tallies with the results presented in Chapters 2 and 3: on several occasions, there was not an effect of Structure on first pass regressions out of the disambiguating verb alone, but there was an effect on the longer disambiguation regions (which contained a spillover region after the

disambiguating verb). There is also evidence of a regression out of the reflexive area part of the sentence, consistent with Slattery et al.'s (2013) findings that misinterpretations lingered and disrupted eye movements in this region: the presence of the reflexive phrase comes as a surprise. However, the scanpath analyses suggested that this only occurs on a subset of trials, perhaps explaining the absence of any effects in the conventional methods in the experiments presented in this thesis.

A second finding emerging from the scanpath analyses was the increased prevalence of the *standard progression* pattern for participants not asked questions. This pattern is consistent with fairly superficial reading of the material: the prototypical scanpath contained no regressions, and the total time spent reading the text was fairly short. This is perhaps to be expected in comma sentences, given their lack of ambiguity – and indeed, there was a main effect of Structure in this direction. However, the interaction with Task reflected the fact that participants not expecting questions (from Experiment 5) showed this pattern more frequently overall, and were almost as likely to show it on garden-path items than on comma items. This supports the conclusion from Experiment 3 that not expecting questions led to a greater preponderance for more superficial reading – even when reading syntactically ambiguous items.

Finally, there were some interesting interactions with memory span. The more superficial *standard progression* pattern was less common on garden-path items in high span participants, indicating that they were less likely to simply continue reading upon encountering a syntactic ambiguity. There was instead an increase in the *selective reanalysis* pattern of returning to the site of the original ambiguity to begin reanalysis – although the interaction between this and Structure did not reach significance.

These analyses took into account all scanpaths – but there may be particular interest in looking at what happens on the approximately half of trials where a regression was made out of the disambiguating region. The next section focuses on these trials.

5.3.2. Analysis of regression path scanpaths

The regression path scanpaths were those starting at a fixation on either the disambiguating verb or the spillover region, followed by a regression to an earlier region, and terminating at a fixation on any region after the spillover region. The cluster analysis initially produced 13 clusters, with prototype scanpaths displayed on Figure 5.3. As before, there were several examples of scanpaths that closely resembled each other, and so these were combined to make four larger clusters, displayed on Figure 5.4. These four scanpaths can be described (and defined) as follows. The top left scanpath shows a progressive eye movement to the spillover region, before regressing from there (**checking**). The top right is a regression out of the disambiguating verb, followed by a return to the disambiguating verb, and then on to the spillover region (**checking and returning**). The bottom left pattern is more brief, with a quick check of the region before the disambiguating verb before proceeding (**rapid regression**); furthermore, there is a regression out of the reflexive region (*changing his clothes*). The final prototype, in the bottom right hand corner, is considerably longer, with several regressions and more detailed re-reading of earlier text (**re-reading**).

The proportion of scanpaths belonging to each cluster were as follows: *checking* (12%)*, checking and returning* (40%)*, rapid regression* (15%) and *re-reading* (33%). Again, caution should be expressed in interpreting results for the *checking* and *rapid regression* patterns, due to the small number of data points. There were no significant effects on *checking*, all $z < 1$, and this pattern was rare in all conditions. There were main

effects of Structure on both the *checking and returning,* $z = 5.06$, and *rapid regression* patterns, $z = 4.08$, but no effects of Task. There was a main effect of Structure on the *re-reading* pattern, $z = 8.30$, and a marginal effect of Task, $z = 1.90$; the interaction was not significant, $z < 1$.

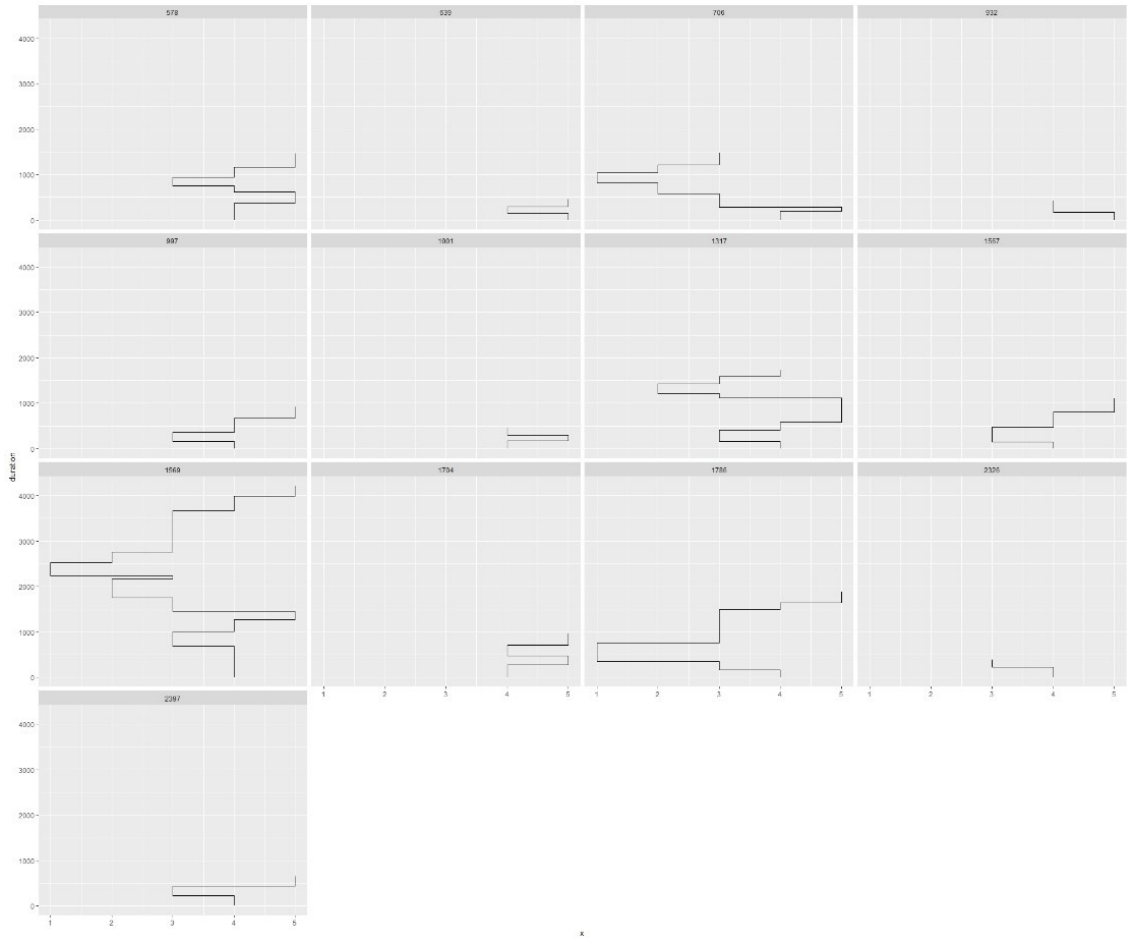

Figure 5.3. Prototype scanpaths for the thirteen clusters from regression path scanpaths.

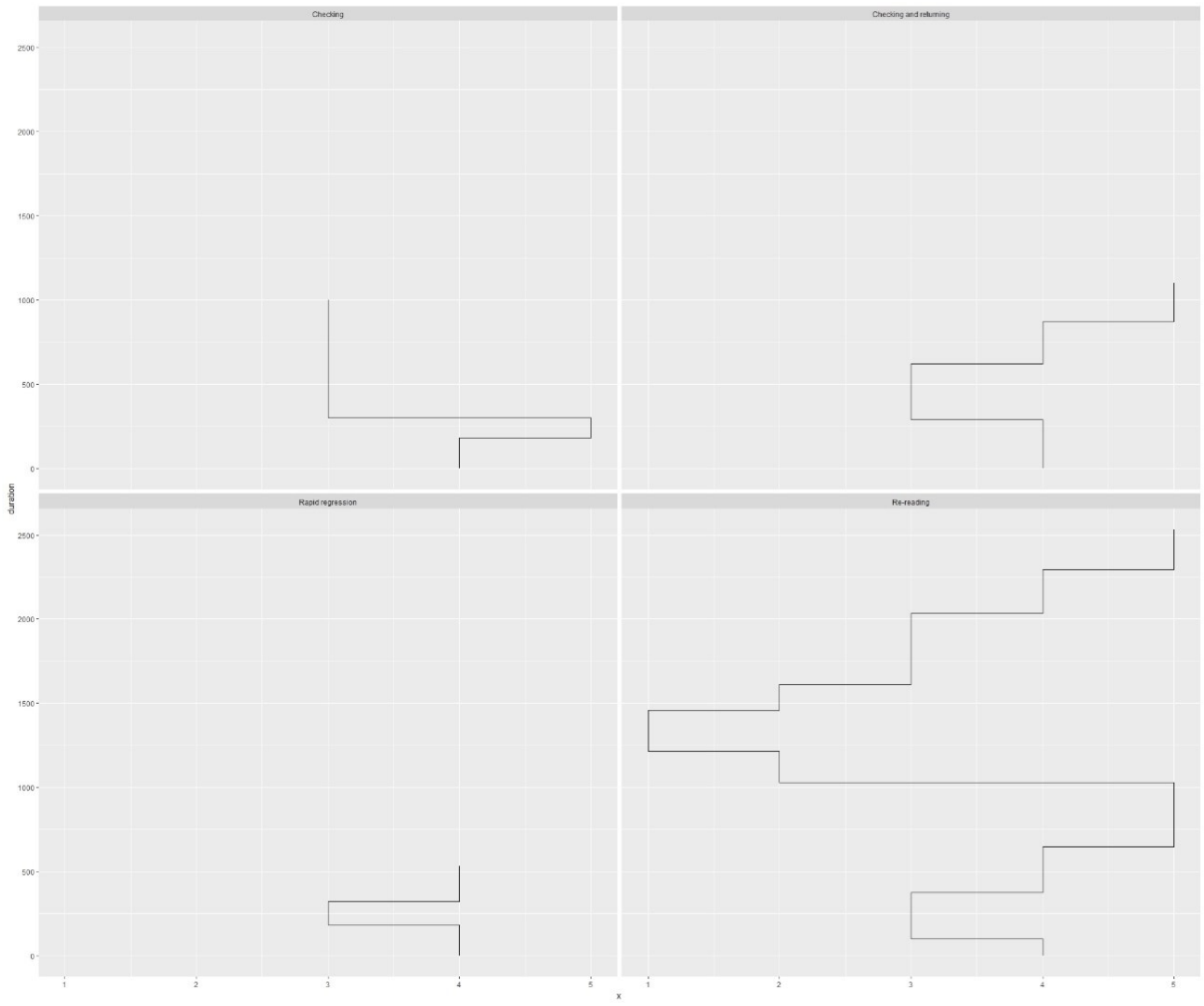

Figure 5.4. Prototype scanpaths for the four larger clusters from regression path scanpaths (top left, *checking*; top right, *checking and returning*; bottom left, *rapid regression*, and bottom right, *re-reading*.

The main effects of Structure reflected the fact that comma trials saw more *checking and returning* (comma 47%, garden-path, 34%) and *rapid regression* patterns (comma 20%, garden-path, 11%), while *re-reading* was more common on garden-path trials (comma 21%, garden-path, 43%). The marginal effect of Task represented the greater prevalence of re-reading in participants presented with questions (comma 23%, garden-path, 47%), compared to participants who were not (comma 18%, garden-path, 38%).

This second analysis showed similar findings to von der Malsburg and Vasishth (2013), following their analysis of a different sentence construction. Participants regressed in a number of qualitatively different ways, from extensive re-reading to a simple, rapid regression just to the left of the disambiguating verb. While this cannot explain *why* they make these regressions, the more extensive regression paths certainly point towards more detailed reanalysis. The analyses demonstrated that the most extensive *re-reading* pattern was more common with syntactically ambiguous items, with the prototypical scanpath showing a regression back to the site of the ambiguity (cf. Frazier & Rayner, 1982; Meseguer et al., 2002). Regression paths for comma items tended to be more rapid checks of what had been read, or they might have represented a pause while processing recent text (Mitchell et al., 2008). The marginal effect of Task on the re-reading pattern adds further credence to the conclusion that the removal of questions led to more superficial reading, with less extensive re-reading seen in those participants expecting questions.

Finally, it is worth adding that memory span was not analysed for this smaller subset, primarily because Task was not found to be a clear predictor of scanpath patterns. It would therefore be more difficult to interpret the results in terms of high working memory counteracting the superficial reading pattern that comes with not expecting comprehension questions.

5.4. Discussion

Two sets of exploratory scanpath analyses were carried out on data from Experiments 2, 3 and 5. The aim was to consider the extent to which load (via a dual task), task demands (the presence or absence of questions), and syntactic ambiguity

affected entire scanpaths – and to consider individual differences in scanpaths even within each group. There were no effects of Load on any measure, across all analyses. This is consistent with the results presented in Chapter 2: Load did not appear to influence eye movements in any clear way in Experiments 2 and 3, despite the effects seen in Experiment 1. Indeed, inspection of cluster proportions by Load demonstrates little effect, other than a slight fall in the *selective reanalysis* scanpath in the 4-back condition. This will be discussed more in Chapter 6, but for now, these analyses offered no evidence of Load affecting eye movements – even if the 4-back condition did adversely affect comprehension, as shown clearly in Experiment 3.

In contrast to the lack of effect of Load, Structure significantly affected scanpath patterns, with a shift towards longer scanpaths containing more regressions (the *selective reanalysis* pattern); furthermore, when regressions were made, they were more likely to be targeted to the site of the ambiguity, and to be relatively thorough (rather than the more rapid regressions seen in comma items). This would support the effects of Structure seen in go-past durations across the experiments in this thesis. The scanpath analyses do not answer the question of why readers sometimes spend longer on their first pass and sometimes make a speedy regression to earlier parts of the text. However, they did demonstrate that for these items, an efficient and well-controlled regression to the site of the ambiguity was a common pattern. Importantly, the pattern is *common*, but not ubiquitous – further questioning fixed-choice theories such as the garden-path model (cf. Meseguer et al., 2002; van Gompel et al., 2000, 2001; von der Malsburg & Vasishth, 2013). Instead, reading patterns after encountering a syntactic ambiguity are variable – while they often involve a regression, they sometimes involve spending longer in the disambiguating region, and sometimes rapidly progressing to later text.

Two factors appear to affect the decision of what to do upon detecting a syntactic ambiguity: task demands, and individual differences in memory span. The first set of analyses demonstrated that when readers were not expecting questions, they showed more superficial eye movement patterns, with fewer regressions, and shorter overall reading times. Furthermore, the aforementioned effect of Structure was attenuated in the no-questions group: there was little difference between the chance of triggering selective reanalysis on a comma item and on a garden-path item. Interestingly, this superficial reading pattern was considerably more apparent in lower-span readers, mirroring the results of von der Malsburg and Vasishth (2013). The second analysis found a marginal effect of Task – demonstrating that when regressions were made, they were less likely to be the extensive re-reads of the text that were common among participants expecting questions. These findings support the conclusions reached in Experiment 5: if a reader expects their comprehension to be tested, they read (and importantly, re-read) a text more carefully. If they simply read, they are more likely to continue through the text without regressing or stopping at the disambiguation point.

Three implications of these additional analyses are worth discussing. First, the results support the idea that reading behaviour, and particularly eye movement behaviour, is adaptive to task demands (Christianson, 2016; Ferreira & Patson, 2007; Lewis et al., 2013; van den Broek et al., 2001). In Chapter 3, it was asserted that the absence of questions led to more superficial reading, but this was based on reading durations on one region of the text. Task effects (questions vs. no questions) were seen here on wider scanpaths, providing further support for this view. This result is consistent with the good-enough approach to sentence processing: extensive reanalysis is conducted if this is necessary, but as doing so is resource-intensive, it is less likely to occur if the task demands do not require this (Ferreira & Patson, 2007; Karimi & Ferreira, 2016; Swets et

al., 2008). Of course, this is also consistent with alternative accounts (e.g., van den Broek et al., 2001), and this will be discussed more in Chapter 6. However, one particularly relevant finding here for the good-enough account was that the selective reanalysis pattern appeared to include a short regression out of the reflexive term in the second sentence. This demonstrates that readers had difficulty processing this phrase (see Slattery et al., 2013), supporting the idea that the initial misinterpretation from the first sentence lingers and disrupts later reading. Taken together, these results support adaptive syntactic processing that does not process text veridically under all circumstances.

A second point relates to individual differences. In line with von der Malsburg and Vasishth (2013), scanpaths varied according to memory span – with high-span readers less likely to show the most superficial scanpath patterns. This supports von der Malsburg and Vasishth's (2013) argument that high-span readers are less likely to leave syntactic interpretations underspecified, since underspecification is an adaptive response to maintain resources (see also Ferreira & Patson, 2007). The high-span readers are therefore more likely to continue and fully process the sentence, even if the task demands do not require it. Their reward was seen in Experiment 5 with better recall even after a 10-minute delay.

Finally, more broadly, the results demonstrate that the syntactic parser does not always do the same thing (von der Malsburg & Vasishth, 2013). There are individual differences in scanpaths, both within and between people, and even taking into account the predictor variables explored here. This challenges theories such as the garden-path model, which predict that selective reanalysis will always be the strategy followed by the parser. At first, the cluster analysis found 9 patterns of overall scanpath, and 13 patterns of reanalysis. These were condensed to four for each subset, mostly in the interests of

parsimony, and to avoid overfitting the data to an intractably large number of similar clusters. However, the exact number of clusters is less important than the overriding observation that scanpaths can be qualitatively different when encountering a syntactic ambiguity. This warrants further investigation to establish exactly why.

In conclusion, the additional analyses presented here found that syntactic ambiguity, task demands, and individual differences in memory span (although not extrinsic load) affected scanpaths when reading. There were notable differences in both scanpaths overall, and the patterns of regressive eye movements made after encountering a syntactic ambiguity. The results provide further support for von der Malsburg and Vasishth's (2011, 2013) scanpath analysis methodology as a powerful and sensitive addition to conventional eye movement measures. As von der Malsburg and Vasishth (2013) showed, this can inform theoretical models of syntactic ambiguity resolution, by considering more data than simply reading durations on critical regions. It is worth ending by stressing again a point made in the introduction to this chapter. These analyses were exploratory, and the data were not designed specifically with these analyses in mind. The analyses rely on the same datasets as in Chapters 2 and 3, and it would be important to test these hypotheses on a new dataset. As such, these analyses demonstrate the importance of further, more targeted research to explore the effects of load, task demands and individual differences on entire scanpaths. Nevertheless, the scanpath analyses provide an insight into individual differences in eye movements beyond conventional measures alone – and provide tentative support for the idea that the variables explored in this thesis affect syntactic ambiguity resolution, and should be incorporated into models that aim to explain this process.

Chapter 6

General Discussion

The aims of this thesis were to explore the processing of syntactically ambiguous sentences, what it means for processing to be *good enough* for the task at hand, and to explore the effects of task demands and individual differences on eye movements and comprehension. In this General Discussion, I will summarise the results of the six experiments outlined in Chapters 2 to 4 (including the additional analyses in Chapter 5), and outline the main conclusions of this research. These results will then be recast against the literature and research questions set out in Chapter 1, looking at the theoretical and methodological implications for sentence processing, and specifically the link between eye movements during reading and comprehension of what has been read. After highlighting some limitations, I will then discuss future directions for this area of research based on the results discussed here.

6.1. Summary of experimental findings

Six experiments have been reported in this thesis. All six have measured both eye movements while reading sentences containing a temporary syntactic ambiguity, and also comprehension of the sentences after having read them. Chapters 2 to 4 each focused on the effects of one main factor on both eye movements and comprehension. The eventual goal was to understand whether syntactic processing is merely *good enough* for the current task demands (Christianson, 2016; Christianson et al., 2001, 2006; Ferreira & Patson, 2007), and so whether altering those task demands would change the extent to which sentences were processed fully and retained veridically. In Chapter 2, the focus

was on the effects of load (via a concurrent n-back task). Experiment 1 introduced the basic paradigm, with participants reading single garden-path sentences, such as (6.1) or (6.2); the first manipulation here was of syntactic ambiguity (if the comma was omitted). Building on Christianson et al.'s (2001, 2006) work, each sentence was followed with a comprehension question such as (6.1a) or (6.2a), designed to tap lingering representations of initial misinterpretations (that the man hunted the deer, or that the father was changing the baby).

(6.1) While the man hunted(,) the deer ran into the woods.

(6.1a) Did the man hunt the deer?

(6.2) While the father changed(,) the baby played that was cuddly played with its toys.

(6.2a) Did the father change the baby?

This reading task was combined with a between-subjects manipulation of load: half of the participants completed a concurrent n-back task. There is considerable debate about the extent to which initial syntactic processing is reliant on general cognitive resources (e.g., Caplan & Waters, 1999; Just & Carpenter, 1992; Macdonald & Christiansen, 2002), and this dual-task condition was used to explore how both online processing and comprehension would be affected by the concurrent task.

I predicted that no-load participants would show similar results to previous experiments, with longer reading durations for the garden-path sentences compared to sentences containing a disambiguating comma. Furthermore, following Christianson et al. (2001, 2006), comprehension was expected to be poorer for the garden-path sentences, with initial misinterpretations not having been fully extinguished. The addition of load was expected to accentuate these effects, with additional re-reading required to parse the sentence, but more comprehension errors as well. In practice, comprehension was not

affected by load: participants with the concurrent 2-back task did not perform any worse on comprehension questions, despite their similarly impressive performance on the n-back task. However, achieving this high rate of comprehension required additional re-reading time. The conclusion was that under load, participants were more likely to produce an underspecified syntactic representation on their first pass – a more superficial reading style to account for the resources allocated to the concurrent task. Upon detecting the syntactic ambiguity, the underspecified parse was insufficient, and more time was spent re-reading the sentences to attain comprehension.

Experiment 1 used sentences presented in isolation. While this is fairly standard in sentence processing experiments (cf. Clifton et al. 2007; Radach et al., 2008), it is less like general reading. It is possible therefore that the stimuli were too simplistic to detect any effect of load on comprehension, especially if the sentences were re-read to maintain adequate comprehension. Furthermore, using single sentences offers no insight into Slattery et al.'s (2013) findings that initial misinterpretations lingered into reading a second sentence. For instance, when presented with sentences such as (6.3), Slattery et al. found that omitting the comma in the first sentence led to longer reading durations on *finish changing his clothes* in the second sentence.

(6.3) While the father changed(,) the baby that was cuddly played with its toys. The father had to finish changing his clothes to pick up and look after the baby.

Slattery et al. concluded that the initial misinterpretation (that the father was changing the baby) had lingered; the idea that the father was getting changed *himself* therefore came as a surprise, leading to inflated reading times.

Experiments 2 and 3 therefore followed up on Experiment 1, maintaining the comma and dual-task manipulations, but presenting sentences in a short two-sentence

passage like (6.3). Experiment 2 again had half of the participants reading with no load, and half with the 2-back task. Experiment 3 used a more difficult, 4-back version of the n-back task – assessing whether it was the mere presence of load that could affect syntactic processing, or whether a more difficult concurrent task would produce more comprehension errors.

Performance on the concurrent n-back task was once more good in both Experiments 2 and 3. Comprehension was poorer in Experiment 2 than it had been in Experiment 1, suggesting that the shift to passages led to a greater tendency to maintain the initial misinterpretation. Like in Experiment 1, the 2-back task had no significant effect on comprehension accuracy in Experiment 2. With the 4-back task in Experiment 3, there was a sharp decline in comprehension specifically for garden-path items. In contrast, and unlike in Experiment 1, load did not notably affect eye movements. While this was not a within-item comparison, a striking difference was that there was no first-pass effect of Structure on the disambiguating verb in the passages used in Experiments 2 and 3 (see also Slattery et al., 2013). The lingering effects seen in Slattery et al. (2013) were not observed in either Experiment 2 or 3.

Chapter 3 moved to the question of how task demands affect syntactic ambiguity resolution, and good-enough effects in comprehension. First, Experiment 4 explored reading of longer, 4-sentence passages – assessing whether the effects seen in Slattery et al. (2013) would be seen with an intervening sentence between the two sentences presented in (6.3). There was also a manipulation of whether the intervening sentence biased towards the initial misinterpretation or not. The results were broadly similar to those seen in Experiment 2: comprehension was poorer than in Experiment 1 (but was not different from that in Experiment 2), and eye movement patterns were similar to the two-sentence passages in Experiment 2 (other than a small effect of neutral comma passages

being read more quickly and with fewer regressions). The manipulation of bias did have an effect on comprehension, with more evidence of good-enough effects when the intervening sentence perpetuated the initial misinterpretation (although the Structure-by-Bias interaction was not significant, and this was only observed in *post hoc* tests).

Since most research in this area (including Experiments 1 to 4) has tested comprehension immediately after presenting each experimental item, Experiment 5 looked at longer-term representations of syntactically ambiguous sentences. This was to see how representations changed over time, and to assess to what extent good-enough effects stemmed from the methodology of asking immediate questions. Participants were given a recognition test for paraphrases of items that had been seen 10 minutes before, after a reading span task during the delay. By using the same stimuli as in Experiments 2 and 3, Experiment 5 also explored how task demands (reading for comprehension vs. reading without any expectation of questions) affected eye movements during reading. The most interesting finding was that good-enough effects of lingering misinterpretations were found in the post-delay paraphrase recognition task. Participants were able to discriminate well between comma paraphrases and foils; for garden-path items, discriminability was poorer, and foils that tapped the initial misinterpretation (e.g., *The father was changing his baby*) were readily accepted as genuine paraphrases. Memory span moderated performance on the paraphrase task, with higher-span participants more able to discriminate the garden-path paraphrases accurately. On the question of task demands, participants in Experiment 5 read more superficially than those who had been expecting questions (with shorter reading durations and fewer regressions), although this was not specific to garden-path items. Once more, this was moderated by memory span, with high-span participants still reading for longer.

Experiment 6 investigated individual differences related to ageing, using the same experimental design as Experiment 1, but with older adults as a comparison to the younger adults used in the first five experiments. This was based on previous work demonstrating differences in eye movements and comprehension with ageing (e.g., Christianson et al., 2006; Kemper et al., 2004; Kliegl et al., 2004; Rayner et al., 2006; Stine-Morrow et al., 2003; von der Malsburg et al., 2015), but building on the fact that little work has compared eye movements and good-enough comprehension effects simultaneously. Older adults performed worse on the concurrent n-back task, but despite (or perhaps, because of) this, they were able to maintain good comprehension – with no difference from younger adults on items containing reflexive absolute transitive verbs (e.g., changed, dressed). They did however show a marked difference in both comprehension and eye movements as a function of verb type, with a greater tendency to answer yes to questions following optionally-transitive verbs (e.g., hunted, ordered). Verb type did interact with load: older adults in the 2-back condition were less influenced by the effects of verb type, suggesting a lesser reliance on text properties. There was also a notable similarity between eye movements in younger adults under a 2-back load, and no-load older adults – although again, this was limited to reflexive absolute transitive items.

Finally, in Chapter 5, data from Experiments 2, 3 and 5 were reanalysed, to assess entire scanpaths rather than just reading durations on critical parts of the text. Previous work has used this approach to investigate attachment ambiguities (von der Malsburg & Vasishth, 2011, 2013); my analyses extended this work to investigate subject-object ambiguities such as (6.1) – (6.3). There is good reason to expect differences in how these two types of ambiguity are processed (as discussed in Chapter 1, and again below). These additional exploratory analyses went beyond results from conventional methods. Reading syntactically ambiguous sentences produced a qualititative shift in scanpaths, compared

to sentences containing a disambiguating comma. There was support for a pattern of selective reanalysis (Frazier & Rayner, 1982; Meseguer et al., 2002), with eye movements returning to the point of ambiguity in order to begin reanalysis. However, this pattern was not ubiquitous, demonstrating variability in responses to syntactic ambiguity. Furthermore, several interactions were found with the task participants were completing – namely, to answer questions in Experiments 2 and 3, or simply to read in Experiment 5. Not expecting questions was associated with a greater proportion of more superficial scanpaths, with fewer and shorter regressions, and less of an increase in re-reading for garden-path passages compared to comma ones. Finally, there was some evidence to suggest that higher-span participants were more immune to this superficiality, maintaining thorough regressive paths even if not expecting questions.

6.2. Overall themes and conclusions

In this section, I will consider the main conclusions across these experiments, and the overall themes that stem from these. These will be considered under the headings of the four questions posed towards the end of Chapter 1.

6.2.1. What is meant by *good-enough* syntactic ambiguity processing?

Given the similarities to previous work by Ferreira, Christianson and colleagues (e.g., Christianson et al., 2001, 2006; Patson et al., 2009; Slattery et al., 2013), this thesis is grounded in the good-enough account of sentence processing. There are several other theories with similar predictions, and the relevance of the results to these theories are discussed in section 6.3 below. Sticking with the good-enough account for now, one issue with the theory that syntactic processing is often only *good enough* is that there is a lack

of clarity about what *good enough* exactly means (Christianson, 2016; Christianson et al., 2001; Ferreira & Christianson, 2016; Ferreira & Patson, 2007; Karimi & Ferreira, 2016; Slattery et al., 2013; Swets et al., 2008; Traxler, 2014; von der Malsburg & Vasishth, 2013).

Arguably, the most promising explanation is a combination of the accounts provided by Slattery et al. (2013) and Karimi and Ferreira (2016). Slattery et al. (2013) suggested that the process of pruning abandoned interpretations is often not completed, causing initial misinterpretations to linger. These misinterpretations can therefore disrupt comprehension, leading to poor comprehension accuracy – even if a garden-path effect of longer reading times has been observed. This account would also explain findings such as Patson et al.'s (2009), who found that when asked to paraphrase garden-path sentences, participants would often produce blurred interpretations that combined elements of the initial misinterpretation with the correct interpretation; the initial misinterpretation lingers and leads to a non-veridical representation of what has been read. Karimi and Ferreira (2016) offered an insight into the processes behind good-enough comprehension, suggesting that the comprehension system aims to reach a consistent interpretation ("cognitive equilibrium"), with both heuristic and algorithmic processing combining to reach this interpretation. Good-enough comprehension errors can occur if the interim output from the heuristic processing is not sufficiently refined by (or, abandoned after) more detailed, algorithmic processing, or if the algorithmic processing is not sufficient to overcome this.

Of relevance to the results of this thesis is the emphasis placed on both individual differences such as in working memory, and task demands (e.g., Ferreira & Patson, 2007; Karimi & Ferreira, 2016; Swets et al., 2008). In several expositions of the good-enough account, it is suggested that a good-enough representation will be relied upon more

readily if the task demands do not require a strictly veridical representation, if cognitive demands (or load) is high, or in individuals without the resources for extensive reanalysis. The results in this thesis provide support for this fundamental prediction, and demonstrate the importance of considering these factors in models of syntactic processing. Eye movements on critical parts of texts, scanpaths and comprehension were affected by: load (using a concurrent task of varying difficulty), task demands (whether or not questions were expected), and individual differences (both in memory span, and more indirectly, with ageing). While individual differences with memory, and to a lesser extent with ageing have been considered before, there is less work looking at the effect of task demands – and few examples of these being incorporated into models of eye movements (but cf. Logačev & Vasishth, 2016a). Together, these findings demonstrate an eye movement and comprehension system that is adaptive to changes in cognitive resources and to the demands of the current task, and that processes text more superficially if this is a viable or necessary option during the current task.

Three examples are representative of this adaptive system. First, in Experiment 1, participants under load adapted to the added demands by reading marginally more superficially on their first pass of the disambiguating region of garden-path sentences. However, as leaving these sentences underspecified is not a viable option (since it produces an incomplete parse), there was more re-reading in order to maintain comprehension. In Experiment 3, where the concurrent task was more difficult, additional re-reading was less viable while simultaneously performing well on the n-back task. No increase in re-reading was thus seen, but comprehension was adversely affected, with a sharp increase in good-enough comprehension errors. In other words, misinterpretations were not pruned efficiently (due to the allocation of resources to the secondary task), and lingered to disrupt comprehension when prompted by the question.

Second, the results of Experiment 5 demonstrated that reading is more superficial when it takes place without the expectation of comprehension questions. As the task does not require extensive re-reading, resources are not wasted on doing this. At the point of the paraphrase verification task 10 minutes later, lingering misinterpretations that were initially built when reading the first sentence were apparent: participants were willing to accept that e.g., the father was changing his baby, despite this not being presented in any item. This was dependent on individual differences in memory load: high-span readers spent longer reading these sentences, and consequently, were better able to discriminate genuine paraphrases from foils.

Finally, the age-related differences described in Experiment 6 also support an adaptive system. Older readers were able to maintain their comprehension (at least, of reflexive absolute transitive items), but did so by prioritising comprehension accuracy over the concurrent n-back task. Older adults spent longer reading and re-reading garden-path sentences, and were rewarded by equivalent comprehension to younger adults. However, they also demonstrated evidence of good-enough effects based on heuristic processing: they were more likely to answer Yes to questions following optionally-transitive items, despite sentences (e.g., *While the man hunted...*) not explicitly warranting a Yes response. Heuristic processing is a less viable option when reading reflexive absolute transitive items (where reanalysis is required after detecting the syntactic ambiguity). As a result, older adults in Experiment 6 showed significantly more re-reading for sentences containing RAT verbs, relative to OPT verbs, indicative of more methodical, algorithmic processing.

6.2.2. What is the effect of task demands on syntactic ambiguity processing?

As discussed earlier, findings from both analyses of conventional eye movement measures (Chapter 3) and scanpaths (Chapter 5) supported a tendency for less detailed reading when questions were not expected (Kaakinen & Hyönä, 2010; Kaakinen et al., 2015; McConkie et al., 1973; Radach et al., 2008; Swets et al., 2008; Wotschack & Kliegl, 2013). The scanpath analyses provided a particularly interesting insight: even when regressions are made out of the disambiguating region, these are less likely to reflect thorough re-reading, and instead tended towards the more rapid regressions to briefly check what had been read previously. Previous sentence processing research (including eye-tracking studies of syntactic ambiguity resolution) has tended to ask simple questions on only a subset of trials – and only to ensure that participants are concentrating on the task, rather than to ascertain the final product of their comprehension. The research presented here adds to previous work demonstrating that the presence of questions affects eye movement patterns, and the decision about whether to include questions (and what questions to include, cf. Swets et al., 2008; Wotschack & Kliegl, 2013) should be carefully considered.

The length of text presented to participants also matters. A first-pass effect of syntactic ambiguity on the disambiguating verb of subject-object ambiguities is well replicated in isolated sentences (Frazier & Rayner, 1982; cf. Clifton et al., 2007). This was not seen in my experiments when garden-path sentences were presented in short passages, replicating a similar finding (which was not commented on) in Slattery et al. (2013). It is possible that the effect was simply underpowered, and some experiments did demonstrate a small, if non-significant, effect in the right direction. This was most notable in Experiment 5, which may raise the question of whether an efficient regression, rather than longer first pass reading, is an adaptive response to the predictable questions in

Experiments 2 to 4. A carefully-controlled comparison of single sentences and different lengths of passages is required, with fewer extraneous variables (i.e., without the load and questions manipulations, which were the focus of the experiments presented here). Nevertheless, these findings support the view that eye movement control is responsive to differing task demands, warranting more research in what is a fairly understudied area.

6.2.3. How do individual differences affect syntactic ambiguity processing?

The scanpath analyses confirmed what may be an obvious finding – that different people (and even the same people, on different occasions) read different sentences in variable ways. There is no one set response on encountering a syntactic ambiguity: on some occasions, readers regress to an earlier part of the text to begin reanalysis, on others they will progress on the expectation of disambiguating in time (or possibly by simply underspecifying), and on others they will show inflated first pass durations. As von der Malsburg and Vasishth (2011, 2013) showed, they may also return and re-read the entire sentence, starting the process again as a less costly alternative to spending time reanalysing the material (see also Lewis, 1998). While this may seem obvious, it contradicts fixed-choice theories of sentence processing, such as the garden-path model (Frazier & Fodor, 1978; Frazier & Rayner, 1982), which predict that a given sentence structure will always be processed initially according to set rules. Even in reanalysis, it is predicted that the parser will return to the site of the ambiguity (Frazier & Rayner, 1982) – a strategy that was common in these experiments, but by no means universal.

Several more systematic sources of individual differences were highlighted by these results. A link between sentence processing and working memory is well supported, even if there is disagreement about how working memory is conceptualised, and about

whether working memory resources affect *initial* sentence processing (Caplan & Waters, 2013; Gordon & Lowder, 2012; van Dyke & Johns, 2012). In Experiment 5, individual differences in memory span predicted accuracy on the paraphrase verification task, with high-span participants more able to distinguish genuine and foil paraphrases. There are two main potential explanations for this finding. First, higher-span participants may simply have stronger representations of the propositions that were expressed in the passages, and are thus able to retrieve these representations more effectively after the 10-minute gap. In support of this view, there was no interaction between memory span group and Structure, suggesting that the beneficial properties of memory span were not specific to recall of garden-path items. This result could be seen as compatible with cue-based retrieval accounts of memory (van Dyke & Lewis, 2003): "better" memory is equivalent to more efficient use of cues to accurate retrieve the correct representation, and inhibit distractors. Alternatively, high-span readers may have been more effective at abandoning misinterpretations while reading the sentences, meaning that the longer-term representations were more veridical to what was written. This view is supported by longer reading durations in high-span readers: more time was spent reprocessing the passages, leading to better accuracy (see also the longer reading durations by readers under load in Experiment 1, which left comprehension unimpaired by the dual task). These two accounts are, of course, not incompatible – and both may work together to inhibit good-enough lingering effects more effectively in higher-span readers.

The effects of ageing in Experiment 6 also provided an insight into the locus of increased good-enough comprehension errors in older adults. By comparing eye movements and comprehension, links could be identified between the absence of a garden-path effect in re-reading measures for OPT items in older readers, and their poorer comprehension performance on these items. The latter effect was seen in Christianson et

al. (2006), who suggested that older adults were more likely to rely on a heuristic and not fully reanalyse OPT items. Evidence from eye movements in Experiment 6 supported this: older adults are less likely to reanalyse OPT garden-path sentences, having not identified any inconsistency between e.g., the man hunting [SOMETHING] and the deer running into the woods. While the results of Experiment 6 cannot definitively explain why, the fact that older adults are more effective at doing this for RAT items, combined with Christianson et al.'s (2006) finding of a link to working memory capacity, suggests that an increase in heuristic processing is an adaptive response to working memory limitations with ageing. This view is also supported by age differences in performance in the dual-task conditions: the difference between eye movements on the two verb types is less apparent, suggesting a greater reliance on heuristic processing, rather than focusing on the detailed syntactic and semantic structure of what is actually being read. With the 2-back task, the older adults even re-read OPT items for longer – but this appears to come at the cost of poor performance on the concurrent n-back task.

6.2.4. What are the links between eye movements, comprehension, and domain-general resources?

The links between eye movements and comprehension were not entirely clear, and the effects of working memory on both were found to be distinct in Experiment 5. For instance, there was little evidence in Experiment 5 for a clear link between reading durations and eventual performance on the paraphrase verification task. Nevertheless, in Experiment 6, the results from eye movements supported the conclusions reached by Christianson et al. (2006) based on comprehension accuracy, reading times and individual difference measures. Furthermore, the n-back task in Chapter 2 led to changes in both eye

movement measures and in comprehension (albeit not in the same experiment). These experiments did not explicitly set out to compare different theories of the role of working memory in sentence processing; nonetheless, it remains apparent that both eye movements during reading and comprehension are related to individual differences in domain-general resources, even if the picture remains unclear.

6.3. Theoretical implications

6.3.1. Implications for the good-enough approach to sentence processing

The results presented here are consistent with several versions of the good-enough approach proposed by Ferreira and colleagues (Christianson et al., 2001; Ferreira & Patson, 2007; Karimi & Ferreira, 2016; Slattery et al., 2013). First, this thesis found evidence of good-enough effects using a combination of methods: eye movements, standard comprehension questions, and a post-delay paraphrase verification task. This provides further support for the view that good-enough effects as seen in Christianson et al. (2001) are not simply an artefact of asking questions tapping the misinterpretation (cf. Christianson, 2016; Kaschak & Glenberg, 2004; Patson et al., 2009; van Gompel et al., 2006). Across experiments, across several manipulations, and importantly across measures, the results were broadly consistent: misinterpretations linger and disrupt comprehension, even after a significant delay. This is then associated with less detailed reprocessing of text under certain task conditions, or individual differences in the detail of reprocessing as a result of individual differences in memory span.

The results of Experiment 3 take the good-enough approach forward, as a high external load led to a stark increase in good-enough errors following garden-path sentences. Participants continued to perform well on the concurrent n-back task, but this

came at a cost of more readily accepting questions that tapped initial misinterpretations. Importantly, general comprehension was not impaired: performance on the comma items was broadly equivalent to Experiments 1 and 2. This result stresses the importance of considering how resource demands can impact on the likelihood of eventual comprehension being less veridical to what has been read.

6.3.2. Links to theories other than the good-enough approach

It has been stressed throughout this thesis that the good-enough approach is not the only theory that predicts non-veridical sentence processing, especially in light of syntactic ambiguity and/or differing task demands (Christianson, 2016; Traxler, 2014). The links between the results of this thesis and three of those theories will be briefly considered.

6.3.2.1. Noisy-channel theories. Noisy-channel theories (e.g., Gibson et al., 2013; Levy, 2008; Levy et al., 2009) suggest that representations may be degraded by "noise", from processing demands, prior expectations and distractors. Misinterpretations can occur if representations are edited to fit with Bayesian priors, and if this editing process disrupts the faithfulness of representations to what was actually written. This account is clearly compatible with the effect of load – the 4-back task produces additional noise, affecting the fidelity of representations of garden-path items. Furthermore, this account can explain the difference between sentences and passages in terms of comprehension: a longer text creates additional noise, and disrupts the quality of comprehension.

As Christianson (2016) notes, it is less clear how these theories explain task effects and demands, as these should not affect the amount of "noise". However, my results could be compatible with noisy-channel theories if the task demands affect the

priority attached to the reading task, and hence the extent to which distractors are prioritised.

6.3.2.2. Now-or-never bottleneck. Christiansen and Chater (2016) suggested that sentences are processed as quickly as possible in line with predictions, and if an initial parse is not quite correct, it is difficult to reanalyse. Representations will therefore be less detailed if cognitive demands prevent a complete parse from being produced on the first attempt. Again, this is compatible with individual differences (more memory capacity leads to more efficient initial parsing). It is less clear how the results of Experiment 1 can be reconciled: the strategic decision to re-read in order to improve comprehension is incompatible with the idea that cognitive demands lead to less detailed processing. Ferreira and Christianson (2016) argued that the now-or-never bottleneck is fundamentally similar to the good-enough approach, albeit with less focus on the output of comprehension (i.e., good-enough comprehension errors), and with less detail on task effects.

6.3.2.3. Standards of coherence. van den Broek et al. (2001) argued that our reading behaviour is determined by a "standard of coherence": if we set a high standard, reading will be more detailed; with a lower standard, we will rely more on more superficial reading, and produce a more gist-like representation. This account is clearly compatible with the differences in eye movements in Experiment 5: when not presented with questions, a lower standard is set. This leads to more superficial reading, and in turn, the inability to discriminate between genuine and foil paraphrases in the verification task. Similarly, the effects of load in Experiment 1 could be explained as the setting of a high standard in the face of the concurrent task. On this view, more resources are allocated to reading (as well as to the n-back task), explaining the increase in re-reading. One question that remains unanswered is the differences with verb type: if sentences are being read at a

set (high) standard, why are optionally-transitive items read less carefully (in older adults), and why is comprehension on these items poorer, with a greater reliance on heuristics?

6.3.2.4. Outstanding questions. Clearly, there is significant overlap between the results of this thesis and theories that predict similar comprehension errors as a result of the manipulations used in these experiments. As noted by advocates of the good-enough approach, it is not entirely clear where these competing theories differ in their predictions (cf. Christianson, 2016; Ferreira & Christianson, 2016). Indeed, Christianson and Chater (2016) respond to Ferreira and Christianson's (2016) commentary by agreeing that the now-or-never bottleneck may be a framework to explain good-enough effects. Similarly, *standards of coherence* are not dissimilar from considering task demands, and how these determine the extent to which sentences are processed more carefully, or left to be merely *good enough*. It is promising to see recent and more detailed expositions of what it means for processing to be good enough (e.g., Karimi & Ferreira, 2016); my results demonstrate the importance of testing these theories on a range of stimuli, a range of participants and under a range of task demands.

6.3.3. Implications for sentence processing

The experiments presented in this thesis did not set out to directly compare theories of sentence processing, and arguably, focused more on reprocessing and reanalysis than on initial syntactic processing decisions. Nevertheless, it is worth noting again the lack of support for fixed-choice syntactic processing theories (e.g., Frazier & Fodor, 1978): the scanpath analyses in Chapter 5 demonstrated a range of eye movement patterns in response to encountering a syntactic ambiguity. Furthermore, while garden-

path effects were apparent in the proportion of regressions out across experiments, these differences were relative (i.e., there was still a non-negligible number of regressions out of comma items), and regressions tended to come from the spillover region as much as (if not more than) the disambiguating verb. This suggests that misinterpretations are sometimes built despite the presence of the comma; conversely, the high rate of comprehension in Experiments 1 and 2 suggests that the omission of the comma does not always cause a misinterpretation to be built. These results are more compatible with race-based models (e.g., Logačev & Vasishth, 2016b; van Gompel et al., 2000, 2001). Logačev and Vasishth's model has the advantage of also taking task demands into account (note it was designed to explain attachment ambiguities, rather than subject-object ambiguities). If the task does not require a full specification of syntax (or if, say, external demands are too high), this may lead to underspecification; underspecification is less of an option, though, if the initial parse of a subject-object ambiguity does not produce a viable parse. Modelling the implications for this syntactic ambiguity would be an interesting area for future work.

These points are also relevant to understanding eye movements during sentence processing. A greater understanding of the links between eye movement behaviour, task demands and individual differences is needed to explain syntactic processing. Experiment 1 demonstrated that extrinsic load (and the need to maintain comprehension despite this) led to an increase in re-reading. In contrast, not being asked questions in Experiment 5 produced more superficial patterns. There is also the question of why readers sometimes spend longer on their first pass, and sometimes quickly regress (cf. Clifton et al., 2007). These factors need to be considered concurrently, to explain their respective effects on eye movements and comprehension (see Engelmann, Vasishth, Engbert, & Kliegl, 2013 for an example of a model combining eye movement control and syntactic parsing).

6.3.4. Implications for understanding individual differences

The experiments here did not set out to examine competing theories of how to conceptualise the role of working memory in sentence processing. Nevertheless, they provide three main insights. First, by looking at how a dual task affects both eye movements and comprehension in the same set of experiments, the results in Chapter 2 explored the extent to which sentence processing was reliant on domain-general resources: theories such as Caplan and Waters' (1999, 2001) postulate that the dual-task should not affect what they call "interpretive", initial syntactic processes, even if they can affect "post-interpretive" processes in reanalysis and eventual comprehension. The effects of load on eye movements in Experiment 1 question this view: although the interaction between Structure and Load in first pass durations was not significant, there was tentative evidence that the addition of load led to more superficial reading on first pass. Even if this result were spurious, it is clear that load affects later processing, with more re-reading in Experiment 1, and a large drop in comprehension in the more difficult 4-back task in Experiment 3. Taken together with evidence of individual differences in Chapters 3 to 5, there is support for a role for working memory in both online processes and in determining eventual comprehension. This is most clear in reanalysis, and particularly in comprehension: Experiment 5 found less of a link between memory span and eye movement measures, despite finding a significant link between memory span and paraphrase recognition. This result concurs with previous research that has found limited evidence for a link between working memory measures and reading durations (e.g., Traxler et al., 2012).

Another interesting finding was the similarity between younger adults under load in Experiment 1, and the older adults in Experiment 6 in terms of eye movements. For reflexive absolute transitive items, the pattern of reading durations was almost identical.

This does not necessarily suggest that these occur for the same reason, but it does offer an opportunity for further exploration of whether a younger adult with good memory span who has these resources reduced by a concurrent task, behaves similarly to older adults in other settings. The result was not pervasive (for example, it was not seen in optionally-transitive items), suggesting that some aspects of age-related cognitive changes may be more strategic.

6.4. Methodological implications

There is a paucity of research comparing eye movements and comprehension concurrently, and especially across task demands and different groups of participants. By considering both, the experiments in this thesis were able to follow syntactic processing from the first pass of ambiguous and then disambiguating text, through to online reanalysis, and into lasting comprehension as tapped by questions. Furthermore, the introduction of the paraphrase verification task in Experiment 5 allowed for an analysis of comprehension after a significant delay, and an understanding of how interpretations of syntactically ambiguous texts change as they degrade over time.

Chapter 1 highlighted the importance of using eye movements (compared to, for instance, self-paced reading methodology), to review real-time online sentence processing. Eye-tracking offers the advantages of being able to break down early and later processing (cf. Clifton et al., 2007; Rayner, 1998), in a reasonably naturalistic way that does not burden working memory more than everyday reading. Using eye-tracking to consider syntactic ambiguity resolution is not novel, and work in this area dates back over 30 years (e.g., Frazier & Rayner, 1982). However, as highlighted by Christianson (2016), less work has considered the eventual product of comprehension by asking questions that

tap initial misinterpretations. When this has been done (e.g., Christianson et al., 2001, 2006; Patson et al., 2009), it has rarely been combined with eye movement measurement. This thesis offered the opportunity to see how eye movements and comprehension measures compare and contrast, especially in their responses to task demands and individual differences. The use of scanpath analyses was particularly interesting for looking at individual differences. At various points, similarities and differences have been identified between effects on eye movements and comprehension. Considering both is therefore imperative to prevent, for instance, drawing the premature conclusion that load had no effect on Experiments 1 and 3 (where only one of the two was affected each time).

As discussed, the use of a paraphrase verification task allowed an insight into longer-term lingering misinterpretations. Previous research has focused on assessing comprehension almost directly after each item, with comprehension questions (e.g., Christianson et al., 2001), paraphrase production (Patson et al., 2009), or syntactic priming (van Gompel et al., 2006). This approach has its benefits: it is easier to interpret results, as there is less chance of confounding interference between items that may affect comprehension. However, some research questions can only be answered with delayed testing, and the methodology used in Experiment 5 offers an interesting paradigm for use in a range of settings. For instance, how do longer-term representations of syntactically ambiguous items change over time? How do they respond to factors such as plausibility, or the length of the original material? Will the distinction between optionally-transitive and reflexive absolute transitive items hold in longer-term representations?

Another advance comes in the use of the n-back task to manipulate extrinsic load. While dual-task studies have been used before in the context of syntactic ambiguity resolution, these studies have focused either on eye movements *or* comprehension. The experiments in Chapter 2 offer an insight into the role of a concurrent n-back task on both

simultaneously. The n-back task is of particular use: it has been demonstrated that similar processes underpin both n-back performance and syntactic ambiguity resolution (Novick et al., 2014). There is potential for future work to consider a range of different n-back tasks, to explore what aspect of the task is interfering with syntactic processing – and what causes the "tipping point" of comprehension failure seen in Experiment 3. Of particular interest would be the level of overlap between the two tasks: my n-back task was designed not to specifically interfere with experimental items, other than by being a verbal task. Changing the nature of the task would help to explain the shared mechanism behind the n-back task and syntactic processing (and eventual good-enough errors).

6.5. Limitations and future directions

6.5.1. Methodological limitations

Several of these limitations were highlighted in the individual chapters, with some broader issues emerging. First, several cross-experiment comparisons were not based on within-item (or within-participant) designs, with minor differences in sample size; this may affect the reliability of these comparisons. In a similar vein, Experiments 1 and 6 had several differences to other experiments, having been run using different software, and with marginally different regions of interest, and a range of stimuli with different lengths and verb types. This does not negate conclusions reached within these experiments (or indeed, the comparison of the two) – but may question some of the comparisons with Experiments 2 to 4. However, the use of linear mixed effects modelling takes into account differences due to item effects, providing more confidence in the results demonstrated here. Nonetheless, the issues raised here preclude cleaner comparisons between identical stimuli across experiments.

A further question is about the representativeness of participants to the wider population. All participants were university students, primarily from the University of Oxford. Additionally, approximately a third of younger participants were psychology students, and a smaller percentage (c. 10%) studied languages and therefore potentially had experience in linguistics. The high performance on the comprehension questions certainly points towards a sample with high reading experience and high verbal ability, including working memory. While this may question the appropriateness of wider implications, this should not take away from the comparisons made between the manipulations in these experiments, or between participants in this experiment. Furthermore, given the difficulty of the n-back task (especially at the 4-back level), and of the paraphrase verification task, a high-performing sample was useful to identify meaningful differences and to avoid floor effects. On a side point, there was also no measure taken of participants' vision. All participants were however asked to confirm they either had normal vision, or that they wore appropriate glasses or lenses.

6.5.2. Theoretical limitations

Several limitations have already been discussed. For example, these experiments did not explicitly set out to compare theories of sentence processing, or of explanations of the role of working memory. As such, the insights offered in these areas are tentative – and a nuanced interpretation of the results could be compatible with more than one theory. Linked to this, it is not possible to say definitively that the effects of task load or ageing are a direct consequence of memory load. This is especially true for ageing, where a range of alternative explanations such as inhibition, speed of processing, and differences in self-regulated language processing could explain the age-related effects observed (see Chapter 4 for more discussion).

As discussed above, several alternative theories such as the noisy-channel approach or now-or-never bottleneck (Christiansen & Chater, 2016; Gibson et al., 2013), are not fully specified, or their distinctiveness from the good-enough theory is not clear. Nevertheless, the support offered for the good-enough approach by this thesis may similarly be consistent with these theories instead. The good-enough approach is however useful for considering these results, given the clear links between the work here, and research by Christianson, Ferreira and colleagues (e.g., Christianson et al., 2001, 2006; Slattery et al., 2013) – even if other theories may also be able to offer credible accounts of these results.

Finally, some results presented in the thesis may be seen as contradictory – for instance, the load effect seen in Experiment 1 on eye movements was absent in Experiments 2 and 3. Load clearly still impacted on results in Experiment 3, based on the results of comprehension accuracy. However, it is unclear why the load effects in Experiment 1 were not replicated. As discussed in Chapter 2, it may be that load has differential effects on the short sentences (where no-load participants do not need to engage in extensive re-reading) versus passages (where even no-load participants re-read, as evidenced by the scanpath data). It would be of use to explore these differences further. This relates to a wider question about strategies used by participants. For example, older adults performed poorly on the concurrent n-back task, with a reasonable proportion of participants being at chance. Manipulating instructions to participants (for instance, to prioritise the n-back task), or tailoring the difficulty of the concurrent task based on individual performance, could be an avenue for exploring these strategies more.

6.5.3. Future considerations

The implications and limitations set out in this chapter offer a number of fruitful areas for future work, as pointed out throughout this thesis. First, there is a need for well-controlled comparisons of several factors explored in this thesis. This would include work on different task demands, comparing different lengths of texts (sentences vs. passages vs. paragraphs), using different load manipulations, and measuring participants with a wider range of individual differences, for instance in comprehension abilities and in working memory. The approach set out here would expect differences in both eye movements and comprehension as a result of all of these, but understanding this in a range of conditions should elucidate the theoretical underpinning of these findings.

Another interesting question is how the manipulations in, for instance, Chapter 2 would affect processing of different syntactic ambiguities. Chapter 2 discussed several different syntactically ambiguous structures that have been previously explored (such as attachment ambiguities; Clifton et al., 2007). There are subtle differences in how these are processed – the load task may therefore be compatible with underspecification where this is possible, as opposed to the increased re-reading seen here (cf. von der Malsburg & Vasishth, 2013 for a similar argument).

Furthermore, it remains unclear what the n-back task is actually doing when it affects syntactic processing. Novick et al. (2014) put forward a specific link between *n-back* task performance and conflict resolution, and found that the link to syntactic ambiguity resolution was not seen for similar executive control tasks (for instance, training on a letter-number sequencing task). To test whether n-back task performance is linked specifically to syntactic ambiguity resolution, an interesting further question is whether a concurrent n-back task would affect reading more generally. For instance, does

it interact with effects such as lexical frequency, predictability, and lexical ambiguity? If the effect is specific to syntactic processing, there should be little difference; if it affects superficiality of reading more generally, a concurrent load may attenuate these effects.

6.6. Conclusion

This thesis has explored a range of issues around the general theme of whether syntactic processing is fully completed, whether the comprehension product of syntactic analysis is fully consistent with what has been read, and how load, task demands and individual differences affect these processes. By looking at eye movements and comprehension concurrently, the six experiments presented here have offered an insight into the entire process of how syntactically ambiguous sentences are analysed, and how they are eventually represented in comprehension. The results have demonstrated that raising cognitive load, asking participants questions, and presenting sentences in longer passages can all affect the extent to which people suffice with a superficial reading of text, rather than engaging in extensive reanalysis. Furthermore, individual differences in working memory (including differences due to ageing) had an impact on both eye movements and eventual comprehension, most notably the extent to which a *good-enough* parse was deemed acceptable. The experiments presented here open up several opportunities for future research, considering different manipulations of load, task demands, and of the nature of the stimuli being read. Two main implications are clear. The first is the importance of not ending data collection when the sentence has been read: comprehension is not guaranteed, even if reanalysis has been observed in the eye movement record. The second is not to treat syntactic (re)analysis as the same process under all conditions, or in all people. Sentence processing appears to be an adaptive, flexible system – sometimes, full analysis and complete comprehension are attained, but sometimes, a less complete analysis is good enough.

References

Aaronson, D., & Ferres, S. (1986). Reading strategies for children and adults: A quantitative model. *Psychological Review*, *93*(1), 89-112.

Altmann, G. T. M. (1988). Ambiguity, parsing strategies, and computational models. *Language & Cognition Processes, 3,* 73-97.

Altmann, G. T. (1989). Parsing and interpretation: An introduction. *Language and Cognitive Processes*, *4*, 1-19.

Altmann, G. (1994). Regression-contingent analyses of eye movements during sentence processing: Reply to Rayner and Sereno. *Memory & Cognition*, *22*(3), 286–290.

Altmann, G. T. M., Garnham, A., & Dennis, Y. (1992). Avoiding the garden-path: Eye movements in context. *Journal of Memory and Language, 31,* 685-712.

Altmann, G. T. M., & Steedman, M. J. (1988). Interaction with context during human sentence processing. *Cognition, 30,* 191-238.

Altmann, G. T. M., van Nice, K. Y., Garnham, A., & Henstra, J. A. (1998). Late closure in context. *Journal of Memory and Language, 38(4)*, 459-484.

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*(4), 390-412.

Baayen, R. H., & Milin, P. (2010). Analyzing reaction times. *International Journal of Psychological Research*, *3(2),* 12-28.

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

Baddeley, A., Chincotta, D., & Adlam, A. (2001). Working memory and the control of action: evidence from task switching. *Journal of Experimental Psychology: General, 130(4)*, 641-657.

Baddeley, A., & Hitch, G. (1974). Working memory. In G. H. Bower (Ed.), *The psychology of learning and motivation*: *8*, (pp. 47–89). New York: Academic Press.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*, 255–278.

Bartlett, F. (1932). *Remembering: A study in experimental and social psychology.* Cambridge: Cambridge University Press.

Barton, S. B., & Sanford, A. J. (1993). A case study of anomaly detection: Shallow semantic processing and cohesion establishment. *Memory & Cognition, 21(4)*, 477-487.

Bates, E., & MacWhinney, B. (1989). Functionalism and the competition model. In B. MacWhinney & E. Bates. (Eds), *The cross-linguistic study of sentence processing* (pp. 3-73). Cambridge, UK: Cambridge University Press.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software, 67(1),* 1-48.

Bever, T. (1970). The cognitive basis for linguistic structures. In J. R. Hayes (Ed.), *Cognition and the development of language.* New York: Wiley.

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

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.

Brébion, G. (2001). Language processing, slowing, and speed/accuracy trade-off in the elderly. *Experimental Aging Research, 27(2)*, 137-150.

Brébion, G. (2003). Working memory, language comprehension, and aging: Four experiments to understand the deficit. *Experimental Aging Research, 29(3)*, 269-301.

Britt, M. A., Perfetti, C. A., Garrod, S., & Rayner, K. (1992). Parsing in discourse: Context effects and their limits. *Journal of Memory and Language, 31(3)*, 293-314.

Bunting, M. F., Conway, A. R., & Heitz, R. P. (2004). Individual differences in the fan effect and working memory capacity. *Journal of Memory and Language*, *51*(4), 604-622.

Burke, D. M., & Osborne, G. (2007). Aging and inhibition deficits: Where are the effects? In C. M. MacLeod & D. S. Gorfein (Eds.), *Inhibition in cognition* (pp. 163-183). Washington, DC, US: American Psychological Association.

Butcher, K. R. & Kintsch, W. (2012). Text comprehension and discourse processing. In I. B. Weiner, A. F. Healy, & R. W. Proctor (Eds.), *Handbook of Psychology, Experimental Psychology* (2nd ed.) (pp. 578–605). Somerset, NJ: Wiley.

Caplan, D., DeDe, G., Waters, G., Michaud, J., & Tripodis, Y. (2011). Effects of age, speed of processing, and working memory on comprehension of sentences with relative clauses. *Psychology and Aging*, *26*(2), 439-450.

Caplan, D., & Waters, G. S. (1999). Verbal working memory and sentence comprehension. *Behavioral and Brain Sciences*, *22*, 77–94.

Caplan, D. & Waters, G. (2002). Working memory and connectionist models of parsing: A reply to MacDonald and Christiansen (2002). *Psychological Review, 109*, 66–74.

Caplan, D., & Waters, G. (2013). Memory mechanisms supporting syntactic comprehension. *Psychonomic bulletin & review*, *20*(2), 243-268.

Carlson, M. C., Hasher, L., Connelly, S. L., & Zacks, R. T. (1995). Aging, distraction, and the benefits of predictable location. *Psychology and Aging, 10,* 427-436.

Carreiras, M., & Clifton Jr, C. (1993). Relative clause interpretation preferences in Spanish and English. *Language and Speech, 36(4)*, 353-372.

Christiansen M. H. & Chater N. (2016). The Now-or-Never bottleneck: A fundamental constraint on language. *Behavioral and Brain Sciences, 39.*

Christianson, K. (2016). When language comprehension goes wrong for the right reasons: Good-enough, underspecified, or shallow language processing. *The Quarterly Journal of Experimental Psychology*, *69*(5), 817-828.

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

Christianson, K., Luke, S. G., Hussey, E. K., & Wochna, K. L. (2016). Why reread? Evidence from garden-path and local coherence structures. *The Quarterly Journal of Experimental Psychology*, *0218*(May), 1–51.

Christianson, K., Williams, C. C., Zacks, R. T., & Ferreira, F. (2006). Younger and Older Adults' "Good-Enough" Interpretations of Garden-Path Sentences. *Discourse Processes*, *42*(2), 205–238.

Clifton Jr, C., & Duffy, S. A. (2001). Sentence and text comprehension: Roles of linguistic structure. *Annual Review of Psychology*, *52*(1), 167-196.

Clifton Jr, C., & Ferreira, F. (1989). Ambiguity in context. *Language and cognitive processes*, *4*(3-4), 77-103.

Clifton, C., & Staub, A. (2008). Parallelism and Competition in Syntactic Ambiguity Resolution, *Language and Linguistics Compass, 2*(2), 234–250.

Clifton, C., Staub, A., & Rayner, K. (2007). Eye Movements in Reading Words and Sentences. In R. van Gompel (Ed.), *Eye movements: A window on mind and brain* (pp. 341-372). Amsterdam: Elsevier.

Clifton, C. J., Traxler, M. J., Mohamed, M. T., Williams, R. S., Morris, R. K., & Rayner, K. (2003). The use of thematic role information in parsing: Syntactic processing autonomy revisited. *Journal of Memory and Language*, *49*(3), 317–334.

Connelly, S. L., Hasher, L., & Zacks, R. T. (1991). Age and reading: the impact of distraction. *Psychology and Aging, 6(4)*, 533-541.

Conrad, R., & Hull, A. J. (1964). Information, acoustic confusion and memory span. *British Journal of Psychology*, *55*(4), 429-432.

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.

Craik, F. I. M. (1983). On the transfer of information from temporary to permanent memory. *Philosophical Transactions of the Royal Society for London: Series B, 302,* 341-359.

Craik, F. I., & Simon, E. (1980). Age differences in memory: The roles of attention and depth of processing. In: L. Poon. (Ed.), *New directions in memory and aging* (pp. 95-112). Hillsdale, NJ: Erlbaum.

Crain, S., & Steedman, M. J. (1985). On not being led up the garden path: The use of context by the psychological parser. In D. Dowty, L. Karttunen, & A. Zwicky (Eds.), *Natural language parsing: psychological, computational, and theoretical perspectives*. Cambridge, UK: Cambridge University Press.

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

Daneman, M., Hannon, B., & Burton, C. (2006). Are there age-related differences in shallow semantic processing of text? Evidence from eye movements. *Discourse Processes, 42,* 177-203.

Daneman, M., & Merikle, P. M. (1996). Working memory and language comprehension: A meta-analysis. *Psychonomic Bulletin & Review*, *3*(4), 422–433.

Daneman, M., Reingold, E. M., & Davidson, M. (1995). Time course of phonological activation during reading: Evidence from eye fixations. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21,* 884-898.

DeDe, G. (2014). Sentence comprehension in older adults: Evidence for risky processing strategies. *Experimental Aging Research, 40*, 436-454.

DeDe, G. (2015). Effects of animacy on processing relative clauses in older and younger adults, *The Quarterly Journal of Experimental Psychology, 68,* 487-498.

DeDe, G., Caplan, D., Kemtes, K., & Waters, G. (2004). The relationship between age, verbal working memory, and language comprehension. *Psychology and Aging, 19(4)*, 601-616.

DeDe, G., & Flax, J. K. (2016). Language comprehension in aging. In H. H. Wright (Ed.), *Cognition, Language, and Aging.* Philadelphia: John Benjamins.

Engelhardt, P. E. (2014). Children's and adolescents' processing of temporary syntactic ambiguity : an eye movement study. *Child Development Research, vol 2014,* Article ID 475315. doi:10.1155/2014/475315.

Engelmann, F., Vasishth, S., Engbert, R., & Kliegl, R. (2013). A framework for modelling the interaction of syntactic processing and eye movement control. *Topics in Cognitive Science, 5,* 452-474.

Evans, W. S., Caplan, D., Ostrowski, A., Michaud, J., Guarino, A. J., & Waters, G. (2015). Working memory and the revision of syntactic and discourse ambiguities. *Canadian Journal of Experimental Psychology*, *69*(1), 136-155.

Farmer, T. A., Misyak, J. B., & Christiansen, M. H. (2012) Individual differences in sentence processing. In M. J. Spivey, M. F. Joanisse & K. McRae (Eds.), *Cambridge Handbook of Psycholinguistics* (pp. 353–364). Cambridge, UK: Cambridge University Press.

Fedorenko, E., Gibson, E., & Rohde, D. (2006). The nature of working memory capacity in sentence comprehension: Evidence against domain-specific working memory resources. *Journal of Memory and Language*, *54*(4), 541–553.

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., & Christianson, K. (2016). Is Now-or-Never language processing good enough?. *Behavioral and Brain Sciences*, *39*.

Ferreira, F., & Clifton, C. (1986). The independence of syntactic processing. *Journal of Memory and Language*, *25*(3), 348–368.

Ferreira, F., & Henderson, J. M. (1991). Recovery from misanalyses of garden-path sentences. *Journal of Memory and Language, 30*(6), 725-745.

Ferreira, F., & Henderson, J. M. (1998). Syntactic reanalysis, thematic processing, and sentence comprehension. In J. D. Fodor & F. Ferreira (Eds.), *Reanalysis in sentence processing* (pp. 73–100). Dordrecht: Kluwer Academic Publishers.

Ferreira, F., & Patson, N. D. (2007). The "Good Enough" Approach to Language Comprehension. *Language and Linguistics Compass*, *2*, 71–83.

Fletcher, C. R., & Chrysler, S. T. (1990. Surface forms, textbases, and situation models: Recognition memory for three types of textual information. *Discourse Processes, 13(2),* 175-190.

Fodor, J. A. (1983). *The modularity of mind: An essay on faculty psychology.* Cambridge, Massachusetts: MIT Press.

Fodor, J. D., & Frazier, L. (1980). Is the human sentence parsing mechanism an ATN? *Cognition, 8*, 417--459.

Ford, M., Bresnan, J., & Kaplan R. M. (1982). A competence-based theory of syntactic closure. In: J. Bresnan (Ed.), *The mental representation of grammatical relations* (pp. 727-796). Cambridge, Massachusetts: MIT Press.

Forstmann, B. U., Tittgemeyer, M., Wagenmakers, E. J., Derrfuss, J., Imperati, D., & Brown, S. (2011). The speed-accuracy tradeoff in the elderly brain: a structural model-based approach. *Journal of Neuroscience*, *31*(47), 17242-17249.

Foss, D. J., & Cairns, H. S. (1970). Some effects of memory limitation upon sentence comprehension and recall. *Journal of Verbal Learning and Verbal Behavior, 9(5)*, 541-547.

Frazier, L. (1985). Syntactic complexity. In D. Dowty, L. Karttunen, & A. Zwicky (Eds.), *Natural language parsing: psychological, computational, and theoretical perspectives*. Cambridge, UK: Cambridge University Press.

Frazier, L. (1987). Sentence process: A tutorial review. In M. Coltheart (Ed.), *Attention and performance XII.* Hillsdale, N J: Erlbaum.

Frazier, L., Carminati, M. N., Cook, A. E., Majewski, H., & Rayner, K. (2006). Semantic evaluation of syntactic structure: Evidence from eye movements. *Cognition, 99(2),* 53-62.

Frazier, L., & Clifton, C. (1989). Identifying gaps in English sentences. *Language and Cognitive Processes, 4,* 93-126.

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

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.

Friedman, N. P., & Miyake, A. (2005). Comparison of four scoring methods for the reading span test. *Behavior research methods*, *37*(4), 581-590.

Gao, X., Levinthal, B. R., & Stine-Morrow, E. A. (2012). The effects of ageing and visual noise on conceptual integration during sentence reading. *Quarterly Journal of Experimental Psychology, 65(9)*, 1833-1847.

Gao, X., Stine-Morrow, E. A., Noh, S. R., & Eskew, R. T. (2011). Visual noise disrupts conceptual integration in reading. *Psychonomic bulletin & review, 18(1)*, 83-88.

Garnsey, S. M., Pearlmutter, N. J., Myers, E., & Lotocky, M. A. (1997). The contributions of verb bias and plausibility to the comprehension of temporarily ambiguous sentences. *Journal of Memory and Language*, *37*(1), 58-93.

Gibson, E., Bergen, L., & Piantadosi, S. T. (2013). Rational integration of noisy evidence and prior semantic expectations in sentence interpretation. *Proceedings of the National Academy of Sciences*, *110*(20), 8051-8056.

Gigerenzer, G., Todd, P. M., & ABC Research Group, T. (1999). *Simple heuristics that make us smart*. Oxford: Oxford University Press.

Glass, J. M., Schumacher, E. H., Lauber, E. J., Zurbriggen, E. L., Gmeindl, L., Kieras, D. E., & Meyer, D. E. (2000). Aging and the psychological refractory period: Task-coordination strategies in young and old adults. *Psychology and Aging, 15,* 571–595.

Göthe, K., Oberauer, K., & Kliegl, R. (2007). Age differences in dual-task performance after practice. *Psychology and Aging, 22*(3), 596-606.

Gordon, P. C., Hendrick, R., Johnson, M., & Lee, Y. (2006). Similarity-based interference during language comprehension: Evidence from eye tracking during reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *32*(6), 1304-1321.

Gordon, P. C., Hendrick, R., & Levine, W. H. (2002). Memory-load interference in syntactic processing. *Psychological Science: A Journal of the American Psychological Society / APS*, *13*(5), 425–430.

Gordon, P. C., & Lowder, M. W. (2012). Complex Sentence Processing: A Review of Theoretical Perspectives on the Comprehension of Relative Clauses. *Language and Linguistics Compass*, *6/7*, 403–415.

Gordon, P.C., Lowder, M.W., & Hoedemaker, R.S. (2016). Reading in normally aging adults. In Wright, H.H. (Ed.), *Cognitive-Linguistic Processes and Aging*, pp. 165-191. John Benjamins Publishing.

Gouvea, A. C., Phillips, C., Kazanina, N., & Poeppel, D. (2010). The linguistic processes underlying the P600. *Language and cognitive processes, 25(2)*, 149-188.

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

Grodner, D., & Gibson, E. (2005). Consequences of the serial nature of linguistic input for sentential complexity. Cognitive Science, 29(2), 261–290.

Hale, J. T. (2011). What a Rational Parser Would Do. *Cognitive Science, 35(3)*, 399–443.

Hamm, V. P., & Hasher, L. (1992). Age and the availability of inferences. *Psychology and Aging, 7*(1), 56-64.

Hartley, A. A. (1992). Attention. In F. I. M. Craik & T. A. Salthouse (Eds.), *Handbook of Aging and Cognition* (pp. 1-49). Hillsdale, NJ: Erlbaum.

Hasher, L., Zacks, R. T., & May, C. P. (1999). Inhibitory control, circadian arousal, and age. In D. Gopher & A. Koriat (Eds.), *Attention & Performance, XVII, Cognitive Regulation of Performance: Interaction of Theory and Application* (pp. 653-675). Cambridge, MA: MIT Press.

Healey, M. K., Hasher, L., & Campbell, K. L. (2013). The role of suppression in resolving interference: Evidence for an age-related deficit. *Psychology and Aging, 28(3)*, 721-728.

Hoeks, J. C., Vonk, W., & Schriefers, H. (2002). Processing coordinated structures in context: The effect of topic-structure on ambiguity resolution. *Journal of Memory and Language, 46(1),* 99-119.

Holmes, V. M., Kennedy, A., & Murray, W. S. (1987). Syntactic structure and the garden path. *The Quarterly Journal of Experimental Psychology*, *39*(2), 277-293.

Huestegge, L., & Bocianski, D. (2010). Effects of syntactic context on eye movements during reading. *Advances in Cognitive Psychology*, *6*(6), 79–87.

Huettig, F., Rommers, J., & Meyer, A. S. (2011). Using the visual world paradigm to study language processing: A review and critical evaluation. *Acta psychologica*, *137*(2), 151-171.

Hyönä, J., & Kaakinen, J. K. (2011). Effects of reading goal and reading task on eye fixation patterns. *Studies of Psychology and Behavior, 9*, 23-33.

Jaeger, T. F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language*, *59*(4), 434-446.

Jaeggi, S. M., Buschkuehl, M., Perrig, W. J., & Meier, B. (2010). The concurrent validity of the N-back task as a working memory measure. *Memory*, *18*(4), 394-412.

January, D., Trueswell, J. C., & Thompson-Schill, S. L. (2009). Co-localization of Stroop and syntactic ambiguity resolution in Broca's area: Implications for the neural basis of sentence processing. *Journal of Cognitive Neuroscience, 21(12)*, 2434-2444.

Jegerski, J. (2014). Self-paced reading. In J. Jegerski & B. VanPatten (Eds.), *Research methods in second language psycholinguistics* (pp. 20-49). New York: Routledge.

Johnson, R. E. (2003). Aging and the remembering of text. *Developmental Review*, *23*(3), 261-346.

Jordan, T. R., McGowan, V. A., & Paterson, K. B. (2014). Reading with filtered fixations: Adult age differences in the effectiveness of low-level properties of text within central vision. *Psychology and Aging, 29(2)*, 229-235.

Jost, K., Bryck, R. L., Vogel, E. K., & Mayr, U. (2011). Are old adults just like low working memory young adults? Filtering efficiency and age differences in visual working memory. *Cerebral cortex*, *21*(5), 1147-1154.

Just, M. A., & Carpenter, P. A. (1980). A theory of reading: From eye fixations to comprehension. *Psychological Review, 87,* 329-354.

Just, M. A, & Carpenter, P. A. (1992). A capacity theory of comprehension: individual differences in working memory. *Psychological Review*, *99*(1), 122–49.

Just, M. A., & Varma, S. (2002). A hybrid architecture for working memory: Reply to MacDonald and Christiansen (2002). *Psychological Review, 109,* 55-65.

Kaakinen, J. K., & Hyönä, J. (2005). Perspective effects on expository text comprehension: Evidence from think-aloud protocols, eyetracking, and recall. *Discourse processes*, *40*(3), 239-257.

Kaakinen, J. K., & Hyönä, J. (2007). Strategy use in the reading span test: An analysis of eye movements and reported encoding strategies. *Memory*, *15*(6), 634-646.

Kaakinen, J. K., & Hyönä, J. (2010). Task effects on eye movements during reading. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *36*(6), 1561–1566.

Kaakinen, J. K., Hyönä, J., & Keenan, J. M. (2003). How prior knowledge, WMC, and relevance of information affect eye fixations in expository text. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *29*(3), 447-457.

Kaakinen, J. K., Lehtola, A., & Paattilammi, S. (2015). The influence of a reading task on children's eye movements during reading. *Journal of Cognitive Psychology, 27(5)*, 640-656.

Kamide, Y., Altmann, G. T., & Haywood, S. L. (2003). The time-course of prediction in incremental sentence processing: Evidence from anticipatory eye movements. *Journal of Memory and Language*, *49*(1), 133-156.

Kane, M. J., Conway, A. R., Miura, T. K., & Colflesh, G. J. (2007). Working memory, attention control, and the N-back task: a question of construct validity. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *33*(3), 615-622.

Kaplan, R. (1972). Augmented transition networks as psychological models of sentence comprehension. *Artificial Intelligence, 3*, 77-100.

Karimi, H., & Ferreira, F. (2016). Good-enough linguistic representations and online cognitive equilibrium in language processing. *The Quarterly Journal of Experimental Psychology*, *69*(5), 1013–1040.

Kaschak, M. P., & Glenberg, A. M. (2004). This construction needs learned. *Journal of Experimental Psychology: General*, *133*(3), 450-467.

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), 157–70.

Kemper, S., & Herman, R. E. (2006). Age differences in memory-load interference effects in syntactic processing. *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences*, *61*(6), 327-332.

Kemper, S., Herman, R. E., & Lian, C. H. (2003). The costs of doing two things at once for young and older adults: Talking while walking, finger tapping, and ignoring speech of noise. *Psychology and Aging, 18(2)*, 181-192.

Kemper, S. & Liu C.-J. (2007). Eye movements of young and older adults during reading. *Psychology and Aging, 22*, 84–94.

Kemper, S. & McDowd, J. (2006). Eye movements of young and older adults while reading with distraction. *Psychology and Aging, 21*, 32–39.

Kemper, S., McDowd, J., Metcalf, K., & Liu, C-J. (2008). Young and older adults' reading of distracters, *Educational Gerontology, 34*, 489-502.

Kemper, S., Schmalzried, R., Herman, R., Leedahl, S., & Mohankumar, D. (2009). The effects of aging and dual task demands on language production. *Aging, Neuropsychology, and Cognition, 16(3)*, 241-259.

Kemtes, K. A., & Kemper, S. (1997). Younger and older adults' on-line processing of syntactic ambiguities. *Psychology and Aging, 12,* 362– 371.

Khan, M., & Daneman, M. (2011). How readers spontaneously interpret man-suffix words: Evidence from eye movements. *Journal of Psycholinguistic Research*, *40*(5), 351-366.

Kim, J. H., & Christianson, K. (2013). Sentence complexity and working memory effects in ambiguity resolution. *Journal of Psycholinguistic Research, 42(5),* 393–411.

Kim, S., Hasher, L., & Zacks, R. T. (2007). Aging and a benefit of distractibility. *Psychonomic Bulletin & Review*, *14*(2), 301-305.

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

King, J., & Just, M. A. (1991). Individual differences in syntactic processing: The role of working memory. *Journal of Memory and Language, 30,* 580-602.

Kintsch, W. (1988). The use of knowledge in discourse processing: A construction-integration model. *Psychological Review, 95*, 163-182.

Kintsch, W. (1998). *Comprehension: A paradigm for cognition*. Cambridge, UK: Cambridge University Press.

Kintsch, W., Welsch, D., Schmalhofer, F., & Zimny, S. (1990). Sentence memory: a theoretical analysis. *Journal of Memory and Language, 29,* 133-159.

Kliegl, R., Dambacher, M., Dimigen, O., Jacobs, A. M., & Sommer, W. (2012). Eye movements and brain electric potentials during reading. *Psychological Research, 76(2)*, 145-158.

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*, 262-284.

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

Kuperman, V., Drieghe, D., Keuleers, E., & Brysbaert, M. (2013). How strongly do word reading times and lexical decision times correlate? Combining data from eye movement corpora and megastudies. *Quarterly Journal of Experimental Psychology, 66,* 563-580.

Kuperman, V., & Van Dyke, J. A. (2011). Effects of individual differences in verbal skills on eye-movement patterns during sentence reading. *Journal of Memory and Language*, *65*(1), 42-73.

Laver, G. D. (2000). A speed-accuracy analysis of word recognition in young and older adults. *Psychology and Aging*, *15*(4), 705-709.

Levy, R. (2008). A noisy-channel model of rational human sentence comprehension under uncertain input. In *Proceedings of the conference on empirical methods in natural language processing* (pp. 234-243). Association for Computational Linguistics.

Levy, R., Bicknell, K., Slattery, T., & Rayner, K. (2009). Eye movement evidence that readers maintain and act on uncertainty about past linguistic input. *Proceedings of the National Academy of Sciences of the United States of America, 106(50)*, 21086–21090.

Lewis, R. (1998). Reanalysis and limited repair parsing: Leaping off the garden path. In J. Fodor & F. Ferreira (Eds.), *Reanalysis in Sentence Processing* (pp. 247–285). Dordrecht, The Netherlands: Kluwer Academic Publishers.

Lewis, R. L. (2000). Falsifying serial and parallel parsing models: Empirical conundrums and an overlooked paradigm. *Journal of Psycholinguistic Research*, *29*(2), 241-248.

Lewis, R. L., Shvartsman, M., & Singh, S. (2013). The Adaptive Nature of Eye Movements in Linguistic Tasks: How Payoff and Architecture Shape Speed-Accuracy Trade-Offs. *Topics in Cognitive Science, 5,* 581-610.

Lewis, R. L., & Vasishth, S. (2005). An activation-based model of sentence processing as skilled memory retrieval. *Cognitive science*, *29*(3), 375-419.

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

Liversedge, S. P., Paterson, K. B., & Pickering, M. J. (1998). Eye movements and measures of reading time. In: G. Underwood (Ed.), *Eye guidance in reading and scene perception* (pp. 55-75). Oxford: Elsevier.

Liversedge, S. P., Pickering, M. J., Branigan, H. P., & van Gompel, R. P. (1998). Processing arguments and adjuncts in isolation and context: The case of by-phrase ambiguities in passives. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 24(2)*, 461-475.

Logačev, P., & Vasishth, S. (2016a). A Multiple-Channel Model of Task-Dependent Ambiguity Resolution in Sentence Comprehension. *Cognitive Science*, *40,* 266-298.

Logačev, P., & Vasishth, S. (2016b). Understanding underspecification: A comparison of two computational implementations. *The Quarterly Journal of Experimental Psychology*, *69*(5), 996–1012.

Luke, S. G., & Christianson, K. (2016). Limits on lexical prediction during reading. *Cognitive Psychology*, *88*, 22-60.

Lustig, C., May, C. P., & Hasher, L. (2001). Working memory span and the role of proactive interference. *Journal of Experimental Psychology: General, 130(2)*, 199-207.

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., Pearlmutter, N. J., & Seidenberg, M. S. (1994). The lexical nature of syntactic ambiguity resolution. *Psychological review*, *101*(4), 676-703.

MacDonald, M. C., & Seidenberg, M. S. (2006). Constraint satisfaction accounts of lexical and sentence comprehension. In M. J. Traxler, & M. A. Gernsbacher. (Eds.), *Handbook of psycholinguistics* (pp. 581-611). London: Elsevier.

Macmillan, N. A., & Creelman, C. D. (2005). *Detection theory: A user's guide (2nd ed.).* Mahwah, New Jersey: Lawrence Erlbaum Associates.

Malyutina, S., & den Ouden, D. B. (2016). What is it that lingers? Garden-path (mis) interpretations in younger and older adults. *The Quarterly Journal of Experimental Psychology*, *69*(5), 880-906.

Marslen Wilson, W. (1975). Sentence perception as an interactive parallel process. *Science, 189*, 226-228.

Marslen-Wilson, W., & Tyler, L. K. (1980). The temporal structure of spoken language understanding. *Cognition*, *8*(1), 1-71.

Marslen-Wilson, W. D., Tyler, L. K., Warren, P., Grenier, P., & Lee, C. S. (1992). Prosodic effects in minimal attachment. *The Quarterly Journal of Experimental Psychology*, *45*(1), 73-87.

McConkie, G. W., Rayner, K., & Wilson, S. J. (1973). Experimental manipulation of reading strategies. *Journal of Educational Psychology*, *65*(1), 1-8.

McElree, B. (2001). Working memory and focal attention. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *27*(3), 817-835.

McElree, B. (2006). Accessing recent events. *Psychology of learning and motivation*, *46*, 155-200.

McElree, B., Foraker, S., & Dyer, L. (2003). Memory structures that subserve sentence comprehension. *Journal of Memory and Language*, *48*(1), 67-91.

McGowan, V. A., Paterson, K. B., & Jordan, T. R. (2013). Age-related visual impairments and perceiving linguistic stimuli: The rarity of assessing the visual abilities of older participants in written language research. *Experimental aging research*, *39*(1), 70-79.

McGowan, V. A., White, S. J., & Paterson, K. B. (2015). The effects of interword spacing on the eye movements of young and older readers. *Journal of Cognitive Psychology, 27(5)*, 609-621.

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*, 283–312.

Mehler, J., Bever, T. G., & Carey, P. (1967). What we look at when we read. *Attention, Perception, & Psychophysics, 2(5),* 213-218.

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.

Metzner, P., von der Malsburg, T., Vasishth, S., & Rösler, F. (2016). The Importance of Reading Naturally: Evidence From Combined Recordings of Eye Movements and Electric Brain Potentials. *Cognitive Science*.

Miller, L. M. S., Stine-Morrow, E. A., Kirkorian, H. L., & Conroy, M. L. (2004). Adult age differences in knowledge-driven reading. *Journal of Educational Psychology*, *96*(4), 811-821.

Mitchell, D. C. (1984). An evaluation of subject-paced reading tasks and other methods for investigating immediate processes in reading. In D. E. Kieras & M. A. Just (Eds.), *New methods in reading comprehension research*, (pp. 69-89). Hillsdale, NJ: Lawrence Erlbaum.

Mitchell, D.C. (1994). Sentence Parsing. In M.A. Gernsbacher (Ed.), *Handbook of psycholinguistics* (pp. 375-409). San Diego: Academic Press.

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.

Mitzner, T. L., Touron, D. R., Rogers, W. A., & Hertzog, C. (2010). Checking it twice: Age-Related differences in double checking during visual search. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 54, No. 18, pp. 1326-1330). Sage CA: Los Angeles, CA: SAGE Publications.

Mohamed, M. T., & Clifton Jr, C. (2011). Processing temporary syntactic ambiguity: The effect of contextual bias. *The Quarterly Journal of Experimental Psychology*, *64*(9), 1797-1820.

Mund, I., Bell, R., & Buchner, A. (2010). Age differences in reading with distraction: Sensory or inhibitory deficits?. *Psychology and aging*, *25*(4), 886-897.

Nicenboim, B., Logačev, P., Gattei, C., & Vasishth, S. (2016). When high-capacity readers slow down and low-capacity readers speed up: Working memory and locality effects. *Frontiers in psychology*, *7*:280.

Novick, J. M., Hussey, E., Teubner-Rhodes, S., Harbison, J. I., & Bunting, M. F. (2014). Clearing the garden-path: Improving sentence processing through cognitive control training. *Language, Cognition and Neuroscience*, *29*(2), 186-217.

Novick, J. M., Trueswell, J. C., & Thompson-Schill, S. L. (2005). Cognitive control and parsing: Reexamining the role of Broca's area in sentence comprehension. *Cognitive, Affective, & Behavioral Neuroscience, 5(3)*, 263-281.

Owsley, C. (2011). Aging and vision. *Vision research*, *51*(13), 1610-1622.

Öztekin, I., Güngör, N. Z., & Badre, D. (2012). Impact of aging on the dynamics of memory retrieval: A time-course analysis. *Journal of Memory and Language*, *67*(2), 285-294.

Paterson, K. B., McGowan, V. A., & Jordan, T. R. (2013). Filtered text reveals adult age differences in reading: evidence from eye movements. *Psychology and Aging, 28(2),* 352-364.

Patson, N. D., Darowski, E. S., Moon, N., & Ferreira, F. (2009). Lingering misinterpretations in garden-path sentences: evidence from a paraphrasing task. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *35*(1), 280–285.

Payne, B. R., Grison, S., Gao, X., Christianson, K., Morrow, D. G., & Stine-Morrow, E. A. (2014). Aging and individual differences in binding during sentence understanding: Evidence from temporary and global syntactic attachment ambiguities. *Cognition*, *130*(2), 157-173.

Payne, B. R., & Stine–Morrow, E. A. (2012). Aging, parafoveal preview, and semantic integration in sentence processing: testing the cognitive workload of wrap-up. *Psychology and aging*, *27*(3), 638-649.

Pickering, M. J., & Traxler, M. J. (1998). Plausibility and recovery from garden paths: An eye-tracking study. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 24*, 940–961.

Pickering, M. J., Traxler, M. J., & Crocker, M. W. (2000). Ambiguity resolution in sentence processing: Evidence against frequency-based accounts. *Journal of Memory and Language, 43,* 447-475.

Pollatsek, A., Reichle, E. D., & Rayner, K. (2006). Tests of the EZ Reader model: Exploring the interface between cognition and eye-movement control. *Cognitive Psychology*, *52*(1), 1-56.

Rabbitt, P. (1979). How old and young subjects monitor and control responses for accuracy and speed. *British Journal of Psychology*, *70*(2), 305-311.

Radach, R., Huestegge, L., & Reilly, R. (2008). The role of global top-down factors in local eye-movement control in reading. *Psychological Research*, *72*(6), 675–688.

Radvansky, G. A., & Copeland, D. E. (2004). Working memory span and situation model processing. *The American Journal of Psychology*, 191-213.

Radvansky, G. A., Copeland, D. E., & Zwaan, R. A. (2003). Brief report: Aging and functional spatial relations in comprehension and memory. *Psychology and Aging, 18(1)*, 161-165.

Radvansky, G. A., Zwaan, R. A., Curiel, J. M., & Copeland, D. E. (2001). Situation models and aging. *Psychology and Aging, 16(1)*, 145-160.

Raney, G. E. (2003). A context-dependent representation model for explaining text repetition effects. *Psychonomic Bulletin & Review*, *10*(1), 15-28.

Rayner, K. (1977). Visual attention in reading: Eye movements reflect cognitive processes. *Memory & Cognition, 5(4)*, 443-448.

Rayner, K. (1978). Eye movements in reading and information processing. *Psychological Bulletin, 85(3)*, 618-660.

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin, 124*, 372-422.

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

Rayner, K., Carlson, M., & Frazier, L. (1983). The interaction of syntax and semantics during sentence processing: Eye movements in the analysis of semantically biased sentences. *Journal of verbal learning and verbal behavior, 22(3)*, 358-374.

Rayner, K., Castelhano, M. S., & Yang, J. (2009). Eye movements and the perceptual span in older and younger readers. *Psychology and Aging*, *24*(3), 755–760.

Rayner, K., & Clifton, C., Jr. (2002). Language processing. In D. Medin (Ed.), *Stevens Handbook of Experimental Psychology* (pp 261-316). New York: John Wiley and Sons, Inc.

Rayner, K., & Frazier, L. (1987). Parsing temporarily ambiguous complements. *The Quarterly Journal of Experimental Psychology Section A*, *39*(4), 657–673.

Rayner, K., Garrod, S., & Perfetti, C. A. (1992). Discourse influences during parsing are delayed. *Cognition, 45(2)*, 109-139.

Rayner, K., Reichle, E. D., Stroud, M. J., Williams, C. C., & Pollatsek, A. (2006). The effect of word frequency , word predictability , and font difficulty on the eye movements of young and older readers. *Psychology and Aging, 21*(3), 448–465.

Rayner, K., & Sereno, S. C. (1994). Regressive eye movements and sentence parsing: On the use of regression-contingent analyses. *Memory & Cognition, 22(3)*, 281-285.

Rayner, K., Sereno, S. C., & Raney, G. E. (1996). Eye movement control in reading: a comparison of two types of models. *Journal of Experimental Psychology: Human Perception and Performance*, *22*(5), 1188-1200.

Rayner, K., Warren, T., Juhasz, B. J., & Liversedge, S. P. (2004). The effect of plausibility on eye movements in reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 30,* 1290-1301.

Reichle, E. D., & Drieghe, D. (2015). Using EZ Reader to examine the consequences of fixation-location measurement error. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *41*(1), 262-270.

Reichle, E. D., Warren, T., & McConnell, K. (2009). Using EZ Reader to model the effects of higher level language processing on eye movements during reading. *Psychonomic bulletin & review*, *16*(1), 1-21.

Riby, L., Perfect, T., & Stollery, B. (2004). The effects of age and task domain on dual task performance: A meta-analysis. *European Journal of Cognitive Psychology*, *16*(6), 863-891.

Rothkopf, E. Z., & Billington, M. J. (1979). Goal-guided learning from text: inferring a descriptive processing model from inspection times and eye movements. *Journal of educational psychology*, *71*(3), 310-327.

Salthouse, T. A. (1992). Influence of processing speed on adult age differences in working memory. *Acta Psychologica, 79,* 155-170.

Salthouse, T. A. (1994). The aging of working memory. *Neuropsychology, 8,* 535-543.

Salthouse, T. A. (1996). The processing-speed theory of adult age differences in cognition. *Psychological Review, 103,* 403-428.

Salthouse, T. A. (2003). Interrrelations of aging, knowledge, and cognitive performance. In U. Staudinger, & U. Lindenberger (Eds.), *Understanding human development: Dialogues with lifespan psychology* (pp. 265–287). Dordrecht, The Netherlands: Kluwer Academic Publishers.

Salthouse, T. A. (2004). What and when of cognitive aging. *Current Directions in Psychological Science, 13(4),* 140-144.

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

Sanz, M., Laka, I., Tanenhaus M. (2013). Sentence comprehension before and after 1970: topics, debates and techniques. In M. Sanz, I. Laka, & M. Tanenhaus (Eds.), *Language Down the Garden Path: The Cognitive and Biological Bases for Linguistic Structure*, Oxford: Oxford University Press.

Schmalhofer, F., & Glavanov, D. (1986). Three components of understanding a programmer's manual: verbatim, propositional, and situational representations. *Journal of Memory and Language, 25,* 279-294.

Schotter, E. R., Bicknell, K., Howard, I., Levy, R., & Rayner, K. (2014). Task effects reveal cognitive flexibility responding to frequency and predictability: Evidence from eye movements in reading and proofreading. *Cognition, 131,* 1-27.

Shake, M. C., Noh, S. R., & Stine- Morrow, E. A. (2009). Age differences in learning from text: Evidence for functionally distinct text processing systems. *Applied Cognitive Psychology, 23(4)*, 561-578.

Simon, H. A. (1956). Rational choice and the structure of the environment. *Psychological review*, *63*(2), 129-138.

Slattery, T. J., Sturt, P., Christianson, K., Yoshida, M., & Ferreira, F. (2013). Lingering misinterpretations of garden path sentences arise from competing syntactic representations. *Journal of Memory and Language*, *69(2)*, 104-120.

Slowiaczek, M. L., & Clifton, C. (1980). Subvocalization and reading for meaning. *Journal of verbal learning and verbal behavior*, *19*(5), 573-582.

Smiler, A., Gagne, D. D., & Stine-Morrow, E. A. (2003). Aging, memory load, and resource allocation during reading. *Psychology and aging*, *18*(2), 203-209.

Smith, G. A., & Brewer, N. (1995). Slowness and age: speed-accuracy mechanisms. *Psychology and aging*, *10*(2), 238-247.

Smith, N. J., & Levy, R. (2011). Cloze but no cigar: The complex relationship between cloze, corpus, and subjective probabilities in language processing. In L. Carlson, C. Hölscher, & T. Shipley (Eds.), *Proceedings of the 33rd Annual Conference of the Cognitive Science Society* (pp. 1637-1642). Austin, TX: Cognitive Science Society.

Smith, N. J., & Levy, R. (2013). The effect of word predictability on reading time is logarithmic. *Cognition, 128*(3), 302-319.

Speer, S. R., & Clifton, C. (1998). Plausibility and argument structure in sentence comprehension. *Memory & cognition*, *26*(5), 965-978.

Spivey, M. J., & Tanenhaus, M. K. (1998). Syntactic ambiguity resolution in discourse: modeling the effects of referential context and lexical frequency. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *24*(6), 1521-1543.

Spivey, M. J., Tanenhaus, M. K., Eberhard, K. M., & Sedivy, J. C. (2002). Eye movements and spoken language comprehension: Effects of visual context on syntactic ambiguity resolution. *Cognitive Psychology, 45*, 447–481.

Spivey-Knowlton, M. J., Trueswell, J. C., & Tanenhaus, M. K. (1993). Context effects in syntactic ambiguity resolution: Discourse and semantic influences in parsing reduced relative clauses. *Canadian Journal of Experimental Psychology, 47(2)*, 276-309.

Starns, J. J., & Ratcliff, R. (2010). The effects of aging on the speed–accuracy compromise: Boundary optimality in the diffusion model. *Psychology and Aging*, *25*(2), 377-390.

Staub, A. (2007). The return of the repressed: Abandoned parses facilitate syntactic reanalysis. *Journal of Memory and Language, 57*, 299-323.

Staub, A., Clifton, C., & Frazier, L. (2006). Heavy NP shift is the parser's last resort: Evidence from eye movements. *Journal of Memory and Language, 54(3)*, 389-406.

Staub, A., Grant, M., Astheimer, L., Cohen, A. (2015). The influence of cloze probability and item constraint on cloze task response time. *Journal of Memory and Language, 82,* 1-17.

Staub, A., Grant, M., Clifton Jr., C., Rayner, K. (2009). Phonological typicality does not influence fixation durations in normal reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*, 806-814.

Staub, A., White, S. J., Drieghe, D., Hollway, E. C., & Rayner, K. (2010). Distributional effects of word frequency on eye fixation durations. Journal of experimental psychology. *Human perception and performance, 36(5)*, 1280-1293.

Steedman, M. J., & Altmann, G. T. M. (1989). Ambiguity in context: A reply. *Language and Cognitive Processes, 4*, 105-122.

Stine-Morrow, E. A., Gagne, D. D., Morrow, D. G., & DeWall, B. H. (2004). Age differences in rereading. *Memory & Cognition, 32(5)*, 696-710.

Stine-Morrow, E. A. L., & Miller, L. M. S. (1999). Discourse processing and aging: Resource allocation as a limiting factor. In: S. Kemper & R. Kliegl (Eds.), *Constraints on language: Grammar, memory, and aging.* (pp. 53-76). Boston: Kluwer Academic Publishers.

Stine-Morrow, E. A., & Miller, L. M. (2009). Aging, self-regulation, and learning from text. *Psychology of learning and motivation*, *51*, 255-296.

Stine-Morrow, E. A., Miller, L. M. S., Gagne, D. D., & Hertzog, C. (2008). Self-regulated reading in adulthood. *Psychology and Aging, 23(1)*, 131-153.

Stine-Morrow, E. A., Miller, L. M. S., & Hertzog, C. (2006a). Aging and self-regulated language processing. *Psychological bulletin, 132(4)*, 582-606.

Stine-Morrow, E. A., Miller, L. M. S., & Leno, R. (2001). Patterns of on-line resource allocation to narrative text by younger and older readers. *Aging, Neuropsychology, and Cognition, 8(1)*, 36-53.

Stine-Morrow, E. A., Morrow, D. G., & Leno, R. (2002). Aging and the representation of spatial situations in narrative understanding. *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences, 57(4)*, 291-297.

Stine-Morrow, E. A. L., Ryan, S., & Leonard, J. S. (2000). Age differences in on-line syntactic processing. *Experimental Aging Research*, *26*(4), 315-322.

Stine-Morrow, E. A., Shake, M. C., Miles, J. R., Lee, K., Gao, X., & McConkie, G. (2010). Pay now or pay later: aging and the role of boundary salience in self-regulation of conceptual integration in sentence processing. *Psychology and aging*, *25*(1), 168-176.

Stine-Morrow, E. A., Shake, M. C., Miles, J. R., & Noh, S. R. (2006b). Adult age differences in the effects of goals on self-regulated sentence processing. Psychology and aging, 21(4), 790-803.

Sturt, P., Pickering, M. J., Scheepers, C., & Crocker, M. W. (2001). The preservation of structure in language comprehension: Is reanalysis the last resort? *Journal of Memory and Language*, *45*(2), 283-307.

Sturt, P., Scheepers, C., & Pickering, M. (2002). Syntactic ambiguity resolution after initial misanalysis: The role of recency. *Journal of Memory and Language, 46(2)*, 371-390.

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

Tabor, W., Galantucci, B., & Richardson, D. (2004). Effects of merely local syntactic coherence on sentence processing. *Journal of Memory and Language*, *50*(4), 355-370.

Tabor, W., Juliano, C., & Tanenhaus, M. K. (1997). Parsing in a dynamical system: an attractor-based account of the interaction of lexical and structural constraints in sentence processing. *Language and Cognitive Processes, 12*, 211–271.

Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M., & Sedivy, J. C. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science, 268*, 1632-1634.

Tanenhaus, M. K., Spivey-Knowlton, M. J., & Hanna, J. E. (2000). Modeling thematic and discourse context effects with a multiple constraints approach: Implications for the architecture of the language comprehension system. In M. Crocker, M. Pickering, & C. Clifton (Eds.), *Architectures and mechanisms of language acquisition and processing* (pp. 90–118). Cambridge, UK: Cambridge Univ. Press.

Townsend, D. J., & Bever, T. G. (2001). *Sentence comprehension: The integration of habits and rules.* Cambridge, MA: MIT Press.

Traxler, M. J. (2014). Trends in syntactic parsing: Anticipation, Bayesian estimation, and good-enough parsing. *Trends in cognitive sciences*, *18*(11), 605-611.

Traxler, M. J., Long, D. L., Tooley, K. M., Johns, C. L., Zirnstein, M., & Jonathan, E. (2012). Individual differences in eye-movements during reading: working memory and speed-of-processing effects. *Journal of Eye Movement Research, 5(1)*, 1-16.

Traxler, M. J., Morris, R. K., & Seely, R. E. (2002). Processing subject and object relative clauses: Evidence from eye movements. *Journal of Memory and Language, 47(1)*, 69-90.

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.

Traxler, M. J., Williams, R. S., Blozis, S. A., & Morris, R. K. (2005). Working memory, animacy, and verb class in the processing of relative clauses. *Journal of Memory and Language, 53(2)*, 204-224.

Trueswell, J. C., Tanenhaus, M. K., & Garnsey, S. M. (1994). Semantic influences on parsing: Use of thematic role information in syntactic ambiguity resolution. *Journal of Memory and Language, 33(3)*, 285.

Tun, P. A., McCoy, S., & Wingfield, A. (2009). Aging, hearing acuity, and the attentional costs of effortful listening. *Psychology and Aging, 24(3)*, 761-769.

van den Broek, P., Lorch, R. F., Linderholm, T., & Gustafson, M. (2001). The effects of readers' goals on inference generation and memory for texts. *Memory & Cognition, 29(8)*, 1081-1087.

van Dijk, T. A., & Kintsch, W. (1983). *Strategies of discourse comprehension.* New York: Academic Press.

van Dyke, J. A., & Johns, C. L. (2012). Memory interference as a determinant of language comprehension. *Language and Linguistics Compass*, *6*(4), 193–211.

van Dyke, J. A., & Lewis, R. L. (2003). Distinguishing effects of structure and decay on attachment and repair: A cue-based parsing account of recovery from misanalyzed ambiguities. *Journal of Memory and Language*, *49*(3), 285-316.

van Dyke, J. A., & McElree, B. (2006). Retrieval interference in sentence comprehension. *Journal of Memory and Language*, *55*(2), 157–166.

van Gompel, R. P., & Pickering, M. J. (2007). Syntactic parsing. In M. G. Gaskell (Ed.), *The Oxford handbook of psycholinguistics* (pp. 289-307). Oxford: Oxford University Press.

van Gompel, R. P., Pickering, M. J., Pearson, J., & Jacob, G. (2006). The activation of inappropriate analyses in garden-path sentences: Evidence from structural priming. *Journal of Memory and Language*, *55*(3), 335-362.

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(2)*, 284-307.

van Gompel, R. P., Pickering, M. J., & Traxler, M. J. (2000). Unrestricted race: A new model of syntactic ambiguity resolution. In A. Kennedy, R. Radach, D. Heller, & J. Pynte (Eds.), *Reading as a perceptual process*, (pp. 621-648). Oxford: Elsevier.

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.

van Heuven, W. J. B., Mandera, P., Keuleers, E, & Brysbaert, M. (2014). SUBTLEX-UK: a new and improved word frequency database for British English. *The Quarterly Journal of Experimental Psychology, 67*(6), 1176-1190.

Vasishth, S., von der Malsburg, T., & Engelmann, F. (2013). What eye movements can tell us about sentence comprehension. *Wiley Interdisciplinary Reviews: Cognitive Science, 4(2)*, 125-134.

Verhaeghen, P. (2003). Aging and vocabulary score: A meta-analysis. *Psychology and Aging, 18,* 332–339.

Verhaeghen, P., Marcoen, A., & Goossens, L. (1993). Facts and fiction about memory aging: A quantitative integration of research findings. *Journal of Gerontology*, *48*(4), 157-171.

Verhaeghen, P., Steitz, D. W., Sliwinski, M. J., & Cerella, J. (2003). Aging and dual-task performance: a meta-analysis. *Psychology and Aging, 18,* 443-460.

Vissers, C. T. W., Chwilla, D. J., & Kolk, H. H. (2007). The interplay of heuristics and parsing routines in sentence comprehension: Evidence from ERPs and reaction times. *Biological Psychology, 75(1)*, 8-18.

von der Malsburg, T. & Angele, B. (2017). False positives and other statistical errors in standard analyses of eye movements in reading. *Journal of Memory and Language*, *94*, 119-133.

von der Malsburg, T., Kliegl, R., & Vasishth, S. (2015). Determinants of scanpath regularity in reading. *Cognitive Science*, *39(7*), 1675-1703.

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

von der Malsburg, T. & Vasishth, S. (2013). Scanpaths reveal syntactic underspecification and reanalysis strategies. *Language and Cognitive Processes*, *28*, 1545-1578.

Wanner, E., & Maratsos, M. (1978). An ATN approach to comprehension. In M. Halle, J. Bresnan, & G. Miller (Eds.), *Linguistic theory and psychological reality* (pp. 119–161). Cambridge, MA: MIT Press.

Warner, J., & Glass, A. L. (1987). Context and distance-to-disambiguation effects in ambiguity resolution: Evidence from grammaticality judgments of garden path sentences. *Journal of Memory and Language*, *26*(6), 714-738.

Warrington, K.L., White, S.J., & Paterson, K.B. (2016). Ageing and the misperception of words: Evidence from eye movements during reading. *The Quarterly Journal of Experimental Psychology, 21*, 1-10.

Waters, G. S., & Caplan, D. (1996). The capacity theory of sentence comprehension: Critique of Just and Carpenter (1992). *Psychological Review, 103*, 761-772.

Waters, G. S., & Caplan, D. (2001). Age, working memory, and on-line syntactic processing in sentence comprehension. *Psychology and aging*, *16*(1), 128-144.

Waters, G. S., & Caplan, D. (2003). The reliability and stability of verbal working memory measures. *Behavior Research Methods, Instruments, & Computers. 35,* 550-564.

Waters, G., Caplan, D., & Hildebrandt, N. (1987). Working memory and written sentence comprehension. In M. Coltheart (Ed.), *Attention and performance XII: The psychology of reading* (pp. 531-555). Hillsdale, NJ: Erlbaum.

White, S. J., Warrington, K. L., McGowan, V. A., & Paterson, K. B. (2015). Eye movements during reading and topic scanning: Effects of word frequency. *Journal of Experimental Psychology, Human Perception and Performance, 41*, 233-248.

Whitford, V., & Titone, D. (2014). The effects of reading comprehension and launch site on frequency–predictability interactions during paragraph reading. *The Quarterly Journal of Experimental Psychology*, *67*(6), 1151-1165.

Wochna, K. L., & Juhasz, B. J. (2013). Context length and reading novel words: An eye-movement investigation. *British Journal of Psychology*, *104*(3), 347-363.

Wotschack, C., & Kliegl, R. (2013). Reading strategy modulates parafoveal-on-foveal effects in sentence reading. *The Quarterly Journal of Experimental Psychology*, *66*(3), 548-562.

Zacks, R., & Hasher, L. (1997). Cognitive gerontology and attentional inhibition: A reply to Burke and McDowd. *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences*, *52*(6), 274-283.

Appendix A – Stimuli used in experiments

Experimental items in Experiments 1 and 6

1. While the student read(,) the magazine that was boring **blew** off the desk.
2. While Jack ordered(,) the fish that was silver and black **cooked** in a pot.
3. While Susan wrote(,) the letter that was eloquent **arrived** from Amsterdam.
4. While the farmer steered(,) the tractor that was big and green **lay** in a ditch.
5. While the man hunted(,) the deer **ran** into the woods.
6. While the clown juggled(,) the balls **fell** on the ground.
7. While the reporter photographed(,) the rocket **sat** on the platform.
8. As the lion attacked(,) the baboon **screamed** in terror.
9. While Jim bathed(,) the child **giggled** with total delight.
10. While the chimps groomed(,) the baboons **sat** in the grass.
11. While Frank dried off(,) the car **stayed** on the driveway.
12. While Betty woke up(,) the neighbour **coughed** out loudly.
13. While the builder moved(,) the bricks that were stacked up **fell** to the floor.
14. While Anna dressed(,) the child that was blond and podgy **played** on the bed.
15. While the boy washed(,) the clothes that were folded **remained** very dirty.
16. While the nurse shaved(,) the patient that was in bed **watched** TV quietly.
17. As the maid dusted(,) the picture that was of the family **tipped** off the side.
18. While Kendra parked(,) the van that was brown and green **bumped** the kerb.
19. As Angela cleaned(,) the dog that was spotted and black **stood** in the yard.
20. As the professor lectured(,) the students that were nervous **took** some notes.
21. As Mark vacuumed(,) the curtains **hung** in the window.
22. While the puppy sniffed(,) the kitten **sat** on the sofa.

23. As the athlete wrestled(,) the opponent **shouted** insults to him.
24. While the secretary typed(,) the memo **arrived** in the mail.
25. While the father changed(,) the baby **played** with its toys.
26. While the mother undressed(,) the boy **cried** very loudly.
27. While the parents cuddled(,) the child that was clever **played** happily alone.
28. While the woman showered(,) the dog that was black **ran** into the room.
29. While the thief hid(,) the jewellery **sparkled** in the light.
30. While the girl scratched(,) the cat **stared** at the dog.
31. Since the soldier divorced(,) the woman who was attractive **looked** for a man.
32. While the builder stripped(,) the carpet that was red **looked** quite dirty.

Experimental items in Experiments 2, 3 and 5

1. While Emma undressed(,) the child that was small and happy **played on the** bed. Emma finished **undressing herself** and walked towards the wardrobe.
2. While the girl washed(,) the clothes that were folded **remained very** dirty. The girl then stopped **washing herself** and she then decided to put the clothes on.
3. While the parents cuddled(,) the child that was clever **played alone** happily. The parents stopped **cuddling each other** and went to play with their child.
4. Since the lawyer divorced(,) the woman who was tall **looked for a** new house. The lawyer had **divorced her husband** and asked the other woman to move in.
5. While the builder stripped(,) the carpet that was brown **looked quite** dirty. The builder was cooler after **stripping his shirt** and decided to clean the carpet.
6. While the woman showered(,) the dog that was black **ran into the** room. The woman stopped **showering herself** and went out to play with the dog.

7. While the armies fought(,) the navy that were attacking **found a new** location. The armies stopped **fighting each other** but the naval battle continued.
8. While the friends hugged(,) the nurse that was busy **cared for the** patient. The friends stopped **hugging each other** and thanked the nurse for her work.
9. While Daniel bathed(,) the infant that was excited **laughed with total** delight. Daniel had to finish **bathing himself** and so dried off and checked on his child.
10. While the chimps groomed(,) the gorillas that were scary **ate lots of** lunch. The chimps stopped **grooming themselves** and went over to the gorillas.
11. While Matt dried off(,) the van that was getting old **stayed on the** driveway. Matt eventually **dried himself** off completely and drove himself to work.
12. While Rebecca woke up(,) the neighbour that was in bed **shouted out** loudly. Rebecca was fully **woken up** by now and heard the shouting next door.
13. While the father changed(,) the baby that was cuddly **played with its** toys. The father had to finish **changing his clothes** to pick up and look after the baby.
14. While the mother dressed(,) the boy that was demanding **cried very** loudly. The mother managed to **dress herself** and could then attend to her crying son.
15. While the burglar hid(,) the jewellery that was expensive **sparkled in the** light. The burglar stopped **hiding himself** while the owners went to call the police.
16. While the boy scratched(,) the cat that was black and white **looked at the** dog. The boy stopped **scratching himself** and the dog stared back at the cat.

Experimental items in Experiment 4

Emma felt she had to stay indoors on such a hot day like this. While Emma undressed the child that was small and happy **played on** the bed. The child was cute and giggled **on the bed**/**up at Emma** loudly. Emma finished **undressing and** walked towards the wardrobe.

The girl had been playing outside before she came in. While the girl washed the clothes that were folded **remained very** dirty. The girl knew she would probably **wear/ruin** the clothes again later. The girl stopped **washing herself** and she decided to put the clothes on.

The parents were busy and didn't have much time to relax. While the parents cuddled the child that was clever **played alone** happily. The child often played while his parents **were/slept** at home. The parents stopped **cuddling and** went to play with their child.

The successful lawyer had been married for many years. Since the lawyer divorced the woman who was tall **looked for** a new house. The woman hadn't moved since she got **employed/married** years before. The lawyer **divorced her** husband and asked the other woman to move in.

The builder had been working hard on the house for a while. While the builder stripped the carpet that was brown **looked quite** dirty. The carpet had been in good need of being **cleaned/replaced** for years. The builder was cooler after **stripping and** decided to clean the carpet.

The woman had two dogs but only one of them was black. While the woman showered the dog that was black **ran into** the room. The black dog did not seem to enjoy being **alone/washed** at all. The woman stopped **showering and** went out to play with the dog.

The war continued for years and the enemy didn't stop. While the men fought the enemy that was hiding **found their** new position. The enemy hoped they wouldn't be **caught/killed** on that day. The men stopped **fighting each** other and focused on the enemy.

The friends had been in the hospital for a long time. While the friends hugged the nurse that was busy **cared for** the patient. The nurse was happy to be **thanked/working** while at the hospital. The friends stopped **hugging and** thanked the nurse for her hard work.

Dan only had one child who needed regular care. While Dan bathed, the infant that was excited **laughed with** total delight. The infant was always happy to be **loved/clean** and cared for. Dan completely **dried himself** with a towel and checked on his child.

The chimps lived together with all the gorillas in the new zoo. While the chimps groomed, the gorillas that were scary **ate lots** of lunch. The gorillas were relaxed when the **wardens/chimps** were there. The chimps stopped **grooming themselves** and went over to the gorillas.

Matt had been driving his white van for many years now. While Matt dried off, the van that was getting old **stayed on** the driveway. The van stayed in condition if **checked/washed** regularly. Matt was eventually **completely dry** and he went to drive to work.

Rebecca had lived near her neighbour for twenty years. While Rebecca woke up, the neighbour that was in bed **shouted out** loudly. The neighbour was often disturbed by **sunrise/Rebecca** nowadays. Rebecca was **fully awake** now and heard the shouting next door.

The father loved his newborn baby, even if it was noisy. While the father changed, the baby that was cuddly **played with** its toys. The baby would wriggle around when being **ignored/cleaned** by his dad. The father **changed into** his new clothes and picked up the baby.

The mother had a really big house with many bedrooms. While the mother dressed, the boy that was demanding **cried very** loudly. The boy often kept crying when his mother **worried/changed** him. The mother **finished dressing** and attended to her son.

The burglar had been stealing for many years now. While the burglar hid, the jewellery that was expensive **sparkled in** the light. The jewellery was obvious to the **owners/burglar** in the house. The burglar stopped **hiding while** the owners left and ran away.

The boy had spent all day with his pets - a cat and a dog. While the boy scratched, the cat that was black and white **looked at** the dog. The cat was not happy being **watched/touched** by others. The boy stopped **scratching himself** and the dog stared back at the cat.

Sentences used in Experiment 5 memory span task

Having had a great holiday, Liz knew she'd have to return for a longer period.
On such a sunny day, Fred was incredibly happy that it was finally summer.

It was always amazing to see how the dog always had so much energy.
It was a pity that the one man had nothing, while the other had plenty.

John didn't want a ring made of gold, but one instead made of silver.
The one family spent Sundays in shops, but the other family went to church.

Hannah was very pleased that her daughter had decided to become a doctor.
As the election approached, everyone agreed it was time for a new leader.
With so many options, it was very difficult for Tim to make the right choice.

Having looked on all the shelves, Janet finally found the packet on the bottom.
The paint Jim had chosen wasn't right, and he needed to find a new colour.
Ben had been out all day and couldn't wait to find out the result.

Sally knew she had to write a cheque, but couldn't remember the amount.
The one queue was totally full with people, while the other queue had nobody.
Lewis wanted to help somehow, but wasn't sure of the right course of action.

It was late in the year and Alice needed to finish writing the report.
Tony didn't think the team stood a chance of winning so late in the season.
It was possible that after all this time, the athlete might finally beat their record.
Some people enjoy working outdoors, while others are happier staying in an office.

While having fun is always great, it is important to look after your health.
The department was good at what it did, but not sticking to a budget.
Some cars stay close to the edge, but it's better to stay near the centre.
Ali knew the importance of staying sensible and keeping his feet on the ground.

Sandra had an idea Paul wasn't happy, but she just didn't know the reason.
It was unclear why the clothes had been thrown around all over the street.
The town centre was very busy, which suggests there was probably a market.
Terry was the eldest child and Barry the youngest, with Robert in the middle.

Owen kept running, but his brother was tired and had to stop for a minute.
The one child looked like her mother, while the other resembled her father.
The computers had been down all day, and there was a fault with the system.
On a warm day, there was nothing better than going down to the garden.
The three people all met up on Saturdays and went to see their friend.

There will probably be a lot more use of technology in the future.
While there were lots of animals at the zoo, the keeper was the only person.
The event attracted a lot of single people but also the odd couple.
Some children used the snow for a day off, but others made it to school.
It was a risky decision to make, but the Chief Executive took a chance.

The burglary had happened late at night, but they still called the police.
The office worker called their colleagues ages ago but still had no answer.
Bill was expecting a quick reply, but he instead had to wait a moment.
Golf can be enjoyable but it involves a lot of walking around the course.
The couple spent some time alone, but far more time with their family.

Appendix B: Additional analyses from Experiment 1

Accuracy

Accuracy ~ Structure * Load * Verb * Length + (Structure*Verb | | Participant) + (1 | Item)

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	2.62691	0.31950	8.222	< 2e-16 ***
Structure	-0.63483	0.25586	-2.481	0.013095 *
Load	-0.16095	0.36620	-0.440	0.660286
Verb	1.97686	0.54404	3.634	0.000279 ***
Length	0.90076	0.52635	1.711	0.087021 .
Str:Load	0.16407	0.51374	0.319	0.749451
Str:Verb	-0.41335	0.40117	-1.030	0.302841
Load:Verb	-0.29432	0.46385	-0.635	0.525750
Str:Length	0.02015	0.39541	0.051	0.959360
Load:Length	0.18582	0.39366	0.472	0.636905
Verb:Length	1.36146	1.05063	1.296	0.195024
Str:Load:Verb	0.31324	0.80571	0.389	0.697449
Str:Load:Length	0.04129	0.79300	0.052	0.958471
Str:Verb:Length	-1.04576	0.78793	-1.327	0.184437
Load:Verb:Length	-0.79376	0.78436	-1.012	0.311547
Str:Load:Verb:Length	0.92567	1.57965	0.586	0.557877

First pass duration

log(firstpass) ~ Structure * Load * Verb * Length + (Structure*Verb*Length | | Participant) + (Structure*Load | | Item)

	Estimate	Std. Error	t value
(Intercept)	5.406079	0.030472	177.41
Structure	0.047542	0.021072	2.26
Load	-0.055422	0.055246	-1.00
Verb	-0.015365	0.031025	-0.50
Length	0.005528	0.033191	0.17
Str:Load	-0.072349	0.042437	-1.70
Str:Verb	0.015172	0.035077	0.43
Load:Verb	0.056522	0.034589	1.63
Str:Length	-0.029227	0.033440	-0.87
Load:Length	-0.073842	0.041857	-1.76
Verb:Length	-0.057163	0.062384	-0.92
Str:Load:Verb	-0.032264	0.070829	-0.46
Str:Load:Length	-0.040634	0.067581	-0.60
Str:Verb:Length	0.050482	0.076761	0.66
Load:Verb:Length	0.008933	0.070378	0.13
Str:Load:Verb:Length	-0.196892	0.154779	-1.27

Go past duration

log(gopast) ~ Structure * Load * Verb * Length + (Structure*Verb*Length | | Participant) + (Structure | Item)

	Estimate	Std. Error	t value
(Intercept)	5.636857	0.045608	123.59
Structure	0.192000	0.050945	3.77
Load	0.076180	0.080928	0.94
Verb	0.018703	0.052714	0.35
Length	-0.011056	0.051678	-0.21
Str:Load	0.029091	0.083451	0.35
Str:Verb	-0.009277	0.083767	-0.11
Load:Verb	0.065864	0.063393	1.04
Str:Length	-0.008904	0.084262	-0.11
Load:Length	-0.030267	0.059874	-0.51
Verb:Length	-0.056094	0.103326	-0.54
Str:Load:Verb	-0.107485	0.119920	-0.90
Str:Load:Length	0.055609	0.121337	0.46
Str:Verb:Length	0.008555	0.175279	0.05
Load:Verb:Length	0.024895	0.119663	0.21
Str:Load:Verb:Length	-0.595842	0.261103	-2.28

Total reading duration

log(totaltime) ~ Structure*Load*Verb*Length + (Structure*Verb*Length | | Participant) + (Structure*Load | | Item)

	Estimate	Std. Error	t value
(Intercept)	5.820898	0.055077	105.69
Structure	0.211102	0.034220	6.17
Load	0.097140	0.080435	1.21
Verb	0.137927	0.079844	1.73
Length	0.015854	0.079794	0.20
Str:Load	0.160905	0.053909	2.98
Str:Verb	0.045934	0.072971	0.63
Load:Verb	0.046692	0.053210	0.88
Str:Length	-0.124825	0.068434	-1.82
Load:Length	-0.004558	0.052885	-0.09
Verb:Length	-0.055916	0.157668	-0.35
Str:Load:Verb	-0.048859	0.119129	-0.41
Str:Load:Length	0.041727	0.107804	0.39
Str:Verb:Length	0.066629	0.140302	0.47
Load:Verb:Length	-0.146382	0.093581	-1.56
Str:Load:Verb:Length	-0.378421	0.224279	-1.69

Second pass duration

log(secondpass) ~ Structure*Load*Verb* Length + (Structure*Verb*Length | | Participant) + (Structure*Load | | Item)

	Estimate	Std. Error	t value
(Intercept)	5.543735	0.046199	120.00
Structure	0.125552	0.052516	2.39
Load	0.026033	0.075792	0.34
Verb	0.109151	0.071293	1.53
Length	0.064739	0.067101	0.96
Str:Load	0.213715	0.075964	2.81
Str:Verb	0.006883	0.105823	0.07
Load:Verb	0.098501	0.095760	1.03
Str:Length	-0.075171	0.113651	-0.66
Load:Length	0.034081	0.082709	0.41
Verb:Length	0.024015	0.134410	0.18
Str:Load:Verb	-0.107680	0.154042	-0.70
Str:Load:Length	-0.182320	0.175154	-1.04
Str:Verb:Length	-0.102364	0.238410	-0.43
Load:Verb:Length	-0.092867	0.166075	-0.56
Str:Load:Verb:Length	0.160511	0.378458	0.42