Ward, Peter Maurice (2008) *The time course of sentence interpretation.* PhD thesis.

http://theses.gla.ac.uk/56/

Glasgow Theses Service
http://theses.gla.ac.uk/
theses@gla.ac.uk

The Time Course of Sentence Interpretation

Peter Ward

Department of Psychology, University of Glasgow

Submitted for the Degree of PhD to the Higher Degree Committee of the Faculty of Information and Mathematical Sciences, University of Glasgow.

September 2007

Abstract

The investigation of Shallow Processing, also known as Underspecification, and 'Good Enough' processing, is a relatively new branch of psycholinguistics. A growing body of evidence within this field indicates that, in some cases, the comprehension system will fail to build or retain a fully specified representation for linguistic input. As a result, the construction of underspecified representations may lead to erroneous interpretations, and the phenomenon of Pragmatic Normalisation is a central instance of this: comprehenders sometimes construct interpretations that reflect pragmatic knowledge rather than the grammatically licensed meaning of the input. Some researchers have suggested that shallow processing can be explained in terms of the comprehension system using reliable – but essentially statistical – heuristic interpretation processes. This heuristic style of interpretation is in contrast with interpretative processes that construct meaning based on the syntactic structure of a string, and one outstanding question is how these different interpretation processes operate in real time.

In a series of eight experiments this thesis investigated the time course of sentence interpretation via a study of pragmatic normalisation. Experiments 1-6 probed interpretations of syntactically unambiguous, implausible sentences, replicating some earlier studies and reporting surprisingly high levels of unlicensed interpretations. Experiments 2-8 used a variety of implausible constructions to investigate the temporal relation of syntax-based interpretation to heuristics-based interpretation. Both self-paced reading and eyetracking data are supportive of a processing model in which syntax informs the interpretation process first, but is later overruled by pragmatic constraints. Investigations into the conditions for shallow processing indicate a role for memory and syntactic complexity, and the opportunity to reread implausible material. An investigation into the impact of reading skill on the tendency to normalise implausible sentences yielded inconsistent results, and there is apparently little difference in the processing styles of skilled and less-skilled readers when reading implausible material. The thesis concludes with suggestions for future work to further elucidate the time course of syntactic vs. heuristic interpretation.

Acknowledgments

Special thanks to my supervisors Professor Anthony Sanford and Dr. Patrick Sturt for their guidance and encouragement. Thanks to my wife, Janis, for her love and support throughout. Thanks to my parents who got me this far. Thanks to the Glasgow University Language Group for their helpful comments and advice. Apologies and grateful thanks to the friends and family who kept me going through all my many little crises.

This research was funded by ESRC 1+3 grant PTA-030-2003-01577

Table of Contents

Page No.

Declaration

I declare that this thesis is my own work, carried out under the normal terms of supervision.

………………………………………………

Peter Ward

Chapter 1: Introduction

Shallow processing: Evidence and questions

There currently exists a body of evidence to support the view that comprehenders build mental representations of language that are not fully specified on all dimensions. Dating back to Fillenbaum's work on pragmatic normalisation (1971, 1974), studies have increasingly shown that language comprehenders often fail to build, or to retain, fully accurate representations for linguistic input in terms of either semantics or syntax, and fail to make adequate use of syntactic information when computing the meaning of a string. The various individual cases have come to be referred to as examples of 'shallow processing' (A. J. Sanford and Sturt, 2002), underspecification (A. J. Sanford and Graesser, 2006) and 'Good Enough representations' (Christianson, Hollingworth, Halliwell and Ferreira, 2001; Ferreira, Ferraro and Bailey, 2002), and there is a growing awareness that complete theories of language comprehension will now need to account for this behaviour (A J Sanford and Graesser 2006; Ferreira, 2003).

The idea that the Human Sentence Processor (HSP) may habitually derive underspecified interpretations for linguistic input may seem surprising. As Ferreira (2003) has pointed out, a central assumption within the study of language processing has been that in order to determine the meaning of a string, the comprehension system first computes its syntactic structure, and that the meaning of a string will then be syntactically determined. Therefore, assuming a string is syntactically unambiguous, the scope for misinterpretation ought to be extremely small. Indeed, the assumption has been that only rarely would the comprehension system assign a meaning to a string that is at odds with its syntactic frame (Frazier and Clifton, 1996). Within the study of ambiguity resolution – a primary concern within psycholinguistics – models differ in terms of the type of information they allow the parser to draw on in the initial stages of processing. But both 'syntax first' models (Frazier and Rayner, 1982; Ferreira and Clifton, 1986) and 'constraint-based' models (MacDonald, Pearlmutter and Seidenberg, 1994; Tanenhaus, Spivey-Knowlton, Eberhard and Sedivy, 1995) focus on how the parser solves problems in its initial syntactic analysis in order to derive the correct interpretation, and so neither type of model has found the necessity of parsing – the construction of a syntactic representation – to be particularly controversial in its status as necessarily prior to interpretation (Ferreira 2003, p165).

However, in developing the Good Enough approach to language processing, Christiansen, Ferreira and colleagues have argued that the comprehension system may actually need to be guided by cognitive heuristics at least some of the time (Ferreira, seminar given at the University of Edinburgh, June 2005) and pointed out that non-algorithmic processing is accepted as a feature in other cognitive domains – for example, underspecification in visual perception (Simons and Levin 1997), and the 'Fast and Frugal Heuristics' approach to decision making (Gigerenzer, 2000; Gigerenzer, Todd, and the ABC Reasearch Group, 1999). Algorithmic processes are those which operate strictly by a relevant set of rules and would, if allowed to run their course, guarantee the correct solution to a task (e.g. Bayesian calculations). Heuristic processing, by contrast, runs according to statistical probabilities and would be expected to generate the correct solution a useful, or significant proportion of the time. An example would be the field of Sequential Decision Making which models decision making under natural conditions unfavourable to fully rational thought (Henderson, Falk, Minut, Dyer, & Mahadevan, 2001, apply this to gaze control in human vision). Given the conditions under which much natural language processing takes place, for example poverty of input, the occurrence of dysfluencies, and frequent ungrammaticalities in production, a comprehension system running purely on algorithmic processes would reliably crash upon encountering the smallest deviation from its operating rules, and may not always be able to operate algorithmically due to time pressures. There is also the matter of the computational difficulties that would arise in the attempt to build fully specified representations of, for example, sentences employing multiple quantification (Hobbs and Schieber, 1987). And obviously, the comprehension system *is* adept at deriving the intended meaning from a string that is syntactically incomplete or incoherent. On the Good Enough view, the purpose of the comprehension system is to derive an interpretation for a string and in many cases a partial interpretation will suffice. The argument is not that the comprehension system *mainly* operates using meaning-based heuristics, ignoring information as useful as grammar, and there are few proponents of 'semantics-only' theories of comprehension (Ferreira, 2003, p192). But taking all of this into account, it has been argued that models of language comprehension may now need to include architectural components to account for non-algorithmic processing (Ferreira, 2003, p168).

This review will first recount some well-known examples of so-called shallow processing, and focus on instances of comprehenders apparently constructing interpretations that are at odds with those specified by syntax. It will continue the discussion of the evidence by looking at some of the main areas of interest within this field of research. It will then state the aims of this thesis in terms of the questions that remain open in the literature and the most promising lines of investigation.

Evidence for shallow processing

The famous Moses illusion (Erikson and Mattson, 1981) has shown many times that readers and listeners will fail to build a fully specified meaning for the sentence, *how many of each type of animal did Moses put on the ark?* The anomaly of course lies in the fact that it was actually Noah who put the animals on the ark. But participants will frequently answer ‘two’, rather than pointing out the erroneous assumption in the question, demonstrating that they had failed to extract the correct meaning from a syntactically unambiguous and non-complex string. A similar effect was reported by Barton and A J Sanford (1993) who observed substantial failures to report the anomaly in their ‘air crash’ scenario. Participants who read *where should the survivors be buried?* detected the anomaly (that survivors should not be buried) only 59% of the time. This finding has recently been replicated by Daneman, Lenertz and Hannon (2007) who used a more explicit procedure for reporting anomalies and still observed an average of only 67% detection of anomalous noun phrases that were internally coherent (e.g. *tranquillising sedatives* as opposed to the internally incoherent *tranquillising stimulants*). Other studies using intentional monitoring procedures still indicated that this type of anomaly is very difficult to detect (Kamas, Reder and Ayers, 1996; Reder and Kusbit, 1991).

Fillenbaum's work with memory for discourse demonstrated that what readers retained was not a perfectly accurate memory of a passage, but rather, the gist of a passage was retained while details were dropped or elided (1971, 1974). Significantly, there was evidence of unusual and anomalous parts of a text having been rendered more acceptable, or 'pragmatically normalised'. An example is the sentence, *don't print that or I won't sue you*, which has the correct paraphrase, 'if you refrain from printing that, I will sue you'. On recollection, participants tended to paraphrase this to mean something like, 'if you do print that, I will sue you', suggesting that the correct interpretation had perhaps not been made in the first place, or had been made and was subsequently overridden by more global semantic constraints (the 'normalised' meaning is certainly the more frequent occurrence). Further evidence comes from so-called 'depth charge' sentences (Wason and Reich, 1979; Natsopoulos, 1985) such as *no head injury is too trivial to be ignored*. There is a general tendency for this sentence to be interpreted as meaning, 'no matter how trivial it seems to be, every head injury should be treated', although this is not supported by the grammar. (A correct paraphrase is actually, 'no matter how trivial a head injury seems to be, it should be ignored'.) In order for the human sentence processor to derive the preferred, incorrect interpretation, it must perform incomplete local semantic analysis, or correct local analysis must be overridden at a global level (Sanford and Sturt, 2002).

As with this last example, several studies have suggested that when interpreting a string, comprehenders may build a representation that is not supported by the syntactic frame. Duffy, Henderson and Morris (1989) reported evidence suggesting that readers in their study were making semantic connections which were unlicensed by the grammar. In a task which required participants to name the final word in a sentence, Duffy et al. observed the same amount of facilitation (relative to an appropriate baseline) in the sentence *the boy watched the bartender serve the cocktails* as in the sentence *the boy who watched the bartender served the cocktails*. Clearly priming had occurred in each case, but the (syntactically determined) semantic relation between the words *bartender* and *cocktails* is different in each case and it would appear that the meaning representations built by the participants were not sufficiently determined by the syntax to reflect this difference. (However, in a later eyetracking study, Morris (1994) reported contrasting results showing that, given a semantic link between a target word and the words in its context, reading times on the target word were shorter if the *message-level* representation of the sentence was semantically related to the target word than if it was not.) Garnham and Oakhill (1987) studied interpretations of elliptical verb phrases (EVP), a difficult construction requiring the precise reconstruction in memory of a previously read verb phrase. Following a sentence such as *The elderly patient had been examined by the doctor*, participants read an elided VP that was either plausible or implausible. The plausible variant was e.g. *the child had too* and the implausible variant was *the nurse had too.* Participants then answered a question about whether the doctor had examined the child/nurse. When the second sentence had an implausible meaning, i.e. that the doctor had examined the nurse, participants answered correctly only 75% of the time. When an adjunct phrase (e.g. *during the ward round*) occurred between the by-phrase and the second sentence, accuracy dropped even further to 61%. Clearly a fully specified representation for the sentence, built around a correct syntactic analysis, had either not been faithfully built or was not retained, and a greater strain on memory, resulted in higher rates of misinterpretation.

There are also cases of the HSP appearing to prefer ungrammaticality. Ferreira and Swets (2005) discuss the common occurrence in everyday, unplanned speech of sentences such as:

We're afraid of things that we don't know what they are

Sentences like this contain island violations and use a 'resumptive pronoun' (in this case, *they*) where legally there should be only a gap. As such, they are ungrammatical (or at least marginally grammatical) in English, and yet are readily comprehensible by native speakers. Also, Gibson and Thomas (1999) reported the 'missing verb phrase' effect in acceptability judgements of doubly-nested relative clause structures such as,

the ancient manuscript that the graduate student who the new card catalogue had confused a great deal was studying in the library was missing a page

This sentence structure requires three VPs to be grammatical, yet participants rated the sentences as just as acceptable when only two VPs were present, giving the sentence:

The ancient manuscript that the graduate student who the new card catalogue had confused a great deal was missing a page.

This is further evidence that the comprehension system can satisfy itself with meaning-based representations that do not fully take into account the syntactic information available. (Gibson and Thomas discuss their findings in terms of memory limitations these sentences are, after all, long and complex.)

There is a small literature on the perseverance of non-grammatical interpretations in garden path sentences, even after the ambiguity has been resolved and the globally correct interpretation has been extracted. Christianson, Hollingworth, Halliwell and Ferreira (2001; findings replicated 2006) presented their participants with sentences like the following:

while the chef stirred the soup boiled over

As is typical of this type of garden path sentence, the NP *the soup* is initially parsed as the direct object of the verb *stirred*. Subsequent input, however, reveals this analysis to be incorrect, and *the soup* must be reanalysed as the subject of the main clause, *the soup boiled over*. In order to directly probe the participant's interpretation of the sentence, Christianson et al. asked questions probing the interpretation of the (initially misanalysed) subordinate clause, e.g. '*did the chef stir the soup?*'. If interpretation was based on a complete syntactic reanalysis then participants should never answer 'yes' to this question. However, participants responded 'yes' up to 51% of the time, and even up to 43% of the time when the incorrect interpretation could not be maintained as an inference, as in the case of a sentence like, *while the chef stirred the soup thawed on the counter*, where it is inconceivable that the chef was stirring the soup. Not only had participants derived a syntactically unlicensed interpretation, but confidence ratings indicated that participants were highly confident in the interpretations they had supplied – they were almost as confident in their incorrect ‘yes’ responses as in their correct ‘no’ responses. This effect, though somewhat smaller, was observed with a class of verbs called Reflexive Absolute Transitive (RAT) verbs (Trask, 1993). RAT verbs form a unique class in that, in the absence of a direct object, they must obligatorily be understood as reflexive. Thus, in a sentence such as

While Anna dressed the baby that was small and cute spit up on the bed

once the NP *the baby* has been correctly analysed as the subject of the matrix clause, the syntactic and thematic role assignment properties of the verb *dressed* absolutely prohibit any interpretation – based on inference or general reasoning – that involves Anna dressing the baby. The implicit object of *dressed* can only be a reflexive, thus determining the meaning that Anna dressed herself. Yet participants responded ‘yes’ approximately 60% of the time to the question, ‘did Anna dress the baby?’. Christianson et al. concluded that the initial misanalysis had never been fully reanalysed and had persevered, allowing the HSP to hold contradictory interpretations of the sentence. Michael and Gordon (2003) tested very similar items in an eyetracking experiment and challenged the conclusion that misinterpretations were due to initial misanalysis, reporting that the role of inference in overall interpretation was the key to whether or not a garden path sentence would be fully reanalysed. That is, garden path sentences would be more or less reanalysed depending on the success of global inferences. But their results still indicated that that readers would misanalyse a sentence such as

while Sally rode her pony rested in its stall, incorrectly answering ‘true’ 25% of the time in response to the question ‘did Sally ride her pony?’. (Michael and Gordon did not extend their investigation to the RAT class of verbs). The idea that syntactically unlicensed interpretations can persevere receives further support from Sturt (2003), who presented readers with items like,

the explorers found the South Pole was right at their feet/ was out of reach

in a segment-based, self-paced reading study. The segment *...was right at their feet* is consistent with the initial likely misanalysis in which *South Pole* was the direct object of *found*, while the segment *...was out of reach* is inconsistent with the early parsing error. Results indicated that readers would spend longer reading the segment *...was out of reach*, suggesting that the interpretation resulting from the initial misanalysis had persevered, interfering with the ultimate interpretation despite being syntactically unlicensed. (Sturt, 2007, replicated this result in an eyetracking study.)

In a paper already referred to, Ferreira (2003) conducted a series of experiments aimed at testing a heuristic proposed by Townsend and Bever (2001), namely, the preference of the comprehension system to interpret any noun-verb-noun string as *agent*-verb-*theme* in its thematic structure (or *proto-agent* – verb – *proto-patient*, as with Dowty, 1991). Experiment 1 demonstrated that passive sentences were frequently misinterpreted (approximately 25% of the time) when they contained implausible ideas, e.g. *the dog was bitten by the man*. The problematic factor appeared to be the necessity of assigning thematic roles in an atypical order (*theme-verb-agent*), and experiments 2 and 3 ruled out the possibility that the misinterpretations were due to the surface form frequency of the sentences. As well as the test of a specific heuristic, these studies were significant in that they reported high rates of systematic misinterpretation with unambiguous, relatively non-complex sentences.

There are several studies, then, showing that comprehenders can fail to extract the correct meaning of a word (e.g. *Moses*, *Survivors*) in otherwise unchallenging contexts; that they can construct meaning-based interpretations that do not fully reflect the syntactic information available; and that they can even prefer sentences that are ungrammatical to their fully grammatical versions.

Conditions for shallow processing

With a number of studies having investigated the phenomenon of shallow processing under different conditions, we are in a position to outline some conditions which are likely to elicit shallow processing and normalised interpretations. Contributing to, or narrowing, this list should be a main concern of work in this area. To begin with the best known example – the Moses illusion – Bredart and Modolo (1988) showed that detection rates could be increased by recasting the original sentence in a focused cleft construction and directly probing its interpretation, i.e. *it was Moses who put two of each kind of animal on the ark. True or False?*. So the semantic illusion effect seems to be best served by the critical impostor word being out of linguistic focus. A further lesson from this example is the importance of semantic relatedness between the correct word and its replacement. Erickson and Mattson (1981) demonstrated that detection rates were improved when *Adam* was used instead of *Moses*, and readers were never fooled when *Nixon* was used. Barton and Sanford (1993) reported that detection rates were influenced by the fit of an anomalous word to the overall scenario described in the context ('fit' was understood in terms of statistical fit rather than a match in the semantic content of the words). Hence readers fell for the *survivors* anomaly because of the good overall fit of *survivors* with a plane crash scenario. Detections rates increased when *survivors* was used in the context of a bicycle crash scenario. (These three factors were combined in a recent study by Nieuwland and Van Berkum (2005), see below.) Hannon and Daneman (2004) reported that reading skill was a factor in anomaly detection. Readers classified as less-skilled on the Nelson Denny Reading test (Form E: Brown, Bennett, & Hanna, 1981) fared worse overall than skilled readers with the type of anomaly used by Barton and Sanford (1993), and struggled especially with anomalous NPs that were themselves locally incoherent (see below for further details). As we have seen, there is evidence that non-canonical role assignments, as with passive sentences may elicit misinterpretation if the sentences contain implausible ideas (Ferreira, 2003). The missing verb phrase effect (Gibson and Thomas, 1999) strongly suggests a role for syntactic complexity in underspecification, and, as the authors themselves discussed, a role for memory. Although some evidence for shallow processing comes from studies using surprisingly un-complex materials (e.g. Ferreira, 2003), the idea that syntactically challenging sentences are susceptible to shallow processing receives support from studies showing that there is a tendency for comprehenders not to fully reanalyse difficult garden path sentences unless absolutely necessary (Christianson refs, 2001, 2006). If heuristics are only employed when the parser is having difficulty, then we would expect more evidence of shallow processing with complex sentences. As yet, only Garnham and Oakhill (1987) have included a memory/complexity manipulation in a study reporting normalised interpretations, and further work should test these ideas more explicitly, via manipulations of complexity and memory load, and using methodologies that vary in the constraints they place on memory.

Time course

Kim and Osterhout (2005) conducted an ERP investigation into the differential ‘control’ exerted by syntax and semantics during the course of sentence interpretation, and concluded that, in certain syntactically ambiguous circumstances, the semantic properties of a sentence are determined independently of its syntax and can even guide parsing operations. They presented their participants with sentences such as

The hearty meal was devouring the kids.

Syntactic cues unambiguously support an agent interpretation of *the meal*, but semantic cues suggest a theme assignment (*meal* is an ideal theme for *devouring*). The logic of their experiment was that if comprehenders assigned the role of agent to *meal*, as might be expected given its position as subject and the argument structure of ‘devour’, it would render the main verb semantically anomalous and thus elicit an N400 effect in the ERP waveform. On the other hand, if *meal* was assigned a theme interpretation, which would be inconsistent with the grammar, then the main verb would be rendered syntactically anomalous and elicit a P600 effect in the waveform. The results showed that the main verbs in such sentences were associated with the P600 effect, and the authors concluded that the semantic link between *meal* and *devouring* had determined the interpretation and led participants to perceive a syntactically well formed sentence to be ungrammatical. (A further experiment ruled out any account of the results based on a simple animacy contrast. Sentences such as *the dusty tabletops were devouring…*elicited an N400 rather than a P600, indicating that the P600 observed in their experiment 1 had not simply been due to the inanimate subject noun causing an early commitment to a passive main verb form.)

The N400 and P600 effects are not universally accepted as being straightforward indices of semantic and syntactic anomaly respectively; however, if Kim and Osterhout's results can be interpreted in this way, and the P600 reflects a processing cost associated with syntactic anomaly, then this suggests a processing strategy in which semantics exert primary control on interpretation. A more recent ERP study (Nieuwland and van Berkum, 2005) reported a similar result with a 'change deafness' experiment, in which an animacy violation – a clear breach of acceptable semantics – elicited a P600 rather than an N400. In this case the authors interpreted the effect as representing a delayed anomaly detection rather than a response to a syntactic violation; however, the interpretation remained possible that what caused the delay in detection was the very early operation of an interpretative heuristic based on semantic association (the violator word, *suitcase*, was closely related to the expected *tourist*). This 'semantics-first' idea is in line with the comprehension model proposed by Townsend and Bever (2001), the Late Assignment of Syntax, or LAST model. This model contends that the comprehension system first analyses input according to semantic associations and syntactic habits (such as the N-V-N strategy), and only then checks the input against the time-consuming, but more reliable, syntactic algorithms available.

Clearly then, a major issue relates to the time course of heuristic processing, particularly in its relation to the, syntax-based interpretation which comprises a major assumption of most theories of comprehension. Ferreira (2003) argued for the operation of both heuristic and algorithmic processes, but accepted that current evidence cannot help us understand the temporal relationships between them.

There are several possibilities. Syntax-based interpretation could operate first, but see its output overridden or ignored by heuristic processes that operate based on pragmatic knowledge and the semantic relations present in a string. There is some experimental support for this account. Sturt (2003) reported ungrammatical interpretations in a study investigating the timing of binding constraints on the interpretation of reflexive anaphors. Participants were more likely to misinterpret if the gender of the anaphor matched the stereotypical gender of an antecedent, even though that antecedent was not legally grammatical in terms of binding constraints. Eyetracking data indicated that even though participants ultimately misinterpreted, there was evidence of very early application of grammatical constraints. The non-grammatical interpretation had therefore been generated either later online, or offline. Regarding the timing of grammatical constraints in general, there is recent evidence that grammatical structure building (of one sort at least) happens very early and with a high degree of precision, even in cases where the surface word order could be expected to make this difficult. Phillips (2006) reported self-paced reading data which demonstrated that the comprehension system incrementally posits gaps in just those environments where parasitic gaps are acceptable. It could be, then, that grammar is indeed the primary input into the interpretation process, and that interpretation is therefore primarily an algorithmic activity.

There is, however, some evidence (cited earlier) that semantics – and therefore heuristic processes – are ‘in control’ of the early generation of an interpretation. Kim and Osterhout (2005) argued for this account, and Nieuwland and Van Berkum (2005) allowed for its possibility. So it could be that the grammar of a sentence is used only after an interpretation has been generated on the basis of semantics/pragmatics, perhaps as a standard checking mechanism, or because the heuristic output has generated an error signal.

A third alternative is that the two types of processing operate in parallel, with one stopping as the other reaches completion. So in the case of a sentence that is ultimately misinterpreted, perhaps the parsing operations never managed to finish and a syntax-based interpretation was therefore never generated. The first two alternatives tend to suggest a competition model, in which syntax and semantics both generate interpretations and, where they differ, one must be chosen in order to arrive at a final interpretation. The third alternative is more an ‘either/or’ model, in which only one interpretation is arrived at (syntax or semantics-based), either because one type of processing did not reach completion, or because it was never begun in the first place.

By examining the online processing of sentences such as those used by Ferreira (2003), it should be possible to tell at what point the syntax is informing the interpretation. Evidence that parsing is operating in the earliest stages might bolster a ‘syntax-first’ model; this evidence in conjunction with the kind of ungrammatical interpretations outlined above might suggest an account in which interpretive output from early parsing operations is overridden by semantic constraints and schematic knowledge (as in Sturt 2003, above). On the other hand, an absence of parsing evidence, either early-on or a total absence, would support an account, like that of Kim and Osterhout (2005) and Townsend and Bever (2001), in which heuristic-based semantic processing is dominant.

Individual differences

If the use of heuristics is indeed a component of the language comprehension system then their operation may be subject to individual differences. For instance, if it is the case that semantic heuristics only operate under especially difficult syntactic conditions, less-skilled readers may be more prone than skilled readers to interpreting input relying on non-syntactic cues. There is recent ERP evidence that readers with a low working memory capacity, as indexed by a low reading span, will experience difficulty with syntactically complex sentences similar to the difficulty they experience with syntactically ambiguous sentences (Bornkessel, Fiebach, and Friederici, 2004). And if heuristic-based comprehension is more typical of less-skilled readers then reading skill may influence online processing of material likely to elicit normalised interpretations, perhaps resulting in less-skilled readers taking longer than skilled readers to detect an implausible phrase. Some recent studies have analysed reading skill as a factor in shallow processing. Hannon and Daneman (2004) tested interpretations of ‘incidental anomalies’ such as Barton and Sanford’s air crash scenario and observed that both skilled and less-skilled readers were prone to errors, but less-skilled readers were significantly more susceptible, and only this group had particular difficulty with locally anomalous NPs (e.g. *tranquillising stimulants*). Daneman, Lennertz and Hannon (2007) replicated these findings. Both these studies included measures of processing time (sentence reading time, 2004; eyetracking, 2007) but processing results were averaged across reading skill as it yielded no significant results in any analysis. However, these studies used only one type of anomaly and the processing measures may not have been suitably subtle to capture any differences caused by reading skill. The online (eyetracking) detection measures involved only initial looking times and look-back times on the anomalous region, and so could not reveal precisely when an anomaly was detected if it wasn't detected immediately. Daneman, Hannon and Burton (2006) eyetracked similar items with older and younger readers and reported that older readers were no more susceptible to shallow processing than younger readers, but that they were more adept at detecting locally coherent anomalous NPs (e.g. *tranquillising sedatives*) than younger readers. They proposed that in these cases older readers were able to draw on their more developed linguistic knowledge. While these age-related findings are interesting, it would be worthwhile to continue testing for reading-skill differences and to analyse the effects of reading skill on the interpretation and processing of several different types of materials in order to draw firmer conclusions.

Thesis aims

This thesis has the following aims: First, it will attempt to replicate reports of shallow processing and *normalisation* which constitute evidence against syntax-only accounts of interpretation and support the Good Enough or Shallow view of language processing. In particular it will use materials similar to those used by Ferreira and Stacey (unpublished manuscript), Ferreira (2003), and Garnham and Oakhill (1987). These studies reported substantial amounts of normalised interpretations and should thus be a fruitful starting point for investigation. Experiments 1-3 will test interpretations of materials similar to those used in the Ferreira studies, while experiment 4 will test materials similar to those used by Garnham and Oakhill, adding a manipulation of *voice* to attempt further investigation into the heuristic theory associated with implausible passive constructions. Second, it will employ methodologies enabling an examination of the time course of this phenomenon. If there is a growing acceptance of the idea that the language processor uses semantic heuristics, leading to normalised, non-syntactic interpretations, the question remains open as to when these heuristics operate, particularly in their relation to syntax-based interpretation processes. The use of self-paced reading and eyetracking in conjunction with implausible materials should reveal the point at which syntactic interpretation occurs, and thus indicate whether, under our conditions, this happens at the earliest stages, or whether it is delayed relative to any putative heuristic processes. A trend for 'late' anomaly detection or even the failure to detect anomalies online would suggest the prominence of early, non-syntactic interpretation. A trend towards immediate or very early detection, especially in difficult/complex materials, would argue for the primary operation of syntax-led interpretation. This latter trend would suggest that semantic or heuristic-based interpretation is not a primary process, but rather operates at a global level, interfering at a later stage with early interpretations based on syntax. (A further advantage of using online processing measures is that they provide a more reliable index of anomaly detection than replying on self report (Daneman, Hannon and Burton, 2006).) Third, it will attempt to narrow the list of conditions under which normalisation is likely to be observed. In particular, it will examine the effect of syntactic complexity and memory constraints on normalisation. Experiments 2-3 will manipulate syntactic load to examine whether normalisation is more likely under high-load conditions. Further, a range of methodologies will be used: Experiment 1 will use the questionnaire format, experiments 2-4 will use word-by-word, self-paced reading, and experiments 5-8 will use eyetracking. The questionnaire format obviously allows free and natural reading. Eyetracking likewise allows for natural reading and rereading, but the questions which probe the interpretation of a passage are normally presented without the opportunity to consult that passage, and thus present a greater challenge to memory. Word-by-word, self-paced reading rules out the opportunity for making regressive eye movements and is thus, methodologically, the most challenging reading format in terms of memory constraints. Testing implausible materials under these different reading conditions should allow useful conclusions here. Fourth, there is the question of individual differences in relation to shallow processing, which has been addressed in the literature only recently. Experiments 4-6 will examine interpretations with regard to reading skill and, with experiments 7-8, will also examine the effects of reading skill on the moment-by-moment, online processing and detection of implausible and anomalous material. The question of how reading skill affects this sort of processing has received only slight attention and has so far failed to yield any significant findings. Examining effects of reading skill over several different anomaly types and using more comprehensive processing measures could allow a more robust contribution to the literature. Fifth, again focusing on conditions for normalisation and shallow processing, it will investigate the online effects of one theory of normalisation. Barton and Sanford (1993) reported differences in anomaly detection depending on the statistical fit of an anomalous word or phrase with its preceding context, and proposed that one factor driving interpretation is the fit of a word to its context. One question is whether this factor is active at the earliest stages of interpretation, affecting immediate, local processing of anomalous words and phrases. Experiments 7-8 will test anomalies of differing severity, manipulating their fit with a preceding context. Under free reading conditions, reading times will be used as an index of anomaly detection to determine whether a well-fitting context can interfere with, i.e. delay, the syntax-based interpretation processes that would lead to detection, relative to a neutral context. If statistical associations between words are involved in interpretation, and act at a very early stage, then in the type of sentence prone to being normalised we might expect to see a later anomaly detection following a well-fitting context.

Chapter 2:

The Interpretation of Implausible Sentences

Introduction

Ferreira (2003) and Ferreira and Stacey (unpublished manuscript) investigated interpretations of non-canonical constructions such as passives and clefts. Despite the sentences always being syntactically non-ambiguous, results indicated that implausible sentences, e.g. *the dog was bitten by the man*, were misinterpreted up to 25% of the time. Readers incorrectly judged implausible sentences to be plausible (Ferreira and Stacey experiment 1) and listeners made errors on a thematic role identification task (Ferreira, 2003, experiment 1). The misinterpretations were quite clearly made with reference to pragmatic constraints: it's not impossible that a man would bite a dog, but our real world knowledge argues strongly against it, preferring the more common situation in which dogs bite men, and the misinterpretations reflected that kind of pragmatic bias.

While pragmatic bias is in many ways very reasonable, the important point is that the misinterpretations were made at the expense of a clear and reliable guide to correct interpretation, namely, syntax. While Ferreira points out that her results by no means argue for a non-syntactic approach to interpretation in general, the considerable error rates, and the nature of the errors, indicate that the dictates of syntax with regards to meaning were being either ignored or overridden. And as mentioned in the literature review, this is a state of affairs at odds with traditional understandings of how the comprehension system generates meaning. The syntactic approach to interpretation appeared to be losing out to a non-algorithmic, heuristic style of interpretation.

The particular heuristic that Ferreira favoured in this case was one based on the canonicity of thematic role assignments. In the case of passives, the canonical *agent-verb-patient* assignment order must be reversed to give *patient-verb-agent*. The argument is that the comprehension system, taken in by the pragmatic cues, assigns the roles in the most common order so that the sentence *the dog was bitten by the man* is assigned the (incorrect) meaning DOG-BIT-MAN. Ferreira suggests this heuristic, called the N-V-N strategy, is one such non-algorithmic interpretation technique commonly employed by the comprehension system. Her results (from both experiments mentioned above) indicated that readers and listeners fared significantly worse when interpreting implausible passives than when interpreting the same sentence in the active voice.

In a series of three experiments we attempted to replicate these results with similar sentences, using both a plausibility judgement task (experiment 1) and a thematic role judgement task (experiments 2 and 3). In an additional investigation, central to this thesis, we included an online measure of reading time in order to examine the time course of interpretation as well as the ultimate interpretations made by readers. If the comprehension system does indeed employ heuristic interpretation strategies, then a question fundamental to their investigation will be how they operate in real time, particularly with regard to interpretation processes based on the computation of syntactic structure. Within the design of experiments using implausible or anomalous material, evidence of disruption in the processing record, at or after an anomaly, provides evidence of correct online interpretation informed by the grammar. In the experiments reported by Ferreira (who deliberately did not include measures of processing time), it is possible that readers and listeners were simply not aware of the implausible nature of what they were reading/hearing. Measures of online processing will allow us to test that possibility by monitoring for online disruption, and will provide more evidence at this early stage in the investigation of how syntactic and heuristic processes operate in real time.

Experiment 1

Experiment 1 aimed to replicate the basic findings of Ferreira and Stacey's (unpublished) Experiment 1: that readers would systematically misinterpret implausible sentences, and demonstrate systematic difficulty with passive sentences compared with active sentences. In their experiment, participants read active and passive sentences that were either plausible or implausible, and demonstrated a tendency to rate an implausible sentence as being plausible when it was passive in form (e.g. the dog was bitten by the man), but not when it was active. Accuracy in judging plausibility was virtually 100% for active and passive plausible sentences, and active implausible sentences, but participants judged passive implausible sentences to be plausible 26% of the time. The present experiment collected plausibility judgements on a modified set of Ferreira and Stacey's 'biased reversible' sentences (see Materials section for an explanation of this term).

Method

Participants

24 participants from the University of Glasgow student population took part in this experiment and received payment or course credit for their participation.

Materials and design

The present study used a modified set of Ferreira and Stacey's 'biased reversible' sentences. This type of sentence can be defined as a simple transitive sentence in which the verb's arguments can be switched without producing a strict anomaly, although one arrangement of the arguments is much less plausible than the other. In the example already given, *The dog was bitten by the man*, the order of the arguments *dog* and *man* do not describe an impossible scenario, but the scenario they describe is obviously less plausible than if they were switched to give *the man was bitten by the dog*. Ferreira and Stacey presented sentences like this on-screen, instructing readers to read them until they were confident they had understood them; when readers had done this, they rated them for plausibility.

On examination of Ferreira and Stacey's items, several failed to meet the relevant criteria of one arrangement of articles being implausible but not anomalous. For instance, *the horse was thrown by the rider* was felt to be anomalous rather than simply unlikely/implausible. Further items were designed to produce a set of 37 experimental items. An item could be in either the active or passive voice and could be either plausible or implausible; Table 1 presents an item in each of its 4 conditions. This produced a Voice*Plausibility, 2X2 within-subjects design.

The 37 experimental items were interspersed with 38 filler items, modelled on the experimental items. All filler items were designed to be semantically plausible. Experimental items and fillers were split across four lists according to a latin square rotation. There were six participants in each of the four list-groups. In order to control for possible practice effects, each list was divided into two halves and the order of presentation reversed for half of the lists, so that three participants in each group saw the two halves of the list in one order, and three participants saw the other order.

Table 1: Example experimental item in all 4 conditions

Condition	**Material**
Active Plausible	*The lawyer sued the builder for one million pounds*
Active Implausible	*The builder sued the lawyer for one million pounds*
Passive Plausible	*The builder was sued by the lawyer for one million pounds*
Passive Implausible	*The lawyer was sued by the builder for one million pounds*

Procedure

Experimental items and filler items were presented in questionnaire format. Participants were required to read each sentence and indicate their judgement of a sentence's plausibility on a 7-point scale, ranging from highly plausible (1) to highly implausible (7). Rereading was not discouraged, as sentences in Ferreira and Stacey's experiment were presented on-screen, all at once, with participants making their judgement only once they were confident they had understood a sentence.

Results

Two analyses of variance were computed for the plausibility ratings: one that treated participants as a random variable (F_1) and one that treated items as a random variable (F_2). The mean plausibility ratings for the four voice/meaning conditions are presented in Table 2 below.

Table 2: Mean plausibility ratings by condition

Condition	**Plausibility Rating**	
	Mean	(SD)
Active Plausible	1.82	(0.60)
Active Implausible	5.48	(0.62)
Passive Plausible	1.89	(0.58)
Passive Implausible	5.51	(0.73)

Analysis of variance indicated a main effect of meaning, with the two plausible conditions rated as being more plausible than the two implausible conditions ($F_1(1,23) = 729.732$, MSe = 0.435, $p < 0.001$; $F_2(1,36) = 455.161$, MSe = 1.118, $p < 0.001$), but no main effect of voice and no voice*plausibility interaction (all Fs < 1).

A separate analysis (F_2) was computed for a subset of the materials that remained from Ferreira and Stacey's original set (12 items in total) to see if the predicted error pattern would be visible in the original items. The F_2 means are presented in Table 3 below

Table 3: F_2 mean plausibility ratings for subset of 12 materials

Condition	Plausibility Rating	
	Mean	(SD)
Active Plausible	1.83	(0.63)
Active Implausible	5.32	(0.59)
Passive Plausible	1.71	(0.69)
Passive Implausible	5.47	(0.55)

Analysis of variance indicated the same pattern: a main effect of plausibility (F_2 (1,11) = 83.203, MSe = 1.012, $p < 0.001$) but no main effect of voice and no interaction (both *F*s < 1).

Of the 37 experimental item used in the study, 13 were eventually excluded as their plausibility ratings did not go far enough in the predicted direction and it was felt they would not be useful in future experiments that required plausible and implausible sentences (i.e., in one or both of a material's meaning conditions it had been rated within 1.5 rating points of the middle score rather than being judged substantially plausible or implausible. An ANOVA (F_2) was computed for this final subset of 24 items; the F_2 means are in Table 4.

Table 4: F_2 means for the final subset of 24 items

Condition	**Plausibility Rating**
	Mean (SD)
Active Plausible	1.72 (0.53)
Active Implausible	5.89 (0.76)
Passive Plausible	1.90 (0.63)
Passive Implausible	5.96 (0.68)

There was a main effect of plausibility (F_2 (1,23) = 576.701, MSe = 0.705, $p <$ 0.001), no main effect of voice (F_2 (1,23) = 2.002, MSe = 0.181, $p > 0.1$) and no interaction ($F < 1$). Thus, experiment 1 also served as a norming study to gather a set of 24 items whose two meaning conditions were reliably rated as plausible and implausible.

Discussion

The results of experiment 1 were in stark contrast to those it sought to replicate. Whereas Ferreira and Stacey had reported sizeable misinterpretation effects and reliable differences in the interpretations of active and passive sentences, our results show only high levels of correct interpretations in each condition. So no evidence of shallow processing and no evidence of the N-V-N interpretation strategy. In all three set of materials that we analysed (the original 37, the 24 selected by rating, and the 12 ‘Ferreira and Stacey’ materials) there was a reliable effect of the plausibility manipulation such that implausible sentences were judged implausible, and there was a reliable difference between the plausible and implausible conditions.

We will return to this disagreement between studies in the chapter's general discussion.

Experiments 2 and 3

Experiment 2 was a further attempt to replicate the finding of misinterpretation of syntactically non-ambiguous sentences. Having failed to find the predicted effect in experiment 1, we changed both the task and the methodology, and also included, in experiment 3, a syntactic load element to maximise the chances of finding systematic interpretation errors if they are present.

Ferreira (2003, experiment 1) used a thematic role assessment task (Bates, Devescovi, & D'Amico, 1999) to directly probe participants' interpretations of experimental sentences, and again found interpretation inaccuracies with sentences that were passive in form and implausible in meaning: when asked to name the agent of the sentences, participants were less accurate overall in response to passive sentences, and a significant voice*plausibility interaction indicated that they were less accurate with implausible sentences that were in the passive voice than implausible sentences in the active voice (76% correctly answered in the Passive Implausible condition vs. 99% in the Active Implausible condition).

While Ferreira (2003) presented the materials aurally, the present studies employed moving-window, self-paced reading. This allowed for measurement of reading times on individual words, and this method of reading is a fair analogue to hearing speech as the words are encountered one at a time with each word disappearing before the next one appears, i.e. no rereading is possible. We hoped that by looking at reading times we would begin to get an idea of how early the comprehension system detects that something is amiss in the implausible sentence cases. Given that people seem to misinterpret passive implausible sentences a quarter of the time, a lengthened reading time at the point at which the implausibility arises, relative to the plausible equivalent, would indicate that correct grammatical processing had in fact taken place. For example:

(1) *The thief was pursued by the policeman for over an hour*

(2) *The policeman was pursued by the thief for over an hour*

If the correct grammatical processing has taken place, and the verb *pursued* has assigned the theta role of theme to *the policeman* and the role or agent to *the thief*, then the processor should recognise the resulting interpretation as implausible and we might expect longer reading times on the underlined word in (2) than in (1). What would be particularly interesting would be if we observed this increase in processing time coupled with an ultimate misinterpretation – that is, if it appeared that the parser had assigned the theta roles correctly, but the comprehension system derived an interpretation at odds with the parser's output. This part of the investigation would hopefully begin to address the time course of grammatical and heuristic processing, via the question of when the grammar informs interpretation.

Experiment 2, henceforth the ‘low load’ experiment, involved the presentation of 24 of the materials from Experiment 1, presented in a word-by-word moving window format. All experimental and filler items were followed by a question that probed some thematic role in the sentence. The experimental items were always probed for either the agent or the patient role. Experiment 3 – the ‘high load’ experiment – used the same format but the materials were embedded in a longer sentence that increased syntactic load. Having failed to find the predicted effect in experiment 1, it seemed possible that putting the comprehension system under greater strain might be more likely to reveal the effect if it was actually present.

Experiment 2: Low Load Experiment

Method

Participants

24 participants were recruited from the University of Glasgow population and received payment or course credit for their participation. All were native English speakers. All were naïve to the design and aims of the study.

Materials

There were 24 experimental materials interspersed with 48 filler items. Experimental items were simple declarative sentences describing transitive events. They were ‘reversible’, in that the arguments of the main verb could be exchanged without resulting in an anomaly, but they were ‘biased’ in that one arrangement was more plausible than the other, for example, *the policeman pursued the thief for over an hour* (plausible) vs. *the thief pursued the policeman for over an hour* (implausible). The item could appear in either active or passive voice; when this was crossed with the plausibility variable the following four conditions resulted:

1. **Active plausible** *The policeman pursued the thief for over an hour*
2. **Active implausible** *The thief pursued the policeman for over an hour*
3. **Passive plausible** *The thief was pursued by the policeman for over an hour*
4. **Passive implausible** *The policeman was pursued by the thief for over an hour*

In processing terms the critical word is the second NP, as this is the earliest point at which the implausibility can be detected in the Implausible conditions.

The necessary semantic property of the experimental items, namely plausibility, had been established in experiment 1.

The questions that followed experimental and filler items probed one of the thematic roles in the sentence, using the thematic role task employed by Ferreira (2003). The experimental items were always probed for either the agent ('actor') or the patient ('acted-on'); 12 were probed for agent and 12 were probed for patient, so each participant made 12 agent decisions and 12 patient decisions. The fillers were modelled on those used by Ferreira (2003, experiment 1) and were of 4 types: 12 probed the colour of some object in the sentence, 12 asked about the main action in the sentence, 12 asked for the location in which the action described in the sentence took place, and 12 asked about the time at which the action took place. The question was presented after the sentence had been read and was in a two alternative, forced-choice format. Following the example above, the question might appear:

'Actor? Thief Policeman'

The correct answer appeared on the left hand side 50% of the time, and the right hand side 50% of the time. The order of items and fillers was randomised and the items in their four conditions were split between four lists using latin square rotation, such that each list contained every item but in only one of its conditions. Each participant viewed only one list and each list contained the same random order of materials.

Apparatus

The experiment was run using DMDX experimental software[1] on a Dell Optiplex GX270 personal computer. Participants paced themselves through the experiment and made their responses to the questions using a Logitech Dual Action Game Pad.

[1] Software programmed by Jonathan Forster at the University of Arizona. DMDX is a member of the DMASTR family of experimental software developed at Monash University and at the University of Arizona by K. I. Forster and J. C. Forster.

Procedure

An experimental session began with the experimenter reading instructions to the participant. Participants were told they would be reading a number of sentences one word at a time and at a rate determined by them. The experimenter presented written examples of each question type and indicated the corresponding correct answer. Participants were then introduced to the Logitech game pad. One button controlled the presentation of the words and this would be operated by either the left or right index finger depending on whether the participant was left or right-handed (this was ascertained by the experimenter). The questions were answered using two buttons operated by the left and right thumbs: if the participant thought the alternative presented on the left was correct, he pressed the left button, and likewise he pressed the right hand button if he thought the answer on the right was correct.

Participants were seated approximately 60cm from the computer screen and completed a practice session consisting of 16 items with thematic role judgement questions. Three of the items probed for the agent ('ACTOR'), 3 probed for the patient/theme ('ACTED- ON'), 3 probed for action ('ACTION'), 2 probed for colour ('COLOUR'), 2 probed for location ('WHERE') and 2 probed for time ('WHEN'). The experimenter observed the participant's responses during this phase, watching for systematic errors and answering any queries. None of the participants experienced particular difficulties in following the instructions or answering the questions that might have excluded them from participation.

Design

The experiment used a 2x2, within-participants design. Each sentence could be either active or passive in voice, plausible or implausible in meaning. (Participants answered either agent or patient/theme questions, but this factor is collapsed for analysis purposes unless otherwise stated.) Three dependant measures were taken. The focus was on accuracy in answering the questions (and thus the ultimate interpretation), but time taken to respond to the question and reading times on individual words were also analysed.

Results

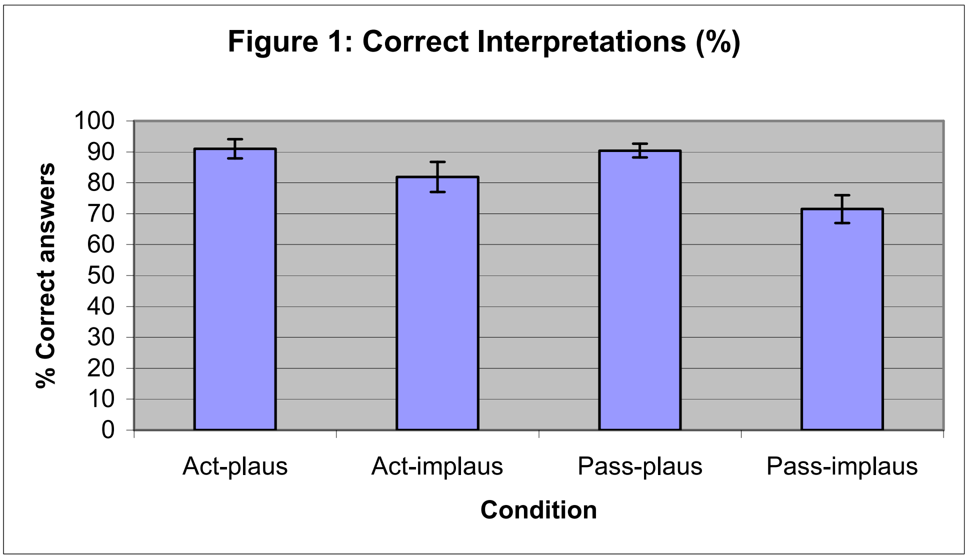
Two analyses of variance were computed for each analysis: one that treated participants as a random variable (F_1) and one that treated items as a random variable (F_2). Except for the analysis of correct answers by question type, analyses collapse across the question type variable.

Accuracy data

We first looked at the percentage of correct answers in each condition. The accuracy means are presented in Table 5 and Figure 1 below.

Table 5: Mean percentage of correctly answered questions in each condition

Condition	**% Correct**	
	Mean	(SD)
Active Plausible	90.97	(13.88)
Active Implausible	81.94	(18.98)
Passive Plausible	90.28	(13.83)
Passive Implausible	71.53	(22.78)



Analysis indicated a main effect of plausibility such that questions following plausible sentences were correctly answered more often than implausible sentences ($F_1(1,23) = 15.8, p < 0.001$; $F_2(1,23) = 10.7,\ p < 0.01$). There was no main effect of voice ($F_1(1,23) = 2.8, p < 0.1$; $F_2(1,23) = 2.9, p > 0.09$) and no voice*plausibility interaction ($F_1(1,23) = 1.7, p > 0.2$; $F_2(1,23) = 2.3, p > 0.1$). The means trend shows poorest performance in the passive implausible condition, although this trend was not borne out by a significant interaction. As we had predicted a difference between the two Implausible conditions, we carried out a direct comparison. However, a one-way ANOVA indicated that there was only a marginal difference by items (F_1 (1,23) = 2.929, MSe = 444.595, $p > 0.1$; F_2 (1,23) = 3.742, MSe = 347.977, $p = 0.066$). It seems that while interpretation was affected by schematic knowledge, the effect was the same for both active and passive implausible sentences.

Decision time data

Next we looked at the time taken to answer the questions. Although we had failed to find the predicted effect with passive implausible sentences in the interpretation data, perhaps the effect would be evident in the time it took participants to respond to the questions. The mean decision times (milliseconds) are presented in Table 6.

Table 6: Mean decision times (milliseconds)

Condition	**Decision Time (msec)**	
	Mean	(SD)
Active Plausible	2228	(615)
Active Implausible	2570	(818)
Passive Plausible	2938	(1075)
Passive Implausible	3039	(1134)

Analysis of variance indicated that there was a main effect of voice (F_1(1,23) = 23.7, $p < 0.001$; F_2(1,23) = 8.3, $p < 0.05$) with slightly longer decision times for passive sentences. There was a marginal effect of plausibility, with a tendency for longer decision times after implausible sentences ($F_1(1,23) = 4.2$, $p = 0.051$; $F_2(1,23) = 3.2$, $p = 0.086$), and no voice*plausibility interaction ($F_1(1,23) = 1.2$, $p > 0.2$; $F_2(1,23) = 1.2$, $p > 0.2$). There was therefore no strong evidence of particular difficulty with the passive implausible sentences as reported by Ferreira (2003).

Reading time data

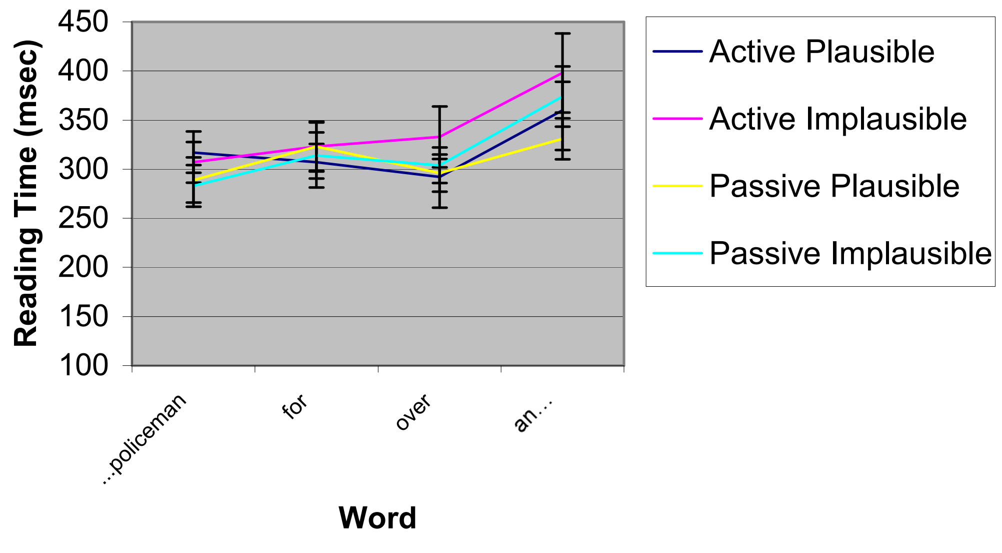
If the comprehension system is correctly applying the rules of grammar, we might expect to find lengthened reading times at the point at which an implausibility arises. For example, in the sentence *the policeman was pursued by the thief for over an hour*, the word *thief* is the point at which the correct interpretation jars with world knowledge (thieves do not normally pursue policemen). We looked at reading times on this critical word and, to allow for spillover effects, we also analysed reading times on the following three words[2]. Mean reading times for the critical word and the following three words are presented in Table 7 and Figure 2 below. Reading times less than 100ms and greater than 4000ms were excluded.

[2] One item was excluded from the Critical Word + 3 analysis as it had only two words following the critical word.

Table 7: Mean reading times (in milliseconds) on the critical word and the following three words

Condition	Decision Time (msec)							
	Critical Region		Critical Region + 1		Critical Region + 2		Critical Region + 3	
	Mean	(SD)						
Active Plausible	317	(106)	307	(91)	292	(89)	360	(141)
Active Implausible	307	(101)	323	(125)	333	(152)	398	(198)
Passive Plausible	289	(113)	323	(121)	296	(92)	331	(102)
Passive Implausible	283	(105)	314	(115)	304	(88	374	(151)

Figure 2: Reading times (msec) on the critical word and following 3 words



Analysis of variance indicated a main effect of voice, significant by participants, at the critical word with longer reading times on this word in passive sentences (F_1(1,23) = 12.905, MSe = 1221.400, $p < 0.005$; F_2 (1,23) = 4.027, MSe = 3913.766, $p > 0.2$). There were no other significant effects on this word ($F < 1$ or $p > 0.2$). There were no significant effects at all on the first word following the critical region ($F < 1$ or $p > 0.2$). At two words after the critical word there was a main effect of plausibility (significant by participants), with longer reading times in implausible sentences ($F_1(1,23) = 6.799$, MSe = 2115.869, $p < 0.05$; $F_2(1,22) = 3.270$, MSe = 4397.017, $p = 0.084$). There was no main effect of voice (both Fs < 1) and no interaction ($F_1(1,23) = 1.278$, MSe = 4933.314, $p > 0.2$; $F_2(1,23) = 1.306$, MSe = 4827.246, $p > 0.02$). The same pattern of effects was found on the third word after the critical word, which was also the last word in the sentence: a main effect of plausibility, significant by participants, with longer reading times in implausible sentences ($F_1(1,22) = 5.160$, MSe = 7801.300, $p < 0.05$; $F_2(1,23) = 3.950$, MSe = 11108.421, $p < 0.06$). There was no main effect of voice and no interaction ($F < 1$ or $p > 0.2$).

It seems that readers are slowing down on the second argument in passive sentences, perhaps reflecting a greater overall processing cost for passive constructions. The effect of the plausibility manipulation was not evident at this point. The plausibility manipulation registered two and three words downstream, when readers slowed down in response to an implausible sentence. There was no interaction however, indicating that the slow-down was not affected by whether the implausible sentence was active or passive in voice.

Reading times on incorrectly answered trials – sparse data problem.

To find lengthened reading times at, or just after the point at which the implausibility is established, would indicate that correct grammatical processing had in fact taken place. If this effect could be observed on just those trials that had been answered (i.e. interpreted) incorrectly, then this would be evidence that syntactic interpretations can be constructed online without informing the final interpretation. Unfortunately this analysis was not possible with the low load data: there simply were not enough participants who had given incorrect answers in all four conditions (there were only four participants who were eligible to be included in an analysis) and so analysis of variance on the reading times was not possible.

Agent vs. Patient/theme answers.

We analysed the percentage of correct answers by question type to see if readers had struggled particularly with assigning either the agent or theme role. But there was no main effect of question type (both *F*s < 1) and question type did not interact with either voice (both *F*s < 1) or plausibility ($F_1(1,23) = 2.003$, MSe = 288.849, $p > 0.1$; $F_2 < 1$). Thus, participants did not have special difficulty with assigning either the agent or the theme role. This is contrary to Ferreira (2003), who found that participants had significantly more difficulty assigning the theme role than the agent role.

Discussion

The main findings from Experiment 2, then, are that readers were less accurate in a thematic role judgement task when responding to sentences that described implausible events, compared with plausible events. The nature of the misinterpretations – providing answers indicative of pragmatic normalisation – reflects the influence of schematic knowledge. However, this misinterpretation effect was not influenced by whether the sentence was active or passive, and there was no especial difficulty associated with passive implausible sentences. The decision times indicated some relative difficulty in responding to passive sentences compared with active sentences, and implausible sentences took only marginally longer to respond to than plausible sentences. Analysis of reading times on individual words revealed a slow-down on the second argument (the critical word) in passive sentences relative to active sentences. Effects of the plausibility manipulation were only evident downstream of the critical word, with a slow-down on the second and third words after the critical word in implausible sentences. Again, there was no particular difficulty associated with passive implausible sentences, and no evidence that grammatical constraints were not being applied early in the interpretation process.

Experiment 3: High Load Experiment

Experiment 3 attempted to draw out any systematic misinterpretations of passive implausible sentences (relative to implausible actives) by embedding the materials from experiment 2 in a more complicated sentence frame. Accuracy levels were generally high in experiment 2 and participants might be more likely to misinterpret passive implausible sentences, relative to the other sentence types, if there was an added comprehension and memory strain. Recall that Garnham and Oakhill (1987) reported a drop in accuracy when their critical phrase was distanced from the comprehension task via an additional adjunct phrase. Therefore, the 24 items from experiment 2, in their four conditions, were embedded in complex sentences and presented to participants in another self-paced reading experiment, with the same thematic role judgement task used to directly probe interpretations.

Method

Participants

24 participants were recruited from the University of Glasgow population and received payment or course credit for their participation. None of the participants had participated in experiments 1 or 2. All were native English speakers.

Materials

The 24 items from experiment 2 were embedded in a syntactically complex sentence frame borrowed from Eastwick and Phillipps (1999, experiment 3). An example of a new material in its four conditions is given below in Table 8:

Table 8: Example experimental item for experiment 3

Condition	Experimental Sentence
Active Plausible	*The jury heard that the testimony revealing that the policeman pursued the thief for over an hour should not influence their decision*
Active Implausible	*The jury heard that the testimony revealing that the thief pursued the policeman for over an hour should not influence their decision*
Passive Plausible	*The jury heard that the testimony revealing that the thief was pursued by the policeman for over an hour should not influence their decision*
Passive Implausible	*The jury heard that the testimony revealing that the policeman was pursued by the thief for over an hour should not influence their decision*

The 24 experimental items were interspersed with 48 filler items – these were taken from experiment 2 and modified to be of a similar length and complexity to the high load experimental items.

As in experiment 2, all items were probed for a thematic role, with the fillers being probed in the same proportion for Time, Location, Action and Colour. Twelve experimental items were probed for Agent and 12 were probed for Patient/theme. The format was again two alternative, forced choice, and in the case of the experimental items, alternatives were always drawn from the 'experimental clause' (that is, the original sentence from experiment 2). So in the example given above, the question might read:

'Actor?   Thief   Policeman'

Apparatus

The new materials were presented using DMDX presentation software on the same Dell Optiplex GX270 computer. Participants again used a Logitech game pad to pace themselves and to record their responses.

Procedure

The procedure was identical to that in experiment 2. The items in the practice trials were adapted to be of similar length and complexity to the experimental items and fillers. Otherwise the sessions were run in exactly the same manner. No participants experienced difficulties during either the practice or experimental sessions.

Results

Analyses will again collapse across the question type variable unless otherwise stated.

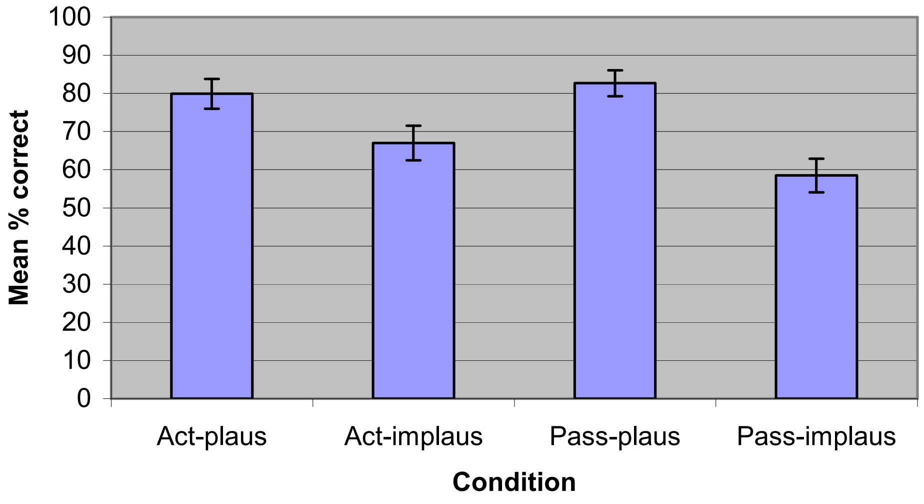
Interpretations: Percentage of Correct answers

We will first look at the percentage of correct answers in each condition. The mean percentages are in Table 9 and Figure 3 below.

Table 9: Mean % correct answers by condition

Condition	**% Correct**	
	Mean	(SD)
Active Plausible	79.86	(19.02)
Active Implausible	66.96	(22.08)
Passive Plausible	82.74	(16.66)
Passive Implausible	58.53	(21.40)

Figure 3: Correct answers (%)



As the pattern suggests, participants in the high load experiment were less accurate overall compared with the low load experiment.

Analysis indicated no main effect of voice ($F_1 < 1$; $F_2(1,23) = 1.117$, MSe = 259.159 $p > 0.3$), but a main effect of plausibility, with greater accuracy in the plausible conditions ($F_1(1,23) = 23.192$, MSe = 356.147 $p < 0.001$; $F_2(1,23) = 28.000$, MSe = 324.074, $p < 0.0001$). There was also a significant interaction, indicating that there was lowest accuracy in the passive implausible condition (F-$_1$(1,23) = 5.344, MSe = 143.603, $p < 0.5$; F_2(1,23) = 3.218, MSe = 291.365, $p = 0.086$). A planned comparison confirmed that there was no difference between the means of the two plausible conditions (both Fs < 1), and that there was a significant difference, by items only, between the means of the two Implausible conditions, such that accuracy was lowest in the Passive Implausible condition (F_1(1,23) = 2.883, MSe = 286.012, $p > 0.1$; F_2(1,23) = 4.715, MSe = 240.340, $p < 0.5$).

This points to a replication of Ferreira's (2003) central finding: that participants would often misinterpret passive sentences that were implausible, relative to the active equivalent. In this case, participants misinterpreted such sentences approximately 41% of the time.

Decision Times

We again analysed the time taken to answer the question in each condition, to see if any sentence type was proving particularly difficult to interpret correctly. The mean decision times are presented in Table 10.

Table 10: Mean decision (msec) times in the high load experiment

Condition	**Decision Time (msec)**	
	Mean	(SD)
Active Plausible	2453	(1479)
Active Implausible	3329	(1569)
Passive Plausible	3324	(1056)
Passive Implausible	3459	(1295)

There was no main effect of either voice (both Fs < 1) or plausibility (both Fs < 1), and there was no significant voice*plausibility interaction ($F_1 < 1$; $F_2 (1,23) = 1.016$, MSe = 620028.366, $p > 0.3$). Apparently none of the sentence types took any longer to answer than any of the others.

Reading time data

Reading times less than 100 milliseconds and greater than 4000 milliseconds were excluded from the analysis. The mean reading times (milliseconds) for the critical word and the following three individual words are in Table 11 below.

Table 11: Mean reading times for the critical word and the following three words.

Condition	**Reading Time (msec)**							
	Critical Region		**Critical Region + 1**		**Critical Region + 2**		**Critical Region + 3**	
	Mean	(SD)						
Active Plausible	374	(103)	423	(134)	383	(108)	367	(99)
Active Implausible	379	(132)	389	(112)	382	(96)	392	(122)
Passive Plausible	371	(136)	382	(120)	378	(114)	355	(114)
Passive Implausible	375	(171)	384	(131)	403	(124)	385	(129)

Analysis of variance was computed for the reading times on each of the words. There were no significant effects at all on the critical word (all Fs < 1). On the first word following the critical word, there was no effect of voice (F_1 (1,23) = 2.775, MSe = 4425.445, $p > 0.1$; F_2 (1,23) = 2.106, MSe = 5831.540, $p > 0.1$), no effect of plausibility (F_1 (1,23) = 1.077, MSe = 5772.694, $p > 0.3$; $F_2 < 1$) and no interaction (F_1 (1,23) = 1.827, MSe = 4201.211, $p > 0.1$; F_2 (1,23) = 2.061, MSe = 3723.541, $p > 0.1$). On the second word after the critical word, there were no significant effects at all (all Fs < 1). On the third word downstream, there was no effect of voice (both Fs < 1) and no interaction (both Fs < 1). But there was an effect of plausibility, significant by participants and marginal by items, with longer reading times in the implausible conditions (F_1 (1,23) = 5.615, MSe = 3295.301, $p < 0.05$; F_2 (1,23) = 3.126, MSe = 5919.228, $p = 0.090$).

Thus, participants slowed down three words downstream of the critical word when the sentence was implausible, a finding also observed with the low load materials in experiment 2.

Reading times on incorrectly answered trials.

To look for effects of any implicit or unconscious detection of implausibility, and thus automatic grammatical processing, we re-analysed reading times on the same words when the participants had answered the question incorrectly. Recall that this analysis was not possible in the low load experiment due to sparse data (N=4). Due to the more demanding nature of the high load materials, however, we had 10 participants who had made incorrect responses in all four conditions and were thus able to run analyses on their data. However, the participants were not evenly distributed across groups, i.e. this subset did not view the four stimulus list in equal proportions. Four participants viewed list 1, three viewed list 2, two viewed list 3 and only one participant viewed list 4. Results of this analysis must therefore be considered tentative. As this type of analysis is partial, we will report analysis of participant means only.

The mean reading times for the relevant word in the 'incorrect trials' are in Table 12 below. Reading times less than 100msec and greater than 4000msec were excluded.

Table 12: Mean reading times on the incorrectly answered trials (N=10)

Condition	**Reading Time (msec)**							
	Critical Region		**Critical Region + 1**		**Critical Region + 2**		**Critical Region + 3**	
	Mean	(SD)						
Active Plausible	352	(74)	404	(212)	363	(105)	376	(114)
Active Implausible	412	(130)	382	(105)	451	(260)	374	(143)
Passive Plausible	383	(222)	358	(81)	430	(110)	327	(115)
Passive Implausible	343	(101)	369	(137)	378	(98)	420	(246)

In the critical region itself there were no significant main effects of voice or plausibility (both *F*s < 1) and no interaction ($F_1(1,9) = 2.581$, MSe 9721.300, $p > 0.1$). One word after the critical word there was also no main effect of voice ($F_1(1,9) = 2.255$, MSe = 3820.074, $p > 0.1$), no main effect of plausibility ($F_1 < 1$) and no interaction ($F_1 < 1$). On the second word after the critical word there was no main effect of voice and no main effect of plausibility (both *F*s < 1). There was a marginal interaction, however ($F_1(1,9) = 4.774$, MSe = 10265.753, $p = 0.57$), with the trend for the longest reading times on this word to be in the active implausible condition (and thus not in the predicted pattern). On the third word downstream from the critical word, there was no main effect of voice ($F_1 < 1$), no main effect of plausibility ($F_1(1,9) = 1.368$, MSe = 15233.600, $p > 0.2$) and no interaction ($F_1(1,9) = 1.892$, MSe = 12079.268, $p > 0.2$).

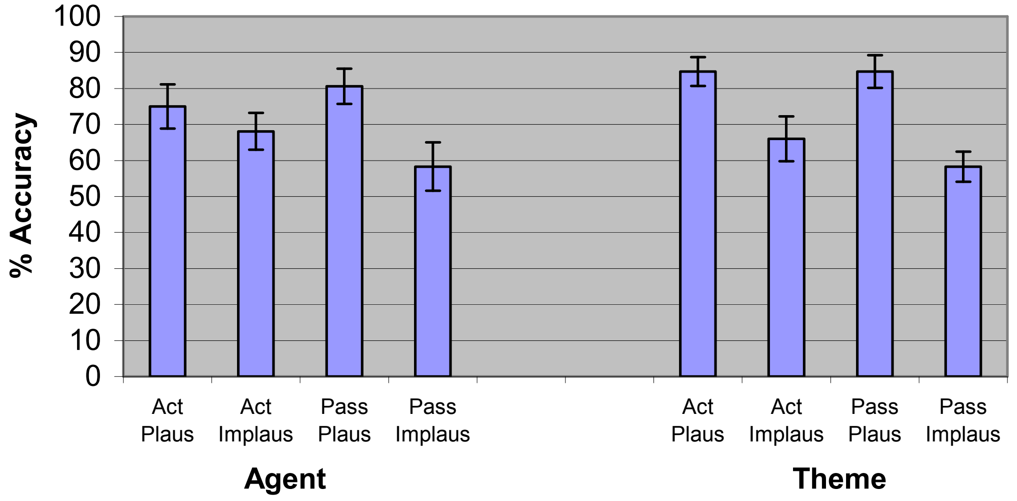
Question type

As with the low load experiment, we also looked at the effect of question type to see whether participants had struggled particularly with the assignment of one of the thematic roles. The means for percentage of correct answers by condition, including the question type variable, are presented in Table 13 and Figure 4 below:

Table 13: Correct answers (%) by condition, question type variable included.

Condition	% Correct	
	Mean	(SD)
Ag. Act Plaus	75.00	(29.90)
Ag. Act Implaus	68.06	(25.02)
Ag. Pass Plaus	80.56	(23.91)
Ag. Pass Implaus	58.33	(32.97)
Pa. Act Plaus	84.72	(19.61)
Pa. Act Implaus	65.97	(30.39)
Pa. Pass Plaus	84.72	(21.93)
Pa. Pass Implaus	58.33	(20.41)

Figure 4: Accuracy (%) by condition, with question type



The three-way ANOVA indicated that there was no main effect of question type (both *F*s < 1) and question type did not interact with either voice (both *F*s < 1), or plausibility ($F_1(1,23) = 1.199$, MSe = 639.524, $p > 0.2$; $F_2(1,23) = 1.812$, MSe = 313.026, $p > 0.1$). There was no three way interaction (both *F*s < 1).

Discussion

Experiment 3 replicated many effects from experiment 2. The first point to note is that while readers again produced normalised interpretations, accuracy appears to suffer even further in the presence of increased syntactic load. The second point is that only under these conditions of heavy load do we see an effect of the voice manipulation on interpretation accuracy: readers were indeed poorer at interpreting implausible passives compared with implausible actives. Lastly, under these taxing load conditions, we still see clear evidence of the online computation of meaning based on syntactic structure. The disruption effects were slightly delayed relative to the low-load experiment, but were still evident prior to sentence wrap-up.

We will now turn to another set of analyses, performed on the pooled data from both the low and high-load experiments.

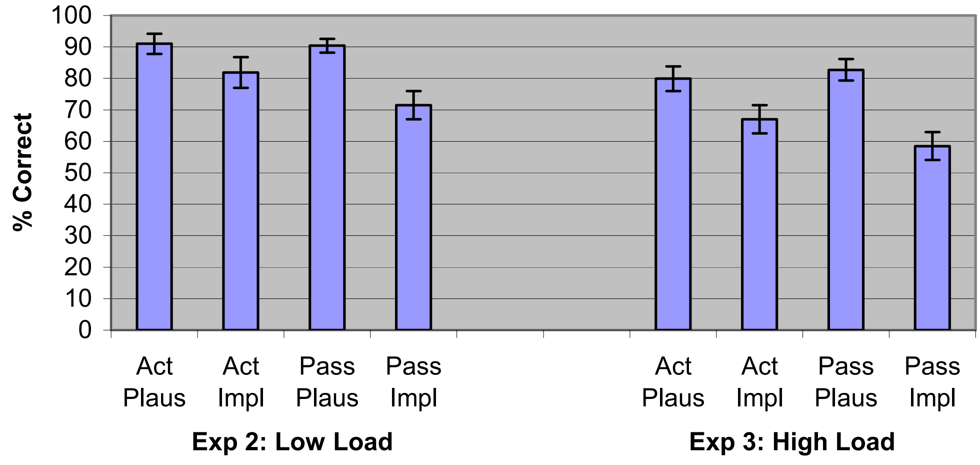
Combined analysis of data from experiments 2 and 3

Due to the identical designs in experiments 2 and 3, we were able to perform analyses that pooled the data from both. Recall that the materials from experiment 2 were embedded, unchanged, in a larger sentence frame, and the questions in experiment 3 always probed exactly the same arguments as the questions in experiment 2. These combined analyses have an advantage of increased power as each observation in experiment 2 has a corresponding observation in experiment 3 (with the exception of the analysis of reading times on the third word after the critical word, as one material in experiment 2 had only two words after the critical word).

Percentage correct answers

For the purposes of comparison, the accuracy data from experiments 1 & 2 are presented in Figure 5.

Figure 5: Experiments 2 & 3 Accuracy Comparison



What is immediately obvious is a trend for poorer accuracy in the high load experiment. An analysis of variance, performed with Load (i.e. experiment) included as a factor, indicated that this main effect of load was significant ($F_1(1,46) = 10.711$, MSe = 611.494, $p < 0.01$; $F_2(1,23) = 40.540$, MSe = 175.869, $p < 0.001$). There was a main effect of plausibility, with higher accuracy in the plausible conditions ($F_1(1,46) = 39.954$, MSe = 325.187, $p < 0.001$; $F_2(1,23) = 28.675$, MSe = 464.976, $p < 0.001$). There was also a significant voice*plausibility interaction ($F_1(1,46) = 5.612$, MSe = 238.685, $p < 0.05$; $F_2(1,23) = 4.874$, MSe = 303.945, $p < 0.05$). Planned comparisons indicated no significant differences between the means of the plausible conditions (both Fs < 1), and a significant difference between the means of the implausible conditions, by participants and items, such that accuracy was lowest in the Passive Implausible conditions ($F_1 = 5.874$, MSe = 362.927, $p < 0.05$; $F_2 = 6.608$, MSe = 184.116, $p < 0.05$). Again, this appears to indicate systematic misinterpretation of this type of sentence, supporting the findings of Ferreira and Stacey (unpublished) and Ferreira (2003).

Decision Times

Next we did a combined analysis of the time taken to answer the questions. There was a main effect of load, with longer decision times in the high load experiment ($F_1(1,46) = 5.849$, MSe = 4260967.136, $p < .05$; $F_2(1,23) = 40.058$, MSe = 664534.547, $p < 0.001$). There was a main effect of voice, significant by participants, and seemingly driven by the means from the low load experiment, with longer decision times after passive sentences ($F_1(1,46) = 8.069$, MSe = 3992070.134, $p < 0.01$; $F_2(1,23) = 3.011$, MSe = 1597551.135, $p = 0.096$). There was also a load*voice interaction, significant by participants, (F_1(1,46) = 8.024, MSe = 517607.433, $p < 0.05$; F_2(1,23) = 3.255, MSe = 1138753.880, $p = 0.084$). This again appears to be driven by the main effect of voice already observed with the low load materials. There was, however, no significant voice*plausibility interaction (both Fs < 1).

Reading times on the critical word and following three words

As there were no departures from the results of the individual analyses, and for the sake of brevity, the results of the combined reading time analyses will not be presented in full. But for the sake of completeness we will present results for the effect of load on each region.

At the critical region itself the main effect of load was significant, with longer reading times in the high load experiment (F_1(1, 46) = 5.850, MSe = 62822.561, $p < 0.05$; F_2(1, 23) = 42.024, MSe = 8884.788, $p < 0.001$). There were no other significant main effects or interactions (all ps > 0.1). At the first spillover word the effect of load was again significant, with longer times in the high load experiment (F_1(1, 46) = 6.510, MSe = 44585.735, $p < 0.05$; F_2(1, 23) = 46.926, MSe = 6180.748, $p < 0.001$). There was an interaction between Load and Voice and Experiment, but it was significant by items only (F_1(1, 46) = 2.600, MSe = 4215.033, $p > 0.1$; F_2(1, 23) = 6.868, MSe = 1595.914, $p < 0.05$). There were no other significant effects (all ps > 0.1). On the second spillover word load was again significant in the same direction (F_1(1, 46) = 8.293, MSe = 42785.648, $p < 0.01$; F_2(1, 23) = 55.194, MSe = 6428.555, $p < 0.001$). The effect of plausibility was also significant by participants, indexing anomaly detection (F_1(1, 46) = 6.637, MSe = 4427.380, $p < 0.05$; F_2(1, 23) = 2.568, MSe =

11443.064, $p > 0.1$). There were no other significant main effects or interactions (all ps > 0.2). By the third spillover word only the main effect of plausibility was significant ($F_1(1, 46) = 9.985$, MSe = 5582.001, $p < 0.05$; $F_2(1, 23) = 9.951$, MSe = 5691.055, $p < 0.01$; all other ps > 0.1). (A four-way analysis including region as a factor (along with Voice, Plausibility and Experiment) indicated that the 4-way interaction was not significant: both Fs < 1.) We can conclude that readers were reading these regions more slowly in the high load experiment, most likely taking a cautious approach with the more difficult sentences.

Reading times on incorrectly answered trials

Combining the small samples from both experiments gives us a sample size of 14 – still lacking in power but worth analysing. An ANOVA, collapsing across load, was performed on reading times on the critical word and the following 3 words as before. On the critical word there were no significant effects of voice of plausibility (both Fs < 1) and no interaction ($F_1(1,13) = 1.678$, MSe = 7989.046, $p > 0.2$). One word downstream there was no effect of voice ($F_1(1,13) = 2.986$, MSe = 3218.733, $p > 0.1$), no effect of plausibility and no interaction (both Fs < 1). Two words downstream there were again no significant effects ($F < 1$ or $p > 0.2$). On the third post-critical word there were no effects of voice of plausibility ($F < 1$ or $p > 0.1$). However, there was a marginal voice*plausibility interaction ($F_1(1,13) = 3.723$, MSe = 9744.974, $p = 0.076$). Comparisons to test for plausibility differences within voice conditions revealed no differences between the two active conditions ($F_1 < 1$) and none between the passive conditions (F_1 (1,13) = 2.855, MSe = 24163.029, $p > 0.1$). Table 14 gives the means for reading times on this word: as can be seen, the Passive Implausible condition, while a full

100 milliseconds greater than the Passive Plausible condition, has a standard deviation more than twice that of the Passive Plausible condition.

Table 14: Reading Times on the third post-critical word, incorrectly answered trials

Condition	**Reading Time (msec)**	
	Mean	(SD)
Active Plausible	364	(123)
Active Implausible	362	(157)
Passive Plausible	319	(108)
Passive Implausible	418	(235)

This analysis therefore offers only a suggestion that, in the case of an ultimately incorrect interpretation, the correct, implausible meaning had been successfully constructed online. (An analysis including load, i.e. experiment, as a factor indicated no significant main effect of load and no load interactions (all *F*s < 1).)

Percentage of correct answers by question type

None of the analyses that included question type as a factor indicated any significant main effects of question type (both *F*s < 1); nor did question type interact with any other factor. While the individual analysis of question type for each experiment did entail an issue of power (only half the items in each experiment probed for each role), this more powerful analysis rectifies the problem. We can say, then, that our results are a departure from Ferreira's, in that out participants didn't display particular difficulty assigning either the agent or theme roles.

General Discussion:

Main findings of experiments 1 – 3

Our plausibility judgement task failed to replicate the findings of Ferreira and Stacey's Experiment 1 (unpublished). There was no evidence at all that participants misinterpreted implausible sentences, and certainly no evidence of a tendency to rate passive implausible sentences as being plausible, relative to the active implausible sentences. Accuracy, as indicated by plausibility judgements, was extremely high. It is not clear why this replication should have failed as rereading was allowed in both experiments, and participants in both experiments had the sentences still available to them as they made their judgements. The presentation methods were different (questionnaire format vs. onscreen reading) but the task conditions were fundamentally the same. There was also very little difference in terms of material load, as the Ferreira and Stacey study used a total of only 84 sentences including fillers. It can therefore be simply put that the results here offer a robust challenge to Ferreira and Stacey's results, and suggest that if the comprehension system does rely on heuristics to interpret pragmatically challenging sentences, it is not at all clear that it will do so under these reading conditions. An explanation as to the radically different results between the two studies can, at this point, only lie in speculation as to differences between the two participant groups.

Turning to experiments 2 and 3, some of the looked-for interpretation effects were observed. In the low load experiment participants were significantly poorer at interpreting sentences when their syntactically licensed meaning was implausible. In the high load experiment, this same behaviour was evident, and also a significant tendency to be less accurate when interpreting implausible sentences when they were in the passive rather than active voice. The analysis of the combined data from experiments 2 and 3 indicated the same effect. The failure to find the active/passive distinction in experiment two differs from Ferreira's results. One possible explanation relates to the differing material loads: Ferreira's participants heard a total of 216 materials including fillers, compared with our 72. This may have produced fatigue in the participants, resulting in comprehension functioning below normal capacity.

Looking at decision times, the low load data and the combined data showed longer times overall for passive materials. There were no other effects of either voice or plausibility which might suggest that normalised interpretations – if applied after syntax-based ones – were applied late online, or were applied so quickly as to be immeasurably fast. (This question of when heuristic interpretations are applied will be taken up in the concluding chapter.)

In terms of whether the question probed the agent or the theme of the sentence, no effects of question type were observed, suggesting that participants did not struggle particularly with any one role assignment. This may be a power issue, as not all items were probed for both agent and theme roles, but the added power from the combined analysis, which showed the same results, doesn't make this a likely explanation. This is a further unexplained departure from the Ferreira studies, whose results showed a particular difficulty in assigning the theme role.

Considering reading times, the only effects to note were main effects of plausibility in the Low-load and Combined analyses at two words downstream of the critical word, when the two implausible conditions had longer reading times;

and in all three analyses (Low, High & Combined) at the third word after the critical word, when the same main effect of plausibility was observed. Thus, we have several robust effects showing that participants were applying syntactic constraints to interpretations made online. In experiment 3, when the trials that were answered incorrectly were looked at, there was a trend towards longest reading times in the passive implausible condition three words after the critical word, suggestive of correct syntactic interpretation. However, there were no significant effects – not especially surprising as these analyses contained data from few participants. The same analysis using reading time data combined from experiments 2 and 3 depicted the same means pattern at the third post-critical word, but also a marginal interaction pointing to a plausibility-related difference between the two Passive conditions. If real, this effect would indicate that participants had computed the correct meaning of the implausible sentences online, but later overridden them in favour of a more pragmatically suitable one either late online, or offline. But again, the low power in this analysis prevents us drawing any confident conclusions.

We have evidence that under challenging reading conditions (that offer a good analogue to speech) readers will misinterpret sentences whose correct meaning is implausible – that is, they will normalise them. Also, under conditions of high syntactic load, readers will not only normalise to a greater extent, but they will misinterpret passive implausible sentences to a higher degree than the equivalent sentence cast in active form. When the parser must assign roles in an atypical order (‘theme-verb-agent’) and the resulting meaning is implausible, it seems there is an increased tendency to generate an interpretation that is more in line with real world knowledge. But again, this tendency is only evident under particularly taxing processing conditions.

The studies presented here are ultimately supportive of Ferreira's findings on normalisation and the particular case of passives, but they strongly suggest limits to their validity and outline some reliable conditions under which syntax-based interpretations are likely to fail.

Chapter 3:

The Interpretation and Processing of Implausible Elliptical Verb Phrases

Introduction

Experiments 1-3 demonstrated that readers can be strongly influenced by pragmatic constraints even when interpreting sentences that are syntactically unambiguous. The differences between experiment 1 and experiments 2 and 3 indicate that this phenomenon is by no means a given, and isn't representative of the normal work of the comprehension system under all circumstances. Observed misinterpretations rates were, however, high, with accuracy as low as 58% in one condition of experiment 3. Analysis of the time course of interpretation showed that readers were indeed aware of the implausible nature of what they were reading, correctly building syntax-based interpretations online. There was some evidence, not conclusive, that readers who had correctly detected an implausible meaning online would go on to misinterpret the sentence, removing the implausibility by normalising it. But the data set was insufficiently powerful to enable a robust analysis of how readers process an implausible sentence prior to normalising it.

The present study is a further attempt to document normalisation and track the time course of interpretation of implausible sentences. The methodology will remain similar to that of experiments 2 and 3 and will use a type of material already known to illicit incorrect interpretations.

Garnham and Oakhill (1987) reported substantial rates of misinterpretation when readers had to interpret a construction that was not only very common, but widely considered to be easily understood. After reading an elliptical verb phrase (EVP) that had an implausible meaning, readers were only correct 75% of the time and their errors reflected a more plausible interpretation suggested by the meaning of the context.

An EVP is a construction whose interpretation depends not only on its own content but on the precise structure and meaning of the antecedent clause which precedes it. In an example given by Garnham and Oakhill, two sentences

John had praised Mary.
Mary had been praised by John.

both have the same truth conditions and hence essentially the same meaning. But when followed by an EVP, e.g. *Sally had too*, the interpretation of the EVP depends on the precise form of the antecedent sentence and will have a different meaning depending on which of our two examples it follows. Garnham and Oakhill's experiment tested the idea that EVPs may be difficult to comprehend after all, given the known problems with a comprehender's memory for surface form ("one of the best-established results in the psycholinguistic literature", p614). Their results showed that when the correct interpretation of an EVP was at odds with a more plausible scenario suggested by the context, readers were often unsuccessful at comprehension (see (3) for an example of Garnham and Oakhill's materials). Interpretation accuracy was even poorer as the distance between the EVP and its antecedent clause was increased via an adjunct phrase (accuracy fell to 61%).

(3) *The elderly patient had been examined by the doctor [during the ward round].*
The nurse had too

Garnham and Oakhill presented their materials in a segment-based self-paced-reading experiment and measured reading times on the EVP and decision times on the comprehension task. Reading and decision times indicated that when the EVP had an implausible meaning readers took longer to read the segment containing the EVP and longer to answer the question that followed. While indicating that the correct interpretation had been formed a significant proportion of the time, the reading time results mask the precise timing of the anomaly detection due to the segment-based presentation. Readers could have produced the correct interpretation first and then generated a normalised interpretation, or vice-versa; either could account for the lengthened reading and decision times if we understand them to reflect a conflict between opposing interpretations. A word-based analysis with an allowance for non-immediate detection of the anomaly (via spillover regions) would be necessary for detailed insight into the time course of the processing of this type of construction. Garnham and Oakhill also attempted an analysis of reading times contingent on response (correct/incorrect) but sparse data prohibited anything beyond a descriptive analysis. Mean reading times for the EVP suggested that a correctly answered Implausible trial was read for longer than an incorrectly answered one.

A look at the means for the Plausible and Implausible trials that were answered incorrectly suggests that reading times were longest in the Implausible condition. Again, robust conclusions were not obtainable, but this pattern suggested that the correct interpretation could be constructed for implausible EVPs but then ultimately be misinterpreted to describe a more plausible scenario – a pattern reminiscent of the account given of experiments 2 and 3 in the previous chapter.

Clearly, these results suggest a fruitful line of investigation. The present study aims to replicate the interpretation results of Garnham and Oakhill with a new set of materials presented word-by-word to allow a more detailed analysis of the time course of syntax-based interpretation. Rather than measuring reading time on a whole clause, the word-by-word format allows us to track interpretation in real time, with early and late online interpretation effects appearing as disruption at different points in the processing stream. Ideally, the new materials, tracked using this methodology, will allow a full response-contingent analysis to test the hypothesis that correct meanings are computed online but later overridden by semantic cues.

The design will also include two new conditions in which plausible and implausible EVPs are preceded by active antecedent clauses to allow further investigation of the interpretation of passives compared to actives under conditions likely to produce normalisation. Recall that, under the heavy syntactic load conditions of experiment 3, readers were poorer at interpreting implausible passives than implausible actives. The active conditions in the present study will allow us to investigate whether this particular difficulty associated with passives is restricted to interpretations of clauses in which the verb's arguments explicitly appear – which has been explained in terms of interference from the N-V-N heuristic – or whether the issue could be understood in simpler terms of memory for syntactic structure and the operation of a plausibility heuristic. Take the following example:

The old woman had been frightened by the mugger in the park yesterday morning. The thug had too according to the news report

In this passage, the NP 'the thug' is signalled by the syntax to be the co-theme of the opening sentence with 'the old woman'. A likely normalised interpretation of the EVP – that the thug had frightened the old woman – would be due to readers opting to assign 'the thug' the more plausible role of co-agent. Importantly, despite the antecedent sentence being passive in form, this outcome could not be explained by the NVN strategy. This strategy would generate the interpretations 'old woman frightened mugger' and 'thug frightened mugger', neither of which are semantically compelling. So a difference in the interpretations of implausible actives and passives similar to that seen in experiment 3 could not be accounted for, here, by passives being more susceptible to plausible interpretations suggested by the NVN heuristic. Any such observed difference in this experiment, therefore, could remove the need to explain poor performance with passives in strictly those terms.

A further aspect of this study will be the introduction of a measure of reading comprehension ability. As discussed in the introductory chapter, recent studies have employed tests such as the Nelson Denny Reading Comprehension Test and have suggested that less-skilled readers are more prone to normalisation than skilled readers (Hannon and Daneman, 2004, Daneman, Lennertz and

Hannon, 2007). But while these studies reported interpretation differences, processing analyses yielded nothing significant when reading skill was included (the data sets were small however). We have the opportunity, then, to replicate the differences between the skill levels in terms of interpretation, and also to extend the investigation of how the different skill levels deal online with implausible material. Could it be that less-skilled readers detect anomalies online as well as, or better than, their skilled counterparts, and exhibit their disadvantage at the decision stage? Or are less-skilled readers more prone to producing normalised interpretations because they do not, in these cases, ever produce the correct ones?[3]

Experiment 4

Method

Participants

32 participants were recruited from the University of Glasgow community and received payment or course credit for their participation. All participants were native English speakers, had normal or corrected-to-normal vision and had not been diagnosed with dyslexia.

[3] Results of individual differences analyses for this and all subsequent experiments will be reported in chapter 6.

Materials

There were 32 experimental materials interspersed with 64 filler materials. Each material consisted of two sentences. The first, opening sentence was in either the active or passive voice, and consisted of a transitive verb phrase followed by an adjunct phrase. The second sentence was an elided verb phrase (EVP) whose successful interpretation depended on the interpretation of the opening sentence. The EVP could have either a plausible or an implausible interpretation. This gives a 2x2 factorial design with two levels of the voice factor (Active and Passive) and two levels of the plausibility factor (Plausible and Implausible). An example of a material in each of its four conditions is given below in Table 15.

Table 15: Example Experimental Material

Condition	Material
Active Plausible	*The mugger had been frightening the old woman in the park yesterday morning.* *The thug had **too** according to the news report.* *Did the thug frighten the old woman?*
Active Implausible	*The mugger had been frightening the old woman in the park yesterday morning.* *The jogger had **too** according to the news report.* *Did the jogger frighten the old woman?*
Passive Plausible	*The old woman had been frightened by the mugger in the park yesterday morning.* *The jogger had **too** according to the news report.* *Did the mugger frighten the jogger?*
Passive Implausible	*The old woman had been frightened by the mugger in the park yesterday morning.* *The thug had **too** according to the news report.* *Did the mugger frighten the thug?*

For the purposes of analysis, the critical region of interest is the final word of the EVP, that is, the word *too*. This is the point at which the EVP is rendered either plausible or implausible and is thus the earliest point at which we could expect to see reading time differences caused by plausibility differences between conditions. In each material, the EVP was followed by a further 5 words. These were added to allow analysis of spillover effects, as normal reading may not be disrupted immediately upon encountering the critical word (as with experiments 2 and 3), and effects may instead be visible further downstream.

Each experimental item was followed by a question (2-alternative, forced choice) that directly probed the interpretation of the EVP, for example:

The mugger had been frightening the old woman in the park yesterday morning. The thug had too according to the news report.

Did the thug frighten the old woman?

Yes <> No

The thematic role judgement task was not used in this experiment, due to the differences in argument structure between the materials of experiments 1-3 and those used here. Within the straightforward declarative clauses of the earlier studies, forced binary choices were justified as there could only be one 'ACTOR' or 'ACTED-ON'. But within the critical passage (antecedent + EVP) there are two NPs that could be correctly identified as the actor or acted-on.

The design of the questions gave rise to some extra considerations. The answer to the questions, i.e. *Yes/No*, was balanced across items so that each condition was tested by an equal number of questions answering 'Yes' and questions answering 'No'. The correct answer to items 1-16 was 'Yes' and the correct answer to items 17-32 was 'No'. This produced an alternation of question types across conditions in terms of whether or not the question suggested an implausible event. The plausibility of the scenario contained in the question (i.e. a mugger frightening a thug) could conceivably result in an answering strategy based on the plausibility of the question rather than the actual item, but the plausibility of the question was also controlled across items (see Table 16 for the balance of plausible and implausible questions).

Table 16: Counterbalancing question types

Condition	Items	Question Type	Correct Answer
Active Plausible	1 -- 16	Question suggests plausible event	Yes
	17 -- 32	Question suggests implausible event	No
Active Implausible	1 -- 16	Question suggests implausible event	Yes
	17 -- 32	Question suggests plausible event	No
Passive Plausible	1 -- 16	Question suggests plausible event	Yes
	17 -- 32	Question suggests implausible event	No
Passive Implausible	1 -- 16	Question suggests implausible event	Yes
	17 -- 32	Question suggests plausible event	No

For items 17-32 which had the correct answer ‘No’, the order of plausible and implausible question types was reversed.

This phrasing of the questions to ensure the balance of answer-type resulted in differences across items in the position of the subject of the EVP when the question was formed. In items 1-16, in the active conditions, the subject of the EVP is in the subject position of the question, but in the passive conditions the subject of the EVP is in object position of the question. In items 17-32, the subject of the EVP was in the object position of the question in the active conditions, while in the passive conditions the subject of the EVP was also in the subject position of the question. This imbalance was unavoidable given the more serious problems envisaged with an imbalance in the plausibility of the question.

The fillers were designed as follows: Thirty-two fillers had opening sentences that were in the active voice, and 32 had opening sentences that were in passive voice. In each set of 32, 16 questions had the correct answer 'Yes' and 16 had the correct answer 'No'. In each block of 16 fillers, only 4 contained an EVP. This allowed a superficial similarity to the experimental materials without greatly increasing the participant's exposure to EVPs and thus limiting practise effects.

Plausibility Norming

Two attempts were made to gather plausibility data on the 32 materials. Despite a close modelling on the Garnham and Oakhill material set, and strong experimenter intuition, plausibility ratings did not conform to predictions. The first attempt presented readers with the materials as they would appear in the main study (antecedent + EVP) and asked for plausibility ratings. The second attempt involved presenting readers with sentences that made explicit the correct interpretation of the EVP in each condition, and gathering plausibility ratings on those. Neither attempt produced data reflective of the plausibility manipulations (compared with, say, the data from experiment 1), with ratings either very close to the 'neutral' score or crossing over it in the wrong direction.

Trusting to intuition, and mindful that Garnham and Oakhill did not gather plausibility norming data for their materials, it was decided to test the materials anyway. To preview the results, the interpretations and reading time data gathered in the experiment itself *did* strongly reflect the intended semantic properties of the 4 conditions. The difficulty in pre-rating the materials therefore points not to their inherent unsuitability, but to difficulties with gathering this type of ratings data.

There is no acknowledged formula or convention for eliciting a true judgement of plausibility, and investigators will ask questions such as, ‘how likely is this to happen?’, or ‘how much sense does this make?’, or simply ‘how plausible is this?’. There are difficulties attached to each. The first may suffer from confusion over whether one is being ask to judge from observed frequency, or being asked how easy it is to *imagine* something happening. Answers to the second question (which strongly confirmed intuition in experiment 1) could again suffer from a confusion between events in real or imaginary worlds, and could conceivably even elicit grammatical acceptability judgements. The third question has the problem of leaving it to the participant to define ‘plausible’. Although it may be standard practice to include a couple of example items to indicate to a participant the definition of plausibility you have in mind, a full material set may never be fully represented by two or three items, and subjective understandings of the concept of plausibility will come into play. In all cases then, the intuition used to create a set of implausible materials may be overruled only because of vagaries in the participant’s understanding of what is required of them, and this is not easy to rectify. There are a number of reasons why something may be unusual, and what may be unusual according to one criteria for judging unusualness may be acceptable according to another. It may also not be easy to say exactly why something is unusual. If a participant switches criteria at any point the result will be an inconsistent set of judgements. The practical course taken here, and subsequently validated by behavioural data, was to trust intuition in the face of the difficulties with gathering reliable judgements.

Apparatus

The experiment was run using DMDX experimental software[4] on a Dell Optiplex GX270 personal computer. Participants used a Logitech Dual Action Game Pad to pace themselves through the experiment and to make their responses to the questions.

Procedure

The experimental session began with some brief verbal instructions from the experimenter about the broad nature of the task. Prior to reading fuller instructions on-screen, participants were familiarised with the Logitech game pad. Having been told that they would be reading sentences one word at a time and at a rate determined by them, participants were shown the button that controlled word presentation. This button differed depending on whether the participant was left- or right-handed (this was ascertained by the experimenter). Questions were answered using two buttons operated by the left and right thumbs. If the participant thought the correct answer was on the left, he was to press the left thumb button, and vice versa. These buttons were indicated to the

[4] Software programmed by Jonathan Forster at the University of Arizona. DMDX is a member of the DMASTR family of experimental software developed at Monash University and at the University of Arizona by K. I. Forster and J. C. Forster.

participant. The participant was then able to read the on-screen instructions and complete a brief practise test consisting of 6 practise items. If the participant reported or exhibited no problems during the practise session, the experiment proper began. None of the participants experienced any problems during the practice session that might have excluded them from participation. The self-paced reading experiment took approximately 30 minutes to complete and participant's were offered the opportunity to take two short breaks.

Having completed the self-paced reading experiment the participants then completed the Nelson-Denny Reading Test (Form E: Brown, Bennett, & Hanna, 1981). This took exactly 20 minutes to complete.

Results and Discussion

This results section will follow the same format as for experiments 2 and 3: Analysis of interpretation accuracy followed by decision time and reading time results.

For each analysis 2 ANOVAS were performed: one by participants (F_1) and one by items (F_2).

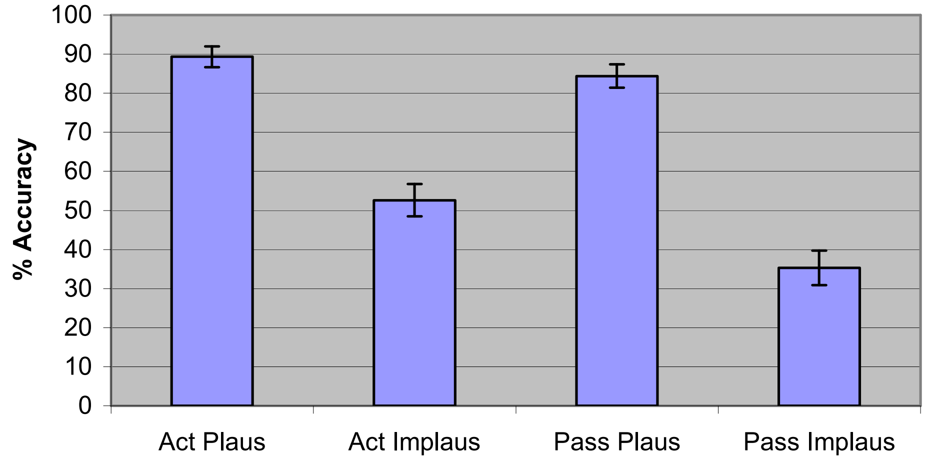
Interpretation Accuracy Results

Two items were removed from the analysis due to typographical errors; analysis of variance was performed on the remaining 30 items. Mean accuracy results are presented in Table 17 and Figure 6, below.

Table 17: Mean % Accuracy Results

	Condition			
	Active Plausible	**Active Implausible**	**Passive Plausible**	**Passive Implausible**
% Accuracy	89.32 (14.96)	52.60 (23.35)	84.38 (16.90)	35.29 (25.00)

Figure 6: Mean Accuracy (%) by Condition



The main effect of voice was significant, with greater accuracy in the Active conditions ($F_1(1, 31) = 11.298$, MSe = 330.824, $p < 0.005$; $F_2(1, 29) = 9.565$, MSe = 340.338, $p < 0.005$). The main effect of Plausibility was also significant, with greater accuracy in the Plausible conditions ($F_1(1, 31) = 113.778$, MSe = 506.762, $p < 0.001$; $F_2(1, 29) = 207.739$, MSe = 265.984, $p < 0.001$). These effects were modulated by an interaction, significant by participants only ($F_1(1, 31) = 6.276$, MSe = 190.955, $p < 0.05$; $F_2(1, 29) = 2.094$, MSe = 487.428, $p >$

0.1). (It's possible that the Items analysis suffered from slightly reduced power due to the removal of two erroneous items). Planned comparisons indicated that there was no significant difference between the two plausible conditions ($t_1(31) = 2.043$, $p > 0.1$; $\mathrm{t}_2(29) = 1.800$, $p > 0.1$), but that there was a significant difference between the two Implausible conditions, with accuracy in the Passive Implausible condition lower than in the Active Implausible condition ($t_1(31) = 14.160$, $p < 0.001$; $\mathrm{t}_2(29) = 6.068$, $p < 0.05$).

This pattern of results is clearly similar to the accuracy results from experiments 2 and 3, though accuracy in the Implausible conditions is considerably poorer. When the EVP had an implausible interpretation readers mistakenly gave answers reflecting a more plausible interpretation. While the accuracy difference between implausible actives and passives has been preserved, the surprising finding is the very low accuracy even in the Active Implausible condition (just over 52%), suggesting that implausible EVPs are simply very difficult to comprehend correctly, with interpretations based on plausibility exerting strong influence. The significant difference between the Active Implausible and the Passive Implausible conditions does provide further confirmation of the particular difficulty in interpreting implausible passives compared with implausible actives. In the case of the Passive Implausible condition *The old woman had been frightened by the mugger in the park yesterday morning. The thug had too according to the news report*, the NP 'the thug' is signalled by the syntax to be the co-theme of the opening sentence with 'the old woman'. Results indicate that readers instead opted for the more plausible role assignment of co-agent roughly 65% of the time.

As outlined in the introduction, these findings cannot be explained purely by the N-V-N heuristic. In the above example this strategy would generate the interpretations ‘old woman frightened mugger’ and ‘thug frightened mugger’, neither of which are semantically compelling. Instead it seems likely that the correct interpretation of the EVP, ‘thug frightened by mugger’, is allowed to become normalised as ‘thug frightens old woman’ due to the availability of a much more suitable filler for the theme role (‘old woman’ is clearly a more plausible theme of ‘frightened’ than is ‘thug’) combined with the good plausibility fit of ‘thug’ as a filler for the agent role of ‘frightened’. The direct comparison of the implausible conditions shows that this is more likely to occur when the antecedent sentence of the EVP is passive, and suggests that this normalising tendency wins out more often when the comprehension system is faced with the more challenging operations required to parse, and in the case of ellipsis, to reconstruct a passive construction.

The fact that the pattern of interpretation results so closely mirrors the results of experiment 3, and therefore of Ferreira’s active/passive findings, could be significant with regard to the ‘weight’ carried by the N-V-N heuristic. While Ferreira’s experiments 2 and 3 were controlled against the argument that passives suffered simply due to being less frequent than actives, the results here demonstrate a disadvantage for implausible passive constructions without the N-V-N strategy explaining it. Garnham and Oakhill (1987) interpreted their accuracy results in terms of the rapid decay of memory for surface structure of the antecedent sentence – at the time of interpreting the EVP, the necessary memory representation was insufficiently specified to allow correct interpretation. (Alternatively, correct interpretation occurred, but was overridden due to the influence of pragmatic cues on the decaying, fragile memory for the antecedent.) If the current results can be interpreted in terms of decay, then the decay is more rapid for passive structures than actives, and the only heuristic we need to posit is a plausibility heuristic. In that case, a heuristic based on role assignment order may have only negligible influence over and above one using pragmatic constraints, and a simple plausibility heuristic may account for a great deal of the misinterpretations observed in the above-mentioned active/passive experiments.

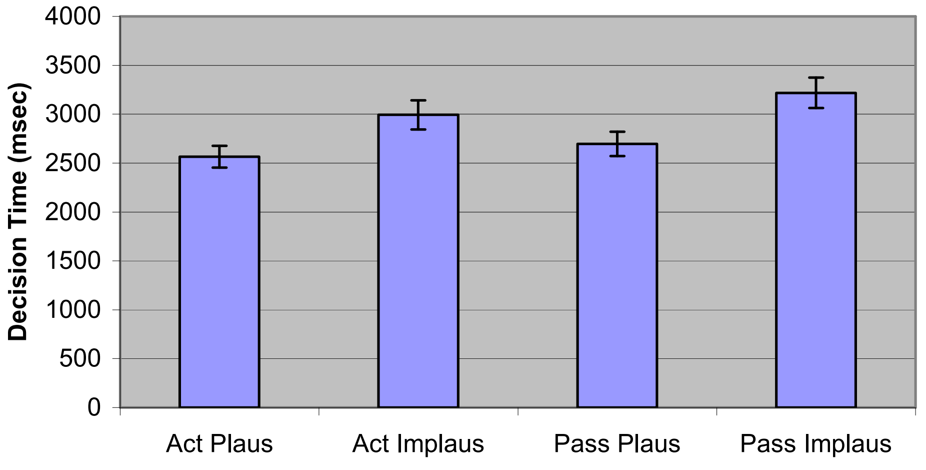
Decision times

Decision time was calculated as the time from the appearance onscreen of the question to the moment the participant pressed the button to answer it. As with the previous analysis, two items were excluded due to typographical error. Mean decision times are presented in Table 18 and Figure 7

Table 18: Mean Decision Time by Condition

	Condition			
	Active Plausible	**Active Implausible**	**Passive Plausible**	**Passive Implausible**
Decision Time (msec)	2565 (636)	2993 (850)	2696 (702)	3218 (884)

Figure 7: Mean Decision Time (msec)



There was no significant effect of the Voice variable, though the effect was marginal by participants with a suggestion of longer decision times following passive conditions ($F_1(1, 31) = 3.491$, MSe = 287617.191, $p < 0.071$; $F_2(1, 29) = 1.525$, MSe = 367285.577, $p > 0.2$). There was a significant main effect of Plausibility such that decision times were longer following the two Implausible conditions ($F_1(1, 31) = 18.055$, MSe = 399766.013, $p < 0.005$; $F_2(1, 29) = 17.066$, MSe = 430939.853, $p < 0.005$). The interaction was not significant (both *F*s < 1).

The longer decision times following the implausible conditions indicate that readers were conflicted a significant proportion of the time, between a plausible and an implausible interpretation, and took longer to confirm their answer than in the Plausible conditions, when they would presumably have generated only the correct interpretation. Clearly, however, the longer decision times in these conditions did not lead to a benefit in terms of accuracy. The accuracy difference between the two implausible conditions was not reflected in the decision times: Readers took no longer to answer in the Passive Implausible condition than the Active Implausible condition relative to their plausible baselines. One explanation for this equality could be that it reflects a proportion of cases in the Passive Implausible condition (least accuracy) in which readers did not generate the implausible interpretation at all, and so were not left deliberating between alternative interpretations at the decision making stage.

Reading time results

The results of the reading time analysis will be presented by Word, beginning with the critical word followed by the following four words to allow for spillover effects. Reading times less than 100 msec and greater than 4000 msec were excluded from the analysis. To preview the results, there was disruption evident online in the implausible conditions but the earliest robust effect was on the second spillover word. There is thus evidence of online, reasonably early application of syntax-based interpretations. There were no significant effects of Voice and the Voice and Plausibility variables did not interact at any point. Mean reading times are presented in Table 19.

Table 19: Mean Reading Time on Critical Word and Spillover Words

			Region		
	Critical Word *…too…*	**Critical Word + 1** *according*	**Critical Word + 2** *to*	**Critical Word + 3** *the*	**Critical Word + 4** *news…*
	Mean (SD)				
Condition					
Active Plausible	314 (125)	328 (133)	329 (130)	282 (78)	289 (76)
Active Implausible	313 (121)	355 (184)	370 (151)	303 (84)	334 (139)
Passive Plausible	301 (102)	337 (172)	309 (93)	296 (88)	313 (93)
Passive Implausible	318 (141)	392 (231)	356 (150)	307 (93)	335 (98)

Critical Word (…had ***too*** according to the news report)

There were no significant effects at all on the critical word (all Fs < 1).

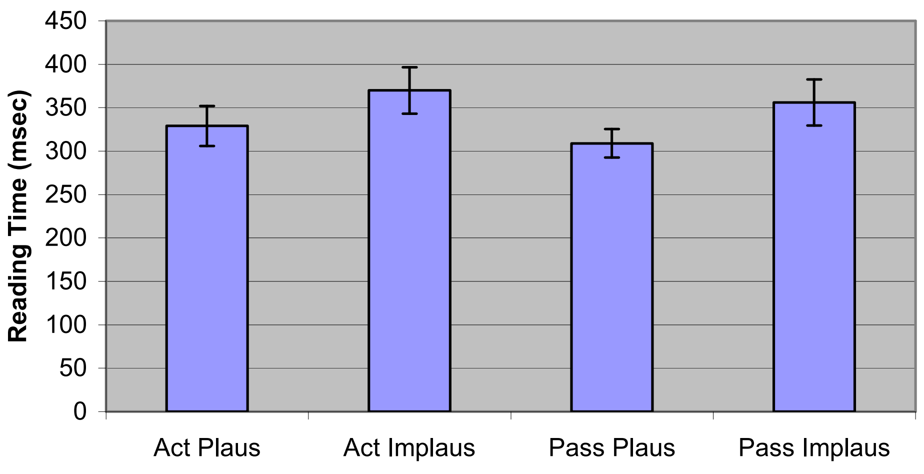
First Spillover Word (…had too ***according*** to the news report)

There was no significant effect of voice (F_1(1, 31) = 1.969, MSe = 8664.017, $p >$ 0.1; F_2(1, 31) = 2.495, MSe = 18017.678, $p > 0.1$). There was an effect of plausibility that was marginal by participants but not significant by items, with longer reading times in the implausible conditions (F_1(1, 31) = 2.999, MSe = 18095.516, $p = 0.093$; F_2(1, 31) = 2.754, MSe = 17396.062, $p > 0.1$). The interaction was not significant (both $F < 1$).

Second Spillover Word (…had too according ***to*** the news report)

There was no significant effect of voice (both Fs < 1). Again, there was an effect of plausibility, now significant by both participants and items, with longer reading times in the implausible condition (F_1(1, 31) = 5.433, MSe = 11436.360, $p < 0.05$; F_2(1, 31) = 7.986, MSe = 7800.838, $p < 0.05$). There was no significant interaction (both Fs < 1) (see Figure 8).

Figure 8: Reading Times (msec) 2nd Spillover Word



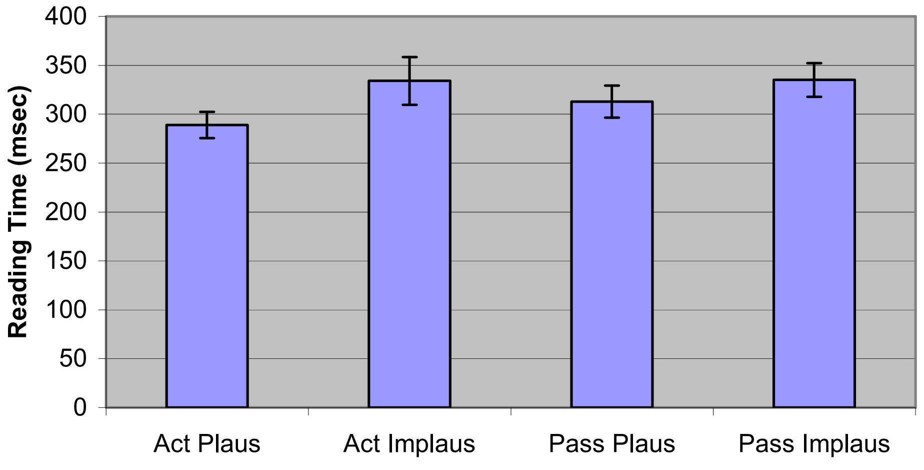
Third Spillover Word (…had too according to ***the*** news report)

There was no significant effect of voice ($F_1 < 1$; $F_2(1, 31) = 1.037$, MSe = 2189.856, $p > 0.3$). The effect of plausibility was this time marginal ($F_1(1, 31) = 4.053$, MSe = 2033.560, $p = 0.053$; $F_2(1, 31) = 2.987$, MSe = 2742.314, $p = 0.094$); there were longer reading times in the implausible conditions. There was no significant interaction (both *F*s < 1).

Fourth Spillover Word (…had too according to the ***news*** report)

There was no significant effect of voice (both *F*s < 1). The effect of plausibility was once again significant, with longer reading times in the implausible conditions ($F_1(1, 31) = 5.635$, MSe = 6273.980, $p < 0.05$; $F_2(1, 31) = 5.923$, MSe = 8662.858, $p < 0.05$). There was no significant interaction (both *F*s < 1) (see Figure 9).

Figure 9: Reading Time (msec) 4th Spillover Word

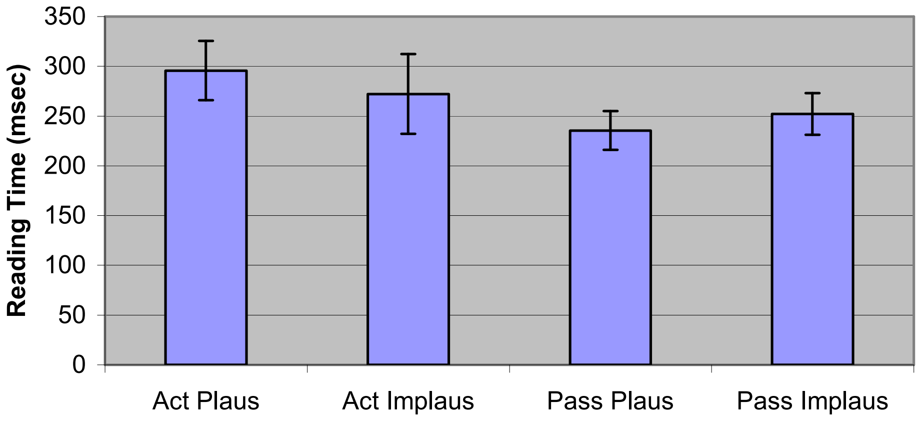


Results by response type

As with experiments 2 and 3, we performed an analysis aimed at investigating whether, in cases when readers incorrectly answered a question, they had previously interpreted the item correctly. So, for example, robust evidence of early disruption in the implausible conditions, relative to the plausible ones, would suggests that syntax informs the interpretation first, but a further interpretation is then generated based on plausibility considerations.

Eleven participants had committed errors in all four conditions and were eligible for an analysis of incorrectly answered trials. The only effect approaching significant was on the critical word itself, which was read for marginally longer in the Active conditions ($F_1(1, 10) = 4.599$, MSe = 3848.525, p = 0.058) (see Figure 10). All other ps > 0.2.

Figure 10: Incorrect Trials
Reading Time (msec) Critical Word



We also conducted an analysis by response type on the Implausible trials only to see if reading times differed depending on whether or not the participant had answered the question correctly for any given trial (N=28). There were no significant effects but inspection of the means (see Table 20) suggests that when a trial was answered correctly participants had slowed down on the second spillover word, (though the very high standard deviation in the Active Right condition indicates considerable variation).

Table 20: Analysis of the Implausible Trails by Response

	Region				
	Critical Word *…too…*	**Critical Word + 1** *according*	**Critical Word + 2** *to*	**Critical Word + 3** *the*	**Critical Word + 4** *news…*
	Mean (SD)				
Condition					
Active Right	327 (209)	378 (237)	446 (541)	292 (96)	320 (112)
Active Wrong	286 (105)	328 (178)	323 (125)	310 (106)	348 (181)
Passive Right	290 (154)	379 (273)	338 (248)	308 (131)	328 (117)
Passive Wrong	311 (150)	362 (231)	327 (145)	306 (105)	321 (86)

A three-way ANOVA including ‘Word’ as a factor was performed and yielded a significant main effect of Region ($F_1(1, 27) = 3.006$, MSe = 37195.066, $p < 0.05$) with longest reading times in the second spillover region, and a marginal effect of Response ($F_1(1, 27) = 3.301$, MSe = 23123.627, $p = 0.080$), with longer reading times when the question was answered correctly.

General Discussion

The surprisingly low accuracy results provide yet more evidence that unambiguous sentences can be radically misinterpreted. Most striking in this particular set of results, perhaps, was the difficulty readers had with interpreting implausible EVPs whose antecedent was active. Granted that EVPs themselves are difficult to interpret, active constructions have reliably elicited high levels of correct interpretation elsewhere. But here, when pragmatic cues suggested an alternative interpretation, actives suffered badly. Overall, our accuracy results were even lower than those reported by Garnham and Oakhill, and we can offer one explanation in terms of the differing methodologies – word-based reading being more taxing than segment-based reading. The evidence continues to mount, then, for a picture of language comprehension in which comprehenders do not interpret according to the best available evidence.

In terms of the processing data, the results were broadly similar to those from experiments 2 and 3. The lack of plausibility effects at the critical word might indicate that the syntactically licensed interpretation (the implausible one) was not applied immediately. However, the marginal effect at the first spillover word, and the robust effect at the second (carrying onto the fourth) spillover word does indicate that anomaly detection, and thus syntax-based interpretation, occurred online and very soon after the occurrence of the anomaly. The fact that Voice did not interact with Plausibility, as it did in the accuracy results, suggests that both active and passive anomalies were detected online (causing equal disruption) and that the processes leading to the ultimate interpretation differences were a later, possibly offline phenomenon. (This interpretation was also proposed for the processing results of the previous two experiments.)

The analysis carried out on the reading time data, by response type, did not yield any significant differences. The power of this analysis was undoubtedly low (N=11) and any conclusion would therefore be tentative, but it would appear that for these readers, an incorrect response entailed a lack of plausibility effects in the reading time data. A similar analysis with only the Implausible conditions likewise failed to reveal any reading time differences on incorrectly answered trials. The only relevant effect was the apparent lengthened reading times when a trial was answered correctly – perhaps not a very surprising result, given that the lengthened reading times would index correct online interpretation. If these analyses were sufficiently powerful we might have been able to conclude that with implausible materials, an incorrect answer meant that the material had never been correctly parsed and interpreted at all, and had in fact been interpreted solely on the basis of semantics. In terms of the relative time course this could suggest an ‘either/or’ model in which a sentence is either interpreted with full reference to the syntactic information, or according to a plausibility strategy. However, given the robustness of the online plausibility effects, and the power concerns with these latter analyses, it is perhaps safer to say that the evidence here supports a model in which the syntactic interpretation is generated first and then overridden by plausibility cues, either online or offline. (Recall that implausible passives did not take longer to interpret, at the decision stage, than implausible actives. One explanation for this was that the equal time taken to decide reflected fewer instances of 'interpretation conflict' in the passive case, presumably due to the correct, syntactic interpretation never having been built. This is a possibility, but it can only be speculation for now and should not displace the syntax-first account as the safest interpretation of our results. It could, after all, be the case that the syntactic interpretation had been built, but in a number of cases had decayed extremely rapidly and exerted no influence at the decision stage.)

These findings and interpretations allow greater insight into processing issues than does the original Garnham and Oakhill study. The two studies are not contradictory, as the early methodology could not allow for investigation into the time course of syntactic vs. heuristic interpretation – indeed, the authors were not concerned to. As just stated, however, we would suggest resolving the time course question in favour of a syntax-first account. Finally, our (low power) analysis was not any more successful than theirs in determining how incorrectly interpreted EVPs are processed online: the sparse data problem continues to hamper investigation into this issue.

Chapter 4

The Interpretation And Processing Of Implausible Sentences: Two Eyetracking Studies

Experiment 5: The Interpretation And Processing Of Elliptical Verb Phrases – An Eyetracking Study

Introduction

The results of experiment 4 demonstrated that while interpretations of implausible elliptical verb phrases (EVPs) could be mistaken to a surprisingly large extent, the detection of an implausible meaning was an online, albeit delayed, phenomenon. It was not possible to conclude firmly that implausible meanings are *always* detected online, although this seemed a reasonable interpretation of the results. Another question not fully resolved, and one we will now pursue, relates to the timing of the observed online anomaly detection. The possibility exists that the delayed detection was related to the methodology rather than being a true reflection of real-time processing. With the repetitive nature of the button-pressing that was necessary for a participant to read the passages, it could be that anomaly detection was immediate but masked by participants pressing the button according to an established rhythm. Thus, the reading disruption caused by the implausibility could have affected the rate of button-pressing only two or three words after it was detected.

This possibility could be investigated (and any problem resolved) by using eyetracking: under natural reading conditions any disruption in the normal reading process would be observable immediately as it occurred. The present study therefore used the 32 materials from experiment 4 and tested them under eyetracking conditions.

There have been a number of eyetracking studies investigating eye movement responses to linguistic anomalies, and attempts made to map correspondences between anomalies on different linguistic levels and particular fixation/movement patterns. Obviously we would like to make predictions about how our anomalies would affect readers if such predictions are warranted.

Two main points may be repeated from a recent review of the topic (Boland, 2004). The first point is that the literature on the subject of how readers' eye movements respond to anomalies in unambiguous sentences is small (p. 56). The second is that it is inconsistent. Discrepancies centre on what type of anomaly will influence first-pass measures on the anomalous word/region, and studies have either reported that this or that type of anomaly influenced first pass measures, or influenced only some first-pass measures. Boland and Blodgett (2002) reported that syntactic anomalies influenced first pass times when they involved phrasal category errors but not a morphological feature error. Both types of syntactic error influenced the first pass regression rates, while semantic errors increased regression path time. Ni, Fodor, Crain and Shankweiler (1998) reported that both syntactic and semantic anomalies increased the first-pass regression rate, but only semantic anomalies increased first-pass reading times, and only then in regions subsequent to the anomalous region. Braze, Shankweiler, Ni, and Palumbo (2002) reported 'later' effects for semantic compared to syntactic anomalies, with semantic anomalies prompting a gradual increase in regressions, peaking at the end of the sentence. Pearlmutter, Garnsey and Bock (1999) reported late, i.e. post-anomaly, effects for subject-verb disagreements, with effects only apparent when the anomalous region was analysed in combination with the following region. On the other hand, Frisson and Pickering (1999), in their study of semantic anomalies using metonymy, reported both early (first-pass) and late effects. To sum up this brief account, there is evidence that syntactic violations tend to appear in first-pass measures, with consistency only in the first-pass regression measure; evidence on semantic-pragmatic anomalies tends to show later influences, with little in the way of first-pass effects. (Boland concludes from this that constraints that control structure building influence first-pass reading time).

There is also a small literature on the effects of 'semantic pre-processing', or 'parafoveal-on-foveal' effects, that may be relevant to our measurements of anomaly detection. While the reported effects are controversial (see Kennedy & Pynte, 2005, for a discussion), several researchers have reported effects of the *pragmatics* of word *n*, on fixation measures on the foveally-fixated word *n*-1 (Murray and Rowan, 1998, Kennedy, Murray & Boissiere, 2004; but see Rayner, White, Kambe, Miller & Liversedge, 2003). Even low-level properties of words are included in the controversy, but it could be worthwhile to examine our pre-critical regions to allow for the possibility of very early, parafoveal anomaly detection.

In terms of predictions, then, we need not be rigid. It seems reasonable not to expect effects in first pass measures on our critical region; but we needn't rule them out. We can certainly expect our pragmatic anomalies to appear in the processing record soon afterwards, though, probably as early as the first spillover region. Given the currently inconsistent state of the literature, the data from the present study, and the upcoming studies, may help to build a more precise picture of the effects of semantic/pragmatic anomalies on readers' eye movements.

Method

Participants

32 participants were recruited from the University of Glasgow student community and were paid for their participation. All participants were native English speakers, had not been diagnosed with dyslexia, and had normal vision or corrected-to-normal vision using soft contact lenses. All participants were kept naïve to the design and goals of the study.

Materials and design

The materials used in this study were the 32 experimental items from experiment 4. The accuracy data gathered in Experiment 4 indicated that the materials were well designed in terms of plausibility and would be suitable for further testing without any changes. Presentation of the materials differed in two ways. First, eyetracking methodology allows the whole passage to be presented at once, as opposed to the word-by-word presentation of Experiment 4. Second, due to restrictions on line length when using eyetracking, the final word of the first sentence was always the first word of the second line in Experiment 5. Otherwise, the materials were identical.

The task was also identical to that of Experiment 4: a 'yes/no' response to a question probing the interpretation of the EVP in the second sentence. As in Experiment 4, the question was not viewed until the passage had been read and had disappeared from the screen.

The experimental materials were mixed with 92 filler passages. Twenty-eight of these were experimental items for experiment 6. A further 32 were fillers matched to the items for experiment 6 and all were modelled after Experiment 6's Plausible items. The final 32 fillers were matched to the experimental items for experiment 5. Within these 32 fillers, 16 were modelled after the Active-Plausible items; half of this group had the correct answer 'Yes' and half had the correct answer 'No'. The other 16 were modelled after the Passive-Plausible items and had the same 50% 'Yes', 50% 'No' split. Experimental items for Experiments 5 (and Experiment 6) were rotated across four lists using a latin square design; participants in each subject group thus saw all 32 items, but each item in only one of its four conditions. In terms of Experiment 5, participants in each subject group would view a total of 16 implausible items. The design of Experiment 6 meant that they would view a further 8 implausible items in each list, although these would bear no resemblance to the Experiment 5 items. The implausible items were thus well hidden among 100 plausible fillers and participants would therefore be unlikely to have had a problematic amount of practice at interpreting them. The interpretation questions were the same as those used in experiment 4.

Apparatus

Eye movements were monitored using a Generation 5.5 Fourward Technologies Dual Purkinje Image Eyetracker. The tracker monitored a participant's gaze location every millisecond, and the software sampled the tracker's output to establish the position of eye fixations and their start and finish times. The tracker monitored eye movements only from the right eye, though viewing was binocular. The passages of text were displayed on a PC VDU screen positioned approximately 80cm from the participants' eyes. The screen displayed approximately four characters per degree of visual angle. Participants' head movements were minimised using a bite bar (prepared individually for each participant), forehead rests and a head strap.

Procedure

Upon entering the lab participants read and signed an instruction and consent form. This briefly explained that they would be reading short passages of text while their eye movements were monitored and recorded. The experimenter also explained the use and preparation of the bite bar and then prepared a new bite bar for the participant. Participants then sat at the Eyetracker and completed a short practice session consisting of some further onscreen instructions related to eyetracking procedure, and three practice materials. Participants were instructed to read at their natural pace and to avoid rereading as much as possible. This last instruction was to avoid excessive rereading of any anomalous phrases or repeated reading of passages to prepare for the questions. Nevertheless, we did expect to see regressions in the eye-movement data, simply because they are a normal and automatic part of the reading process, and it would be especially reasonable to expect rereading when faced with implausible or anomalous, i.e. difficult, material. Three of our reading time measures, *first pass regressions out, regression path time and total reading time* take such rereading into account and will be included in the analysis.

Following the practice session, a brief calibration procedure was carried out. The experiment began with the participant fixating a small, box-shaped character in the top-left section of the monitor that signalled the position of the first character of the upcoming text. The first passage was then presented and participants read through it. The participant then fixated another box-shaped character below and to the right of the last character of text, and pressed either of two hand-held buttons. The question screen was then displayed:

e.g.

Did the vicar bless the bishop?

Yes <> No

Participants responded by pressing the right button if they thought the answer on the right was correct, or the left button if they thought the answer on the left was correct. This constituted one trial and the pattern was repeated throughout all 124 trials. The experimenter checked calibration between trials, and the eye-tracker was re-calibrated if necessary. Following the eyetracking experiment, the participant then completed Form E of the Nelson Denny Reading Test (Form E: Brown, Bennett and Hanna, 1981)

Data analysis

An automatic procedure pooled short contiguous fixations. This procedure merged fixations of less than 80 msec into any neighbouring fixations within a distance of one character, and then deleted any remaining fixations of less than 80 msec.

In order to calculate eye-movement measures, the experimental materials were split into regions. The regions are given for an example item in (4):

> (4).[$_1$The assistant] [$_2$had been serving] [$_3$the woman] [$_4$at the customer service desk.] [$_5$The manager] [$_6$had too] [$_7$and the] [$_8$problem was] [$_9$resolved.]

The first region comprised the first noun phrase (NP), the second region comprised the first verb phrase (VP), the third region comprised the second NP, and the fourth region contained all the words after the second NP, up to the end of the first sentence. The fifth region comprised the third NP, the subject of the elliptical verb phrase (EVP). The sixth region – the critical region – comprised the words *had too* and was the earliest point at which we would expect a slowdown in the Implausible conditions. The remaining five words of each item were split into three regions to allow for examination of any spillover effects caused by the correct interpretation of an implausible EVP. Words of three letters or fewer never formed a single region due to the likelihood of them being skipped, and were either combined with another three letter word (as in the example above) or to an adjacent longer word as the situation allowed.

We will report results for Regions 5-8, focusing mainly on the critical region (region 6).

Reading Time Measures

We calculated eye-movement measures on all regions associated with both early and later processing. Recent studies that have used eyetracking with anomaly detection tasks (Daneman, Lennertz and Hannon, 2007; Daneman, Hannon and Burton, 2006) reported first pass reading time, number of first pass fixations on the target word, look-back time on the target word and number of look-back fixations (regressions) on the target word. Note that, because all measures are taken on the target word only, only a broad division into 'immediate detection' and 'delayed detection' is possible – it would not be possible to tell using this design exactly when detection had occurred unless there was clear evidence of detection in the early measures on the target word. This approach may have been necessary in these studies, which used a very small material set (only three passages), but clearly we need a broader field of analysis to satisfactorily keep track of time course issues. The measures and regions used in the present study, and described now, allow this broader investigation.

The duration of the *First Fixation* in a region is a measure of the very earliest processing in that region, though when applied to a post-critical region, it can also be informative about later, integrative processing. The measure is calculated by taking the duration of the fixation following the first saccade into the region from the left, before any material to the right of the region has been fixated. We also report *Gaze Duration*. This measure involves summing the duration of all fixations made within a region from the time it is first entered from the left to when it is first exited to the left or right. Thus, gaze durations will tend to be longer than first fixation times as they allow for multiple initial fixations within a region. This measure is also informative about early processing, although, as with first fixation, when applied to a post-critical region, it can also be informative about later, integrative processing. In passages involving implausible or anomalous material, these measures may be indicative of anomaly detection, and thus will be applied to post-critical regions to allow for delayed anomaly detection. The term *gaze duration* is preferred when the region of interest consists of a single word; however, this measure is generally known as *First Pass Reading Time* when the region consists of two or more words. As we will be reporting data for both single word regions and larger regions, we will use both terms to refer to this measure.

Percentage of First Pass Regressions Out is a further measure of early processing, with an increase of regressions out of a particular region to reread previously read material indicating an early difficulty in processing that region. The measure is calculated as the percentage of times a reader regresses out of a given region. A related measure is *Regression Path Time*, also known as *Go Past Time*, calculated as the time taken to exit a region to the right after entering it from the left. It thus includes all the time spent rereading material prior to that region if a reader has made a regression out of it, and the sum of re-fixations if the region is fixated again before being exited to the right.

Total Time is the sum of the durations of all fixations made within a region, so it will include gaze duration/first pass times and the sum of any fixations made on the region after the reader has already exited the region. The measure is a good reflection of the overall processing difficulty associated with a given region.

In the cases of first fixation time, first pass time, and regression path time (which can be collectively referred to as first pass measures), a zero reading time is recorded for a region if subsequent regions are read first. Data analysis procedures exclude these zero reading times and calculate mean reading times from the other data points in the design cell.

Results

Interpretations

We will first present the accuracy results from the interpretation questions. Our analysis programme generated error information for each participant, and Table 21 contains the mean number of errors and mean % Accuracy by condition. An ANOVA (F_1) was performed on the accuracy data[5] (see Figure 11).

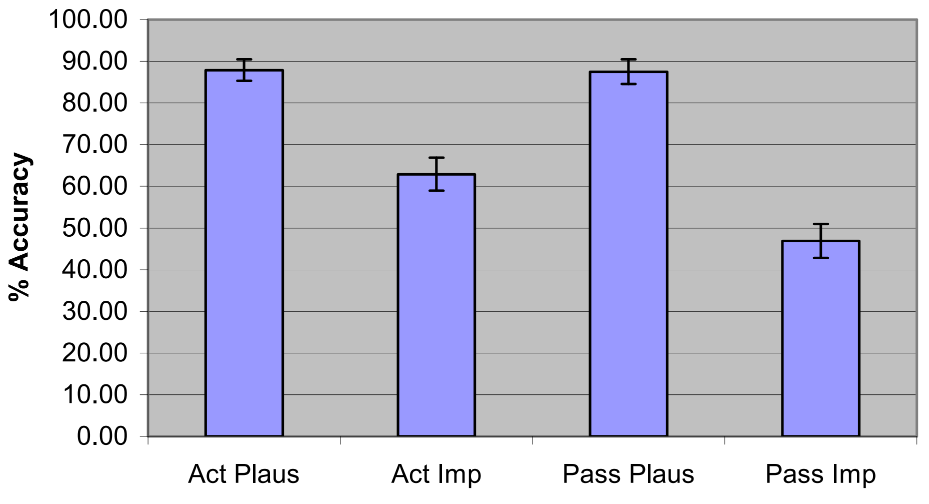
[5] The accuracy data was generated from the eyetracking data by an automatic procedure that provided only participant results.

Table 21: Mean number of errors and mean % Accuracy

Condition	Mean no. of Errors	% Accuracy
Active Plausible	0.97	87.89
Active Implausible	2.97	62.89
Passive Plausible	1.00	87.50
Passive Implausible	4.25	46.88

Analysis of variance indicated that the effect of voice was significant, with greater accuracy in the Active conditions ($F_1(1,31) = 6.452$, MSe = 333764, $p < 0.05$). There was also a significant effect of plausibility, with greater accuracy in the plausible conditions ($F_1(1,31) = 91.752$, MSe = 375.504, $p < 0.001$). These effects were modulated by a significant interaction ($F_1(1,31) = 11.923$, MSe = 163.810, $p < 0.005$). T-tests indicated that there was a significant difference between the two implausible conditions, with greater accuracy in the Active Implausible condition ($t_1(31) = 3.334$, $p < 0.005$). The plausible conditions did not differ ($t_1 < 1$) and were apparently interpreted with equal ease.

Figure 11: Accuracy (%) by Condition



Processing results

For each analysis, two ANOVAs were conducted: One by participants (F_1) and one by materials (or items; F_2). Reported means are for F_1 analyses. Analysis will be presented by region, with *F* values (and interpretations where appropriate) for each reading time measure in a region. Mean reading time results are reported in Table 22, below.

To summarise, it appears that anomaly detection occurs online with both active and passive implausibilities. However, the earliest effects are not robust. The detection-related disruption may begin and end earlier for the Active Implausible condition (detection may occur as early as the critical region itself); an alternative interpretation is that the Passive implausibility is detected first, in the first spillover region, which contains the first robust effects.

Table 22: Mean reading time measures for the critical and following two regions

	Region			
	Region 5 (pre-critical)	**Region 6 (critical)**	**Region 7**	**Region 8**
Measure	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)
First Fixation (msec)				
Act Plaus	272 (143)	287 (56)	266 (78)	257 (42)
Act Implaus	265 (45)	275 (49)	268 (55)	272 (41)
Pass Plaus	261 (48)	280 (53)	269 (55)	253 (41)
Pass Implaus	263 (50)	283 (63)	277 (53)	262 (52)
First Pass (msec)				
Act Plaus	362 (173)	348 (78)	317 (101)	394 (105)
Act Implaus	370 (108)	347 (112)	327 (70)	421 (137)
Pass Plaus	380 (85)	328 (84)	313 (83)	368 (421)
Pass Implaus	379 (172)	339 (113)	353 (90)	383 (368)
First Pass Reg. Out (%)				
Act Plaus	19.34 (14.91)	7.56 (10.70)	21.84 (22.57)	26.06 (26.63)
Act Implaus	18.47 (15.38)	10.94 (16.72)	16.22 (19.34)	24.31 (19.69)
Pass Plaus	14.19 (13.28)	9.81 (15.78)	15.09 (20.43)	21.03 (21.34)
Pass Implaus	18.09 (14.63)	5.59 (12.03)	16.78 (22.36)	24.25 (21.93)
Regression Path (msec)				
Act Plaus	545 (237)	412 (155)	477 (126)	719 (437)
Act Implaus	539 (204)	400 (163)	446 (180)	723 (390)
Pass Plaus	498 (144)	381 (124)	427 (208)	715 (621)
Pass Implaus	539 (249)	386 (189)	495 (223)	790 (700)
Total Time (msec)				
Act Plaus	559 (256))	425 (135)	433 (134)	504 (148)
Act Implaus	592 (264)	488 (203)	451 (122)	527 (130)
Pass Plaus	521 (153)	430 (127)	399 (119)	488 (136)
Pass Implaus	594 (286)	479 (214)	486 (174)	549 (209)

Region 5 (pre-critical)

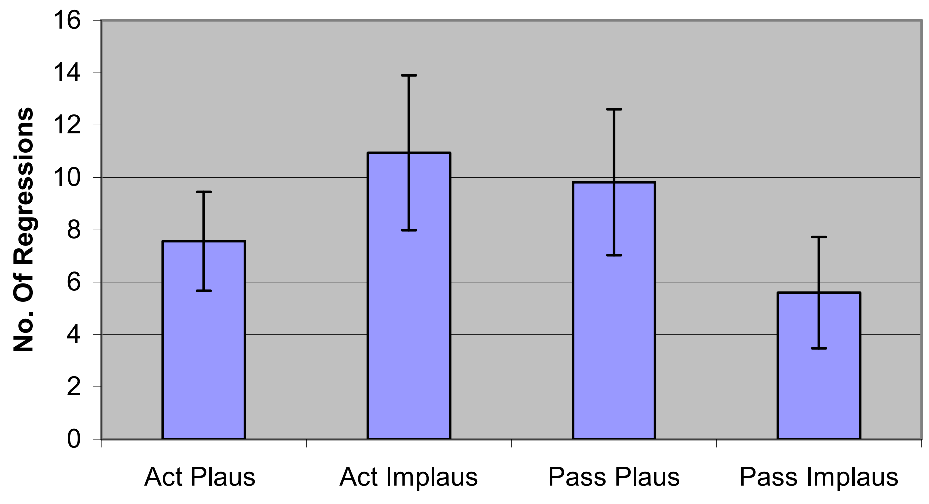
There were no significant effects in either the first fixation or first pass measures (all Fs < 1). In first pass regressions out there was no significant effect of voice ($F_1(1, 31) = 1.230$, MSe = 199.030, $p > 0.2$; $F_2(1, 31) = 2.638$, MSe = 159.403, $p > 0.1$), no significant effect of plausibility (both Fs < 1), and no interaction ($F_1(1, 31) = 1.049$, MSe = 174.318, $p > 0.3$; $F_2 > 0.3$). Regression path analysis showed no significant effect of voice ($F_1 < 1$; $F_2(1,31) = 1.277$, MSe = 27886.955, $p > 0.2$), or plausibility (both Fs < 1) and no interaction (both Fs < 1). In total time there was a significant effect of voice, with longer reading times in the implausible conditions ($F_1(1, 31) = 4.265$, MSe = 20680.242, $p < 0.05$; $F_2(1, 31) = 3.893$, MSe = 26239.726, $p = 0.057$). There is therefore no evidence that the anomaly was detected prior to fixating the critical, anomalous region itself. (But see chapter 6 for some evidence that the less-skilled readers may have detected the Active anomaly in this region).

Critical region (region 6)

While there is no conclusive evidence in this region of anomaly detection for either the active or the passive anomalies, the means are suggestive of disruption in the active conditions but not the passive, and hence, possibly of an immediate detection of the active condition anomaly.

First fixation analysis revealed no significant main effects of either voice or plausibility (all *F*s < 1) and no significant interaction ($F_1(1,31) = 1.249$, $p > 0.2$; $F_2 > 0.7$). There were likewise no significant effects in first pass analysis (voice; $F_1(1,31) = 1.415$, MSe = 4613.257 $p > 0.2$; $F_2(1,31) = 1.212$, MSe = 2624.870, $p > 0.2$; plausibility: both *F*s < 1; interaction: both *F*s < 1). In the first pass regressions out analysis, while there were no significant effects of either voice or plausibility (all *F*s < 1), there was an interaction significant in the items analysis ($F_1(1,31) = 2.291$, MSe = 201.336 $p > 0.1$; $F_2(1,31) = 4.671$, MSe = 113.937, $p < 0.05$). T-tests comparing Active conditions and Passive conditions separately showed a marginally significant difference in the items analysis of the Passive conditions only (t_2 (31) = 1.990, $p = 0.055$; all other *p*s > 0.1). However, the direction of the means in the passive conditions is not at all suggestive of anomaly detection – if the difference is real then, surprisingly, the plausible condition elicited more regressions than the implausible condition. It is not clear why this difference should have emerged, but the weakness of the effect, and the robust effects to be reported shortly, suggest it isn't related to our plausibility manipulation. The means in the Active conditions, while not significantly different, are on the other hand in the expected direction for detection of the anomaly (see Figure 12). In regression path analysis there was no significant effect of voice ($F_1(1,31) = 1.872$, MSe = 8748.435, $p > 0.1$; $F_2 < 1$), plausibility (both *F*s < 1) and no interaction (both *F*s < 1). Total time analysis yielded only a main effect of plausibility, with longer total reading times in the implausible conditions ($F_1(1,31) = 4.784$, MSe = 20849.655, $p < 0.05$; ($F_2(1,31) = 5.879$, MSe = 18845.959, $p < 0.05$; all other *F*s < 1).

Figure 12:
First Pass Regressions Out (%), Critical Region



Region 7 (first spillover region)

First fixation analysis revealed no significant effects at all (all *F*s < 1). In first pass there was no significant effect of voice (both *F*s < 1) but there was a significant main effect of plausibility ($F_1(31) = 5.519$, MSe = 3664.907, $p < 0.05$; $F_2(31) = 4.260$, MSe = 3692.800, $p < 0.05$), and an interaction that was marginally significant in the items analysis ($F_1(31) = 2.171$, MSe = 3309.766, $p > 0.1$; $F_2(31) = 3.190$, MSe = 4456.470, $p = 0.084$) (see Figure 13). The difference between the two passive conditions is indicative of anomaly detection and is (numerically) larger than the difference between the active conditions, though the interaction is not robust. If the difference between the Active means in the critical region (first pass regressions out) were indeed due to an early detection of the anomaly in the active condition, that would explain the difference between the voice conditions here – the explanation being that having begun earlier in the active conditions, the disruption may be settling down earlier. However, it may be more likely that the effect in this region simply indicates that the implausibility has been detected in the passive condition first. There was nothing significant in first pass regressions out (all *p*s > 0.1) but in the regression path analysis there was an interaction, significant by items, suggesting that the disruption was eliciting rereading in the Passive Implausible condition, but not in the Active Implausible condition ($F_1(1,31) = 2.073$, $p>0.1$, Mse = 37830.1; $F_2(1,31) = 4.311$, $p<0.05$, Mse = 47706.8; see Figure 14). T-tests showed only a marginal difference (by items) between the Passive conditions ($t_2(31) = 1.907$, $p = 0.066$; all other *p*s > 0.1). (All other *F*s for this measure were < 1). In total time analysis there was no effect of voice (both *F*s < 1). There was a significant effect of plausibility, with longer total reading times in the implausible conditions ($F_1(1,31) = 10.067$, MSe = 8849.903, $p < 0.005$; $F_2(1,31) = 6.178$, MSe = 14843.273, $p < 0.05$). The interaction was not significant ($F_1(1,31) = 2.173$, MSe = 17319.740, $p > 0.1$; $F_2(1,31) = 3.007$, MSe = 15826.915, $p = 0.093$). (Inspection of the means suggests that the rereading cost was greater in the passive conditions than in the active conditions.)

Figure 13:
First Pass reading time (msec), Region 7

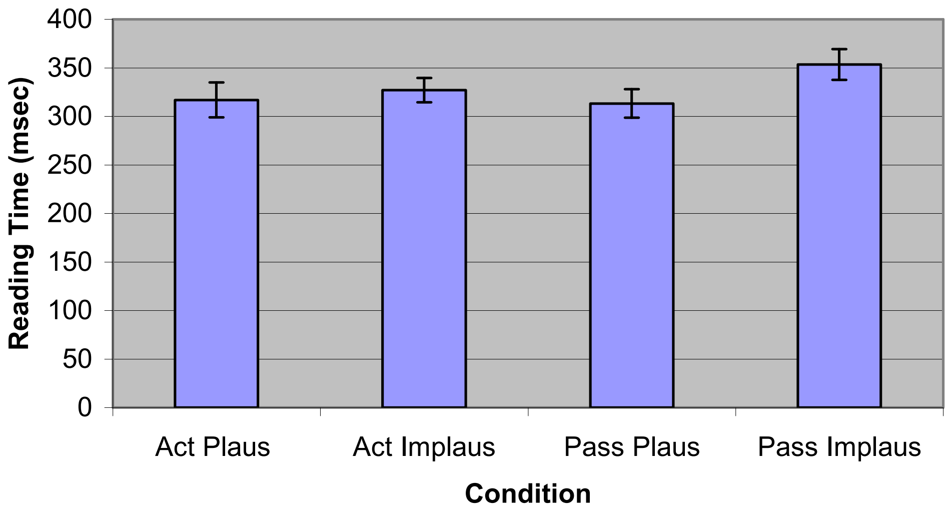
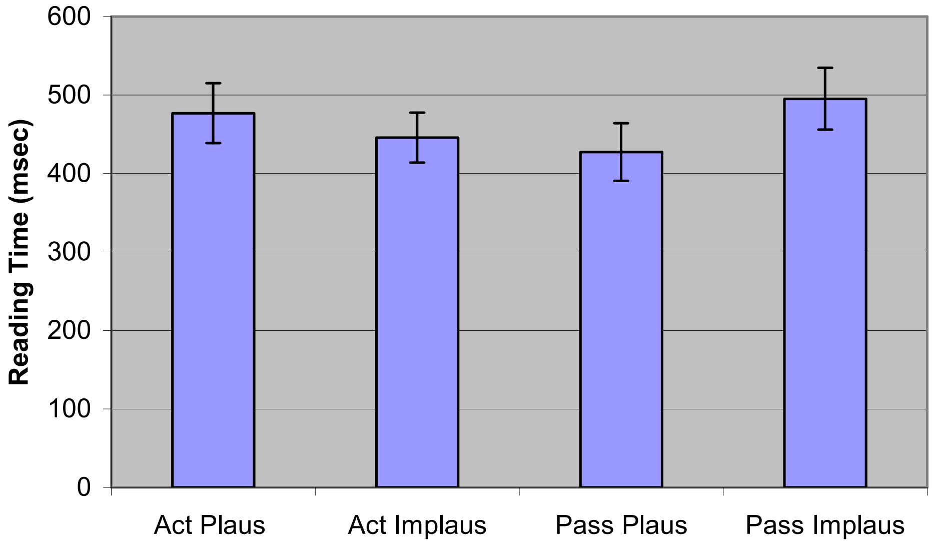


Figure 14:
Regression Path Time (msec), Region 7

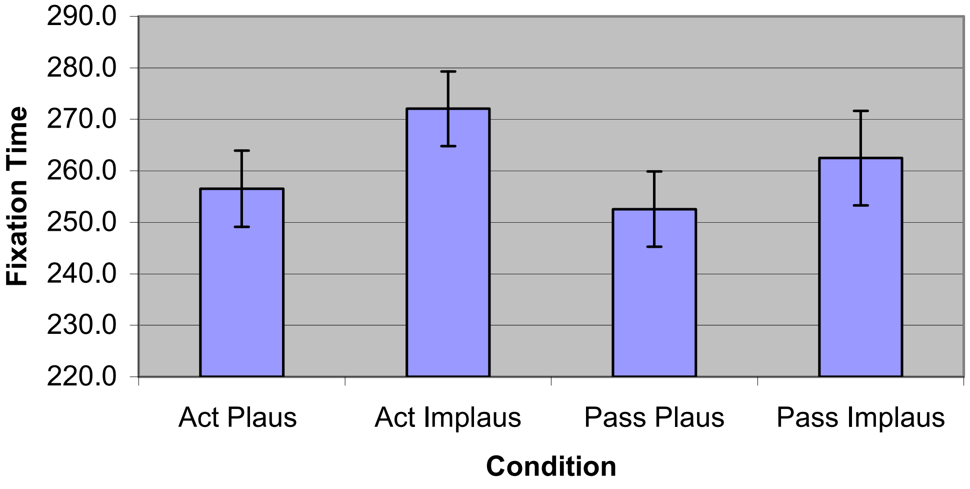


Region 8 (second spillover region)

First fixation analysis revealed no significant main effect of voice ($F_1(1,31)$ = 1.463, MSe = 1006.064, $p > 0.2$; $F_2(1,31)$ = 2.757, MSe = 844.757, $p > 0.1$).

However, there was a main effect of plausibility, significant by participants and marginal by items, with longer initial fixation times in the implausible conditions (F_1(1,31) = 7.562, MSe = 684.580, $p < 0.05$; F_2(1,31) = 3.868, MSe = 1001.113, $p = 0.058$) (see Figure 15). The interaction was not significant (both *F*s < 1). This is the first clear effect indicating disruption in the Active conditions. This plausibility effect was short-lived however, as there was no significant effect of plausibility in the analysis of first pass reading times (F_1(1,31) = 2.338, MSe = 5770.128, $p > 0.1$; $F_2 < 1$). For the first time, there was a significant effect of voice, with longer first pass times in the Active conditions (F_1(1,31) = 4.904, MSe = 6616.394, $p < 0.05$; F_2(1,31) = 5.497, MSe = 6202.633, $p < 0.05$). The interaction was not significant (both *F*s < 1).

Figure 15
First Fixation Time (msec), Region 8



Analysis of first pass regressions out and regression path time yielded no significant effects (all *p*s > 0.1). There was no effect of voice in total time analysis (both *F*s < 1), and the effect of plausibility was significant by participants only, with longer total reading times in the implausible conditions ($F_1(1,31) = 4.842$, MSe = 11390.169, $p < 0.05$; $F_2(1,31) = 2.745$, MSe = 18161.064, $p > 0.1$). The interaction was not significant ($F_1(1,31) = 1.014$, MSe = 11865.060, $p > 0.3$; $F_2 < 1$).

Response-contingent analysis

As with earlier experiments this analysis was not possible due to sparse data problems. There were 12 participants who had made errors in all four conditions, but they were unevenly spread across the 4 experimental lists. Only lists 1, 2 and 4 were represented, resulting in a serious imbalance in the latin square design.

Discussion

Interpretations

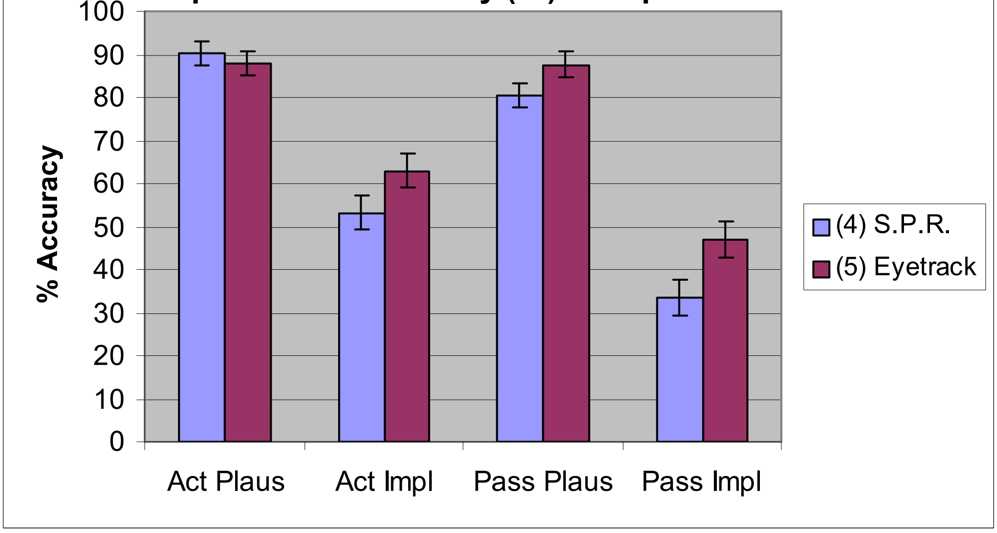
The first point to note is that accuracy results were similar to those in experiment 4, though accuracy was somewhat improved in the present study. The significant main effect of plausibility indicated that readers were poorer at interpreting the implausible materials, and were thus susceptible to shallow processing and normalisation. The presence of a significant interaction indicated that readers had more difficulty interpreting implausible passive EVPs than implausible active EVPs, but were equally successful when interpreting plausible actives and plausible passives. The direct comparison of the two implausible conditions provides further support for the idea that implausible passive structures are particularly difficult to interpret correctly, most likely due to the requirement of assigning thematic roles in an atypical order. We can conclude that this tendency to process implausible passive structures differently to implausible actives is robust, occurring here under conditions of free reading and higher overall accuracy.

Cross-experiment analysis

As the materials used in experiments 4 and 5 were identical in content, we performed a cross-experiment analysis to directly test the appearance of higher accuracy levels. Recall that the differing methods of presentation were expected to differ in terms of the strain they each placed on memory, and to result in improved interpretations in the less demanding, free-reading conditions of experiment 5. The mean accuracy results for both experiments are presented in Figure 16 for ease of comparison.

A mixed ANOVA (including Experiment as a between-subjects factor) yielded a significant effect of the between-subjects factor: readers were more accurate overall in experiment 5 ($F_1(2, 63) = 4.980$, MSe = 617.814, $p < 0.05$). The between-subjects factor did not interact significantly with the two within-subjects factors, but the Plausibility*Experiment interaction was marginally significant ($F_1(2, 63) = 3.117$, MSe = 432.503, $p = 0.082$). A look at the means suggests the difference, if real, lies in the interpretations of the Implausible conditions, with greater accuracy in experiment 5.

Figure 16
Comparison of Accuracy (%) in Experiments 4 & 5



Processing results

As with the results of experiment 4, there is evidence that anomalies were detected online. The difference here is that the detection effects are visible earlier in the processing record – one word earlier – giving support to the idea that the relatively late anomaly detection in experiment 4 was linked to the self-paced-reading methodology (i.e. an artefact of repetitive button-pressing).

Overall, effects were fewer and weaker than those observed in the self-paced reading version, which may indicate that these anomalies, while disruptive, are slightly easier to deal with under free-reading conditions. One slightly surprising feature of the processing data was the fact that, in terms of robust evidence of anomaly-related disruption, effects appeared first for the passive implausible condition, with the disruption in the active conditions not clearly visible until the following region. Keeping a very open mind, the trend in the regressions analysis for the active conditions (critical region) may have indicated immediate anomaly detection. But that would entail a picture in which the disruption died down before the following region, only to re-emerge in the region following that; and this would be difficult to explain. The simpler picture, then, is that the active implausible condition didn't cause any difficulty until relatively late, and the difficulty was on a very small scale. The difficulty with the passive case, though beginning earlier, took longer to resolve.

What this amounts to is evidence that, while the active and passive EVPs can be differentiated in terms of the interpretations ultimately assigned to them, that difference does not translate to online processing behaviour. In our earlier experiments there was no evidence of passives being processed differently to actives, and now we see a difference that we wouldn't have predicted, namely, that 'harder' passive anomalies are detected prior to active, 'easier' ones. If interpretation difficulties do not obviously correlate with processing difficulties, then this bolsters the view that syntactic interpretation – whose processes we observe – is primary, and heuristic processes secondary, dependent, as seems likely, on our memory for what we have read.

So, as there is again clear evidence of online detection of implausible material, the question of the time course of syntax-based processing seems settled at 'early', with the high rates of misinterpretation presumably due to the operation of later interpretative processes based on heuristics. Given that the disruption carries on late into the processing stream, we could maybe go further and say that the heuristics themselves are not likely to operate online. The application of a preferred heuristic interpretation, after the syntactic-interpretation, would be expected to neutralise the processing disruption; and, therefore, in the case of the most heavily normalised condition, to produce a lack of plausibility effects before the end of the sentence. But, as we have seen, the passive implausible condition is still seen to cause difficulty in the penultimate region. We could conclude, tentatively, that heuristic interpretation is an offline phenomenon.

Experiment 6:
The Interpretation And Processing Of Implausible Sentences –
A Further Eyetracking Study

Introduction

Experiment 6 was a further investigation into the interpretation and processing time course of implausible sentences. Eyetracking was used again as the most sensitive measure of anomaly detection and thus the best index of interpretation based on a syntactic frame. Participants completed the Nelson Denny reading comprehension test as a measure of reading skill and an index of reading span. The overall methodology and running of the experiment was therefore very similar to that of experiment 5.

For this study a material set was constructed using verbs that were either pragmatically biased or unbiased towards the noun phrase to which the thematic role of agent was assigned. For example, in (5)

(5)After quizzing the defendant, the lawyer noticed the journalist writing in his notebook

the NP *lawyer* is a more plausible agent of the activity of ‘quizzing a defendant’ than is *journalist*. On the other hand, a reversal of these two NPs to give the sentence

(6) After quizzing the defendant, the journalist noticed the lawyer…

results in the agent role for ‘quizzing’ being assigned to the less plausible *journalist*. In each case, the NP designated as agent is unambiguously identified by the syntax (through control relations). If thematic roles are correctly assigned when reading these two sentences, we would expect to see lengthened reading times on the less plausible agent NP in the second sentence (‘journalist’) relative to the agent NP in the first sentence (‘lawyer’). However, if interpretation is not being driven by syntax at this stage, we might not see any difference in reading times on this critical NP, and see either differences emerging further downstream or not at all. And if final interpretations are being informed by shallow processing based on the plausibility of a given NP occupying a given thematic role, then we would expect to see less accurate responses to questions probing the agent role in the second sentence (6) compared with the first sentence (5).

Method

Participants

32 members of the University of Glasgow student population were paid for their participation. All had normal or corrected-to-normal vision, were native English speakers and had not been diagnosed with dyslexia. All were naïve to the design and goals of the study.

Materials and design

A set of 28 materials was constructed based on the type of sentence described above. Critical sentences consisted of an opening subordinate clause ('*after quizzing the defendant*') followed by a matrix clause whose subject NP constituted the agent of the subordinate clause ('*the lawyer*'), and which specified a further action performed by this agent ('*noticed*'), and the patient (or theme) of this action (*the journalist…*'). In the context of the matrix clause and for the purposes of description, we will term the Agent NP 'NP1', and the Patient (or theme) NP 'NP2'. The choice of NP to fill these slots determined whether the sentence would be plausible or implausible, i.e. if NP1 was a relatively plausible agent of the verb in the subordinate clause then the sentence was plausible; if it was a relatively implausible agent, then the sentence was implausible overall. Another way of saying this is that the verb in the subordinate clause *biases* towards either NP1 or NP2 as a plausible filler of the agent role. The sentence was therefore rendered plausible or implausible by the order of the two NPs contained in the matrix clause, and the two possible variations of the critical sentence will be termed ‘Biased Order 1’ and ‘Biased Order 2’.

As NP1 is the point at which the sentence could become implausible, and since this would mean measuring reading times on different words in the two conditions, it was necessary to control for differing lexical properties of the different words by creating two control conditions. These conditions contained subordinate clause verbs that did not bias towards either NP1 or NP2 as a plausible agent of the action they described, and were labelled Neutral Order 1 and Neutral Order 2. As these conditions contained no biasing verb, they were both equally plausible; the only plausibility difference would be between the two biased conditions. Biased Order 1 is plausible and Biased Order 2 is implausible. This resulted in a 2x2 design: Factor 1 was subordinate verb type (‘Verb’) which could be either Biased or Neutral, and factor 2 was the order of the two matrix clause NPs, either Order 1 or Order 2 (‘Order’). Critical sentences were always preceded by a context sentence that was held constant across all four conditions. Table 23, below, presents an experimental item in each of its four conditions.

Table 23: Experimental Materials by Condition (with questions)

Condition	Sentence
Neutral Verb Order 1 **(Plausible)**	*The courtroom gallery was completely full. After listening to the defendant, the lawyer noticed the journalist writing in his notebook.*
Question	Who listened to the defendant? Journalist <> lawyer
Neutral Verb Order 2 **(Plausible)**	*The courtroom gallery was completely full. After listening to the defendant, the journalist noticed the lawyer writing in his notebook.*
Question	Who listened to the defendant? Lawyer <> journalist
Biased Verb Order 1 **(Plausible)**	*The courtroom gallery was completely full. After quizzing the defendant, the lawyer noticed the journalist writing in his notebook.*
Question	Who quizzed the defendant? Journalist <> lawyer
Biased Verb Order 2 **(Implausible)**	*The courtroom gallery was completely full. After quizzing the defendant, the journalist noticed the lawyer writing in his notebook.*
Question	Who quizzed the defendant? Lawyer <> Journalist

Plausibility data was gathered in an attempt to confirm experimenter intuition. As discussed in the previous chapter, there are certain, currently unresolved, problems with doing this. However, as the method used in experiment 1 had proven highly successful, it was reemployed here in the absence of a more credible alternative. 24

participants rated 28 items; mean plausibility ratings are presented in Table 24.

Table 24: Mean Plausibility Ratings for Each Condition

	Neutral 1	Neutral 2	Bias 1	Bias 2
	Mean (SD)			
Mean Plausibility Rating (7 - point scale)	4.85 (1.36)	4.67 (1.40)	5.34 (1.05)	2.87 (0.74)

There was a main effect of verb type ($F_1(1, 23) = 14.587$, MSe = 0.634, $p < 0.005$; $F_2(1, 27) = 10.996$, MSe = 1.019, $p < 0.005$) and a main effect of Order ($F_1(1, 23) = 34.660$, MSe = 1.108, $p < 0.001$; $F_2(1, 27) = 28.055$, MSe = 1.745, $p < 0.001$). There was also a significant Verb*Order interaction ($F_1(1, 23) = 45.727$, MSe = 0.608, $p < 0.001$; $F_2(1, 27) = 25.490$, MSe = 1.439, $p < 0.001$). Planned comparisons indicated that there was no significant difference between the two Neutral conditions (both *F*s < 1) and that there was a significant difference between the two Biased conditions, with the Biased Order 1 condition rated more plausible ($F_1(1,23) = 76.043$, MSe = 0.773, $p < 0.001$; $F_2(1, 27) = 58.201$, MSe = 0.103, $p < 0.001$). This analysis therefore confirmed experimenter intuition and suggested the materials had the necessary semantic properties for the eyetracking study. The mean ratings for the two neutral verb conditions indicated that the participants judged these to be neither highly plausible nor highly implausible (‘neutral’, in fact). The biasing effect of the verbs in the two biased conditions is clear, with the Biased Order 1 condition receiving the highest (most plausible) ratings overall – the good semantic/pragmatic fit of the verb with its agent led participants to feel it made especially good sense, as might be expected. The Biased Order 2 condition, as predicted, was rated least plausible.

Each experimental material was followed by an interpretation question (two-alternative, forced-choice) that probed the assignment of thematic roles. If a reader read the sentence

After quizzing the defendant, the journalist noticed the lawyer writing in his notebook…

And correctly answered the question

Who quizzed the defendant?
Lawyer < > Journalist

this would indicate that the role of Agent had been correctly assigned to the implausible *journalist*, as signalled unambiguously by the syntax. An incorrect answer, i.e. *lawyer*, would indicate that the agent role had incorrectly been assigned to the more plausible NP, contrary to the information contained in the syntax.

The 28 materials for experiment 6 were randomised according to a latin square design, combined with the materials for experiment 5 and spilt across four experimental lists. As well as the filler materials for experiment 5 there were a further 32 fillers based on the experiment 6 materials, all of which were plausible in meaning. With the experiment 6 materials and filler materials, the correct answer was presented on the left hand side 50% of the time and on the right 50% of the time to guard against answering strategies not based on comprehension.

Apparatus

The apparatus was identical to that described in experiment 5.

Procedure

The procedure was identical to that described in experiment 5.

Data Analysis

Eyetracking procedures and measures were the same as those outlined for experiment 5. As with the previous eyetracking experiment, the materials for experiment 6 were split into regions for the purposes of analysis. An example is given below:

$_1$The courtroom gallery was completely full.| $_2$After listening to the defendant,| $_3$the lawyer| $_4$noticed| $_5$the journalist| $_6$writing in| $_7$his notebook|

Region 1 consisted of the opening context sentence. Region 2 consisted of the entire subordinate clause. Again, for the sake of completeness, we will report reading measures for this region to allow for possible parafoveal detection of the anomaly. Region 3, the critical region, contained the subject NP of the matrix clause; region 4 contained the main verb; region 5 contained the object NP of the matrix clause. Regions 6 and 7 contained the final four words of the critical sentence which served as further spillover regions and tended to consist of a further verb phrase or a prepositional phrase. A straightforward 50-50 division of these four words was sufficient to create regions 6 and 7.

Results

Interpretations

We will first look at the results of the interpretation task. Accuracy means (% correct) can be found in Table 25 and are also presented in Figure 17

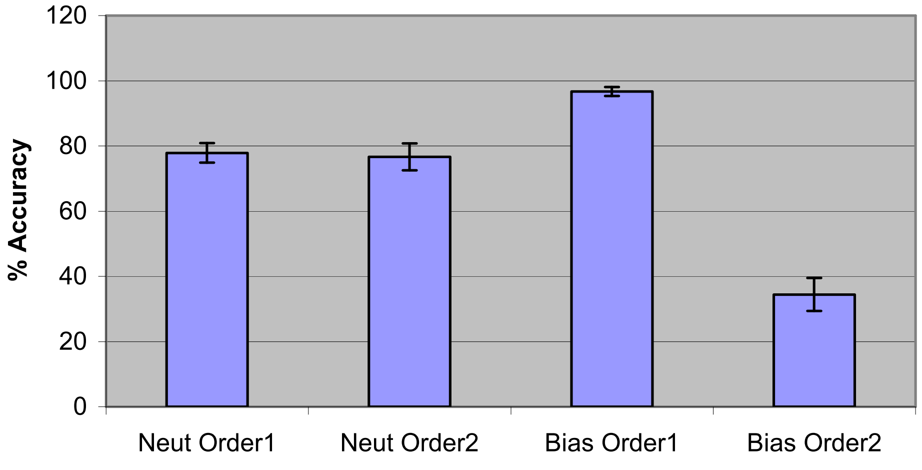
Table 25: % Accuracy in the interpretation task

	Condition			
	Neutral Order 1	Neutral Order 2	Biased Order 1	Biased Order 2
% Accuracy	77.86 (16.92)	76.64 (23.42)	96.68 (7.79)	34.44 (28.42)

Analysis of variance yielded a significant effect of Verb type, with higher accuracy in the Neutral conditions ($F_1(31) = 20.999$, MSe = 208.173, $p < 0.001$). There was also a significant effect of Order, with higher accuracy in the Order 1 conditions ($F_1(31) = 73.732$, MSe = 436.795, $p < 0.001$). These effects were modulated by a significant interaction ($F_1(31) = 63.305$, MSe = 442.576, $p < 0.001$) (see Figure 17). Planned comparisons indicated that there was no reliable difference between the two Neutral conditions ($t_1 < 1$) and that there was a reliable difference between the two Biased conditions ($t_1(31) = 11.759$, $p < 0.001$).

As predicted by the results of the norming study, the Biased Order 1 condition was by far the easiest to comprehend and interpret correctly. The two Neutral conditions were interpreted somewhat less successfully but did not differ between themselves. While the accuracy is well above chance in these conditions, the lack of a biasing (you might say 'helpful') verb resulted in lower accuracy as compared with the Biased Order 1 condition. Accuracy was extremely low in the Biased Order 2 condition – below chance level at 34% (significantly below, as confirmed by a one-sample t-test (t (31) = 3.096, $p <$ 0.005). Readers were clearly not interpreting according to the syntactically specified meaning a high percentage of the time, and this is therefore yet more compelling evidence for shallow processing and pragmatic normalisation. It must also be said that it is very surprising to see such low accuracy under free reading conditions. While no specific predictions were made on expected accuracy levels, it is unlikely that any predictions would have allowed for levels as low as this.

Figure 17: Interpretation Accuracy (%)



Reading time analysis

To preview the results, the earliest evidence of anomaly detection, and thus correct interpretation, was observed in the first spillover region. Effects on the critical region itself were seemingly confounded due to unknown lexical properties differing between the Order 1 and Order 2 critical NPs (e.g. *child* and *mugger*). Having controlled the two groups of critical NP (for the Order 1 and Order 2 conditions) for length and frequency, analyses in this region yielded unexpected main effects of the Order variable and so were clearly confounded by uncontrolled lexical factors (this also affected Region 5 which consisted of the same NPs). However, looking at all results in region 3, there is no reason to think that our predicted effects were masked by this confound. So while results for these regions will still be reported, the main region of interest is region 4, and the plausibility effect in *first fixation time* in this region (with the means following the predicted interaction) strongly suggests early application of syntax-based interpretation. Mean reading time results are presented in Table 26, below.

Table 26: Mean Reading Time Measures For Critical Region and 4 Spillover Regions

			Region			
	Region 2 (pre-critical) after frightening…	**3 (critical)** the child	**4** spotted	**5** the mugger	**6** and ran	**7** off quickly
Measure	Mean (SD)					
First Fixation (msec)						
Neutral Order 1	-	255 (36)	261 (37)	259 (38)	254 (53)	269 (48)
Neutral Order 2	-	252 (41)	255 (45)	267 (69)	245 (44)	269 (47)
Biased Order 1	-	249 (39)	259 (40)	256 (42)	247 (41)	258 (43)
Biased order 2	-	257 (37)	274 (46)	274 (53)	249 (44)	266 (46)
First Pass (msec)						
Neutral Order 1	837 (236)	302 (53)	363 (113)	366 (95)	278 (61)	447 (146)
Neutral Order 2	815 (228)	332 (95)	359 (94)	338 (88)	293 (85)	427 (115)
Biased Order 1	840 (287)	297 (76)	365 (102)	330 (84)	273 (55)	410 (120)
Biased order 2	828 (260)	330 (69)	398 (128)	349 (91)	278 (55)	464 (128)
First Pass Reg. Out (%)						
Neutral Order 1	6.38 (12.96)	19.22 (16.29)	7.91 (11.25)	13.50 (14.68)	17.22 (26.84)	46.06 (31.68)
Neutral Order 2	8.72 (13.69)	17.63 (16.66)	10.09 (12.76)	17.13 (20.65)	9.25 (16.05)	44.72 (35.23)
Biased Order 1	6.75 (13.56)	16.63 (15.30)	4.72 (8.36	12.16 (16.03)	14.19 (23.07)	41.00 (24.52)
Biased order 2	9.19 (13.31)	17.31 (17.76)	8.38 (9.83)	18.16 (23.22)	12.34 (27.37)	49.88 (30.87)
Regression Path (msec)						
Neutral Order 1	894 (265)	455 (179)	396 (122)	430 (134)	376 (168)	938 (673)
Neutral Order 2	918 (262)	445 (175)	444 (177)	424 (125)	336 (130)	966 (845)
Biased Order 1	899 (282)	408 (190)	391 (122)	417 (157)	546 (1154)	734 (464)
Biased order 2	955 (291)	451 (202)	466 (176)	473 (200)	439 (556)	933 (767
Total Time (msec)						
Neutral Order 1	996 (328)	425 (154)	464 (147)	464 (169)	369 (118)	546 (180)
Neutral Order 2	1022 (355)	482 (171)	479 (137)	447 (180)	353 (109)	513 (179)
Biased Order 1	999 (418)	399 (144)	423 (144)	431 (141)	343 (127)	472 (148)
Biased order 2	1027 (354)	407 (209)	522 (210)	440 (144)	344 (80)	530 (154)

Pre-critical Region (region 2)

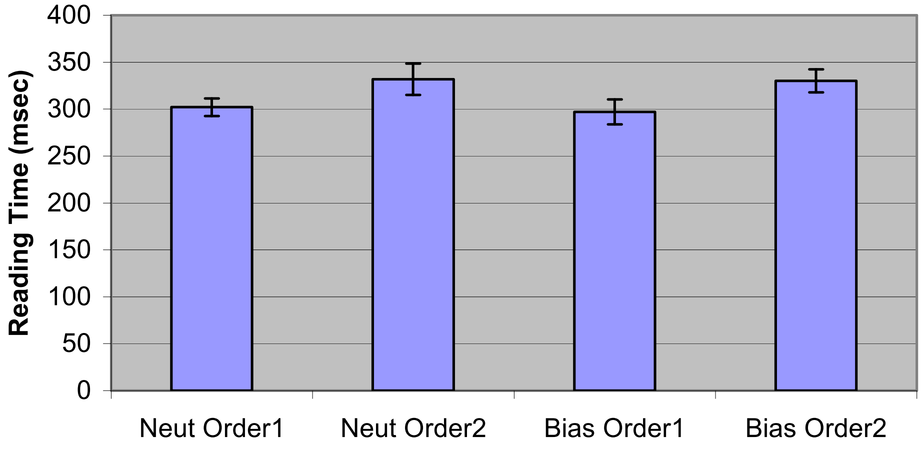
Analysis in this region will not include first fixation analysis due to its length. We begin, therefore, with first pass analysis. This analysis revealed no significant main effects or interactions (all *F*s < 1). In first pass regressions out there was no significant effect of Verb (both *F*s < 1). There was a marginal main effect of the order variable – participants analysis only – with more regressions out of this region in the Order 2 conditions ($F_1(1, 31) = 3.167$, MSe = 57.754, $p = 0.085$; $F_2(1, 28) = 1.152$, MSe = 93.776, $p > 0.2$). See the analysis of the critical region, and following, for a discussion of this unexpected effect. In regression path analysis showed no significant effect of Verb (both *F*s < 1), no effect of Order

(F_1(1, 31) = 2.649, MSe = 19859.232, $p > 0.1$; F_2(1, 28) = 2.031, MSe = 23842.631, $p > 0.1$), and no significant interaction (both *F*s < 1). In total time there was no effect of verb (both *F*s < 1), no effect of order (F_1(1,31) = 1.132, MSe = 20918.562, $p > 0.2$; $F_2 < 1$), and no interaction (both *F*s < 1). There is therefore no indication that the anomaly in the Biased Order 2 condition was detected parafoveally (but see chapter 6 for some evidence that the anomaly may have been detected parafoveally by the less-skilled readers).

Critical Region (region 3)

First fixation analysis yielded no significant effects (Verb and Order main effect: all *F*s < 1; Interaction F_1(1,31) = 1.086, MSe = 1244.894, $p > 0.3$; $F_2 < 1$). In first pass analysis there was no significant effect of verb (both *F*s < 1) and no interaction (both *F*s < 1) but there was a significant effect of order, with longer reading times in the Order 2 conditions (F_1(1,31) = 7.048, MSe = 4487.141, $p < 0.05$; F_2(1,27) = 9.627, MSe = 2870.337, $p < 0.05$) (see Figure 18).

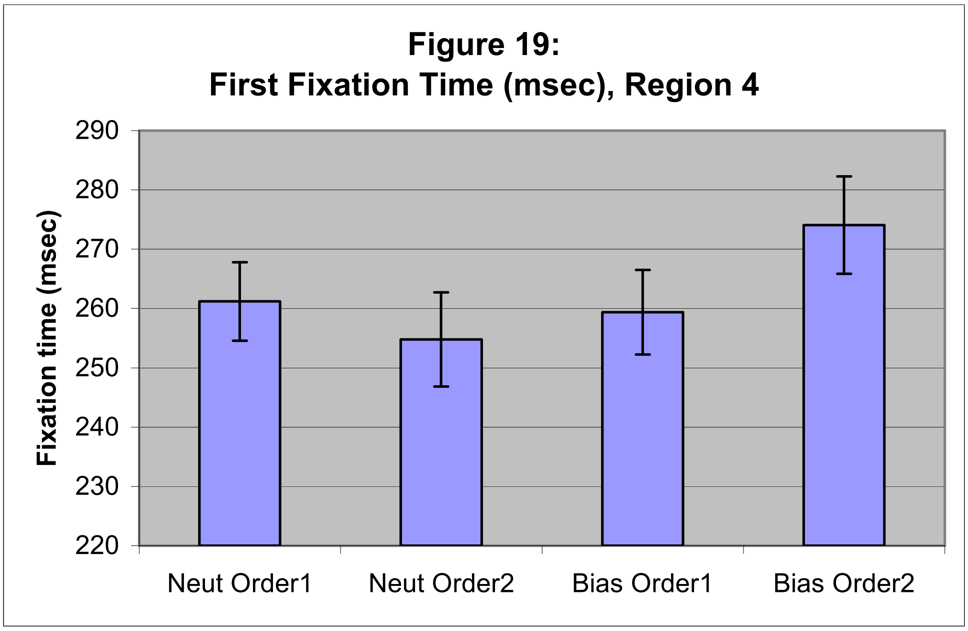
Figure 18:
First Pass Time (msec), Region 3



There were no significant effects in first pass regressions out (all *F*s < 1) or regression path analysis (all *p*s > 0.2). Total time analysis showed no effect of verb ($F_1 < 1$; $F_2(1,27) = 1.366$, MSe = 8040.914, $p > 0.2$), and no interaction (both *F*s < 1). But there was again a main effect of order, with longer total reading times in the Order 2 conditions ($F_1(1,31) = 16.569$, MSe = 10098.786, $p < 0.001$; $F_2(1,27) = 15.378$, MSe = 10393.451, $p < 0.005$).

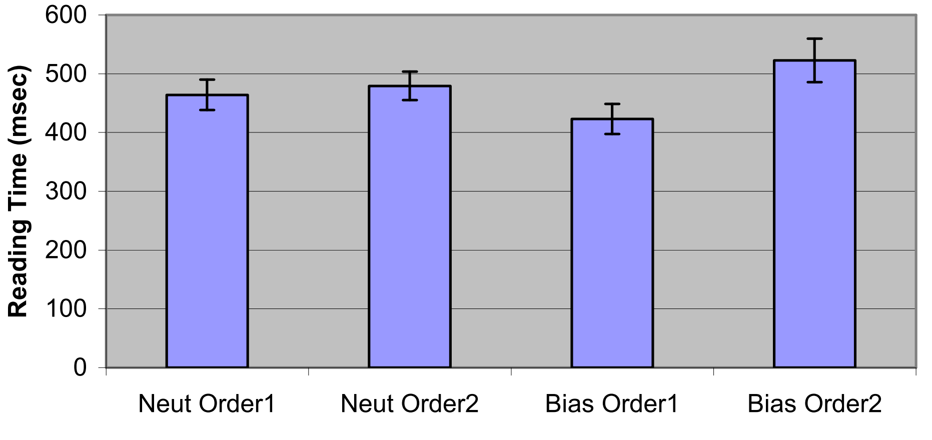
Region 4 (first spillover region)

In first fixation analysis there was verb*order interaction that was marginally significant by participants ($F_1(1,31) = 3.953$, MSe = 900.423, $p = 0.053$; $F_2(1,27) = 2.162$, MSe = 1177.628, $p > 0.1$) (see Figure 19). T-tests indicated there was no significant difference between the two Order 1 conditions ($t_1 < 1$; $t_2 < 1$), and that there was a difference between the two Order 2 conditions, significant by participants and marginal by items, such that there were longer initial fixation times in the Biased Order 2 condition ($t_1(31) = 2.103$, $p < 0.05$; $t_2(27) = 1.796$, $p = 0.084$). This difference between the two Order 2 conditions can be put down to the plausibility difference, reflected in the norming study, and is thus evidence that correct interpretation of the Biased Order 2 condition at this stage had resulted in disruption relative to the Neutral Order 2 condition. The effect of verb was not significant ($F_1(1,31) = 2.279$, MSe = 1071.209, $p > 0.1$; $F_2(1,27) = 1.463$, MSe = 853.411, $p > 0.2$), nor was the effect of order ($F_1(1,31) = 1.005$, MSe = 545.713, $p > 0.3$; $F_2 < 1$).



This plausibility effect did not spillover into the first pass measure (all *p*s > 0.1). In first pass regressions out there was some evidence of a spillover of the Order main effect in region three ($F_1(1,31) = 2.524$, MSe = 108.260, $p > 0.1$; $F_2(1,27) = 3.491$, MSe = 76.557, $p = 0.073$) but no other significant effects (all *p*s > 1). Similarly in regression path time, neither the effect of verb nor the interaction were significant (all *F*s < 1) but the effect of order was significant, with longer regression path times in the Order 2 conditions ($F_1(1,31) = 4.417$, MSe = 27109.397, $p < 0.05$; $F_2(1,27) = 8.857$, MSe = 12711.594, $p < 0.05$). Lastly, total time analysis yielded an effect of Order, marginal by participants and significant by items, with longer total times in the Order 2 conditions ($F_1(1,31) = 4.055$, MSe = 17899.468, $p = 0.053$; $F_2(1,27) = 7.113$, MSe = 10709.649, $p < 0.05$). There was no effect of verb (both *F*s < 1) and the interaction was only marginal by participants $F_1(1,31) = 3.160$, MSe = 10594.362, $p = 0.085$; $F_2(1,27) = 1.916$, MSe = 13455.075, $p > 0.1$) (see Figure 20).

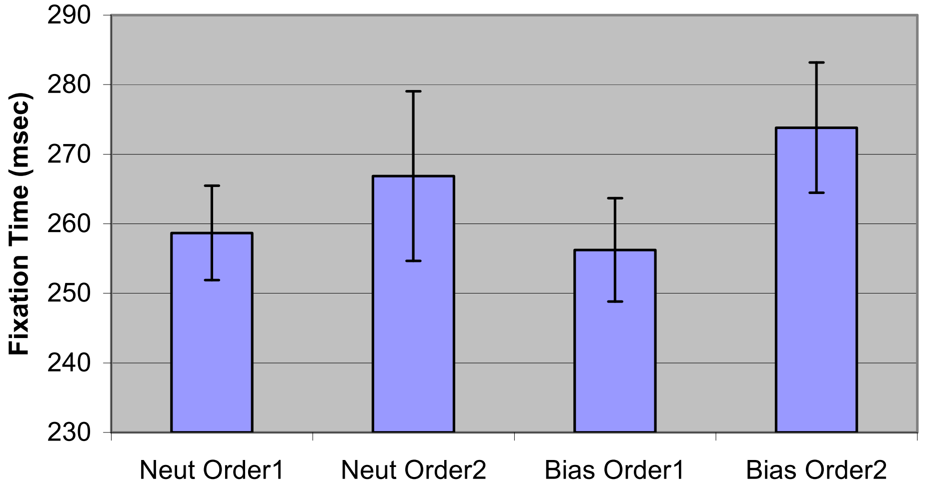
Figure 20:
Total Reading Time (msec), Region 4



Region 5 (second spillover region)

First fixation analysis yielded no significant effects of voice and no significant interaction (all *F*s < 1). There was, however, a main effect of order, with longer initial fixation times in the Order 2 conditions ($F_1(31) = 4.176$, MSe = 1267.149, $p = 0.05$; $F_2(27) = 3.576$, MSe = 1757.583, $p = 0.069$) (see Figure 21).

Figure 21:
First Fixation (msec) Region 5



As the standard errors appeared to show a difference between the two Biased conditions, and since a plausibility difference was predicted here, t-tests were conducted for the Neutral and Biased conditions separately, and indicated that there was no difference between the two Neutral conditions ($t_1 < 1$; $t_2(27) = 1.040$, $p > 0.3$) and that there was a difference between the two Biased conditions, significant by participants and marginal by items, with longer fixation times in the Biased Order 2 condition ($t_1(31) = 2.070$, $p < 0.05$; $t_2(27) = 2.013$, $p = 0.054$). So despite the analysis still being troubled by the NP confound, there is some evidence of spillover from region 4 caused by the plausibility difference between the two Biased conditions. First pass analysis contained no significant main effects (all *F*s < 1) and no significant interaction ($F_1(31) = 3.058$, MSe = 3734.463, $p > 0.05$; $F_2(27) = 1.250$, MSe = 3848.544, $p > 0.2$). Analysis of first pass regressions out contained no significant effect of verb and no significant interaction (all *F*s < 1) but a marginal effect of order, with more regressions in the Order 2 conditions ($F_1(31) = 3.773$, MSe = 196.431, $p = 0.061$; $F_2(27) = 2.576$, MSe = 245.250, $p > 0.1$). Regression path analysis yielded no significant effect of verb (both *F*s < 1) or order ($F_1(31) = 3.050$, MSe = 6426.048, $p > 0.05$; $F_2(27) = 1.720$, MSe = 14392.379, $p > 0.2$) and no interaction ($F_1(31) = 2.280$, MSe = 13813.935, $p > 0.1$; $F_2 < 1$). In total time analysis the was no significant effect of verb ($F_1(31) = 1.565$, MSe = 8243.518, $p > 0.2$; $F_2(27) = 2.690$, MSe = 8413.276, $p > 0.1$), no effect of order and no interaction (all *F*s < 1).

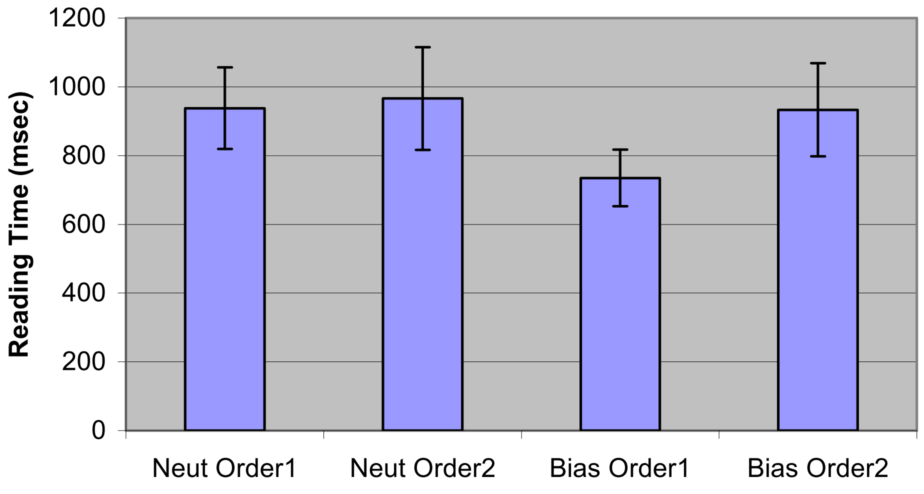
Region 6 (third spillover region)

There were no significant effects in this region and no effects approaching significance.

Region 7 (wrap up)

Effects in this region generally reflected the results of the norming study, generating effects consonant with the overall comprehensibility of the sentences as indicated by plausibility judgements. First fixation analysis yielded no effects of verb ($F_1(1, 31) = 1.328$, MSe = 1154.582, $p > 0.2$; $F_2(1, 27) = 2.293$, MSe = 683.654, $p > 0.1$) or order ($F_1 < 1$; $F_2(1, 27) = 1.187$, MSe = 833.877, $p > 0.2$) and no interaction (both Fs < 1). In first pass analysis there was no effect of verb (both Fs < 1) or order ($F_1(1, 31) = 1.465$, MSe = 5979.764, $p > 0.2$; $F_2(1, 27) = 1.008$, MSe = 5625.509, $p > 0.3$). But there was a significant interaction ($F_1(1, 31) = 5.624$, MSe = 7980.764, $p < 0.05$; $F_2(1, 27) = 4.887$, MSe = 6178.205, $p < 0.05$). T-tests revealed that there was no reliable difference between the two Neutral conditions ($t_1(31) = 1.030$, $p > 0.3$; $t_2(27) = 1.035$, $p > 0.3$) and a reliable difference between the two biased conditions ($t_1(31) = 2.515$, $p < 0.05$; $t_2(27) = 2.065$, $p < 0.05$). In first pass regressions out there was no effect of verb or order (all Fs < 1), and no interaction ($F_1(1, 31) = 2.478$, MSe = 337.157, $p > 0.1$; $F_2 < 1$). In regression path time there was, for the first time, an effect of verb with longer regression path times in the Neutral conditions ($F_1(1,31) = 10.267$, MSe = 43422.547, $p < 0.005$; $F_2(1,27) = 4.807$, MSe = 123808.119, $p < 0.05$). There was also a main effect of order, significant b participants only ($F_1(1, 31) = 4.262$, MSe = 96192.577, $p < 0.05$; $F_2(1, 27) = 1.930$, MSe = 130422.060, $p > 0.1$). These effects were modulated by a significant interaction, significant only in the participants analysis ($F_1(1, 31) = 10.147$, MSe = 22863.770, $p < 0.005$; $F_2(1, 27) = 1.362$, MSe = 121107.620, $p > 0.2$). T-tests confirmed there was no difference between the two neutral conditions (both ts < 1) and a difference, significant by participants and marginal by items, between the biased conditions ($t_1(31) = 2.760$, $p < 0.05$; $t_2(27) = 1.953$, $p = 0.061$) (see Figure 22).

Figure 22:
Regression Path Time (msec), Region 7



A similar pattern of effects was observed in the total time analysis. There was a marginal effect of verb by participants only ($F_1(1, 31) = 3.238$, MSe = 7964.818, $p = 0.082$; $F_2(1, 27) = 2.110$, MSe = 10269.856, $p > 0.1$). There was no effect of order (both Fs < 1) but there was an interaction, significant by participants and marginal by items ($F_1(1, 31) = 17.331$, MSe = 3846.165, $p < 0.001$; $F_2(1, 27) = 3.259$, MSe = 12546.508, $p = 0.082$) (see Figure 23). Comparisons of the neutral conditions yielded no reliable difference ($t_1(31) = 1.445$, $p > 0.1$; $t_2(27) = 1.208$, $p > 0.2$) and comparison of the biased conditions yielded a reliable difference (by participants only) between the biased conditions ($t_1(31) = 2.495$, $p < 0.05$; $t_2(27) = 1.653$, $p > 0.1$)

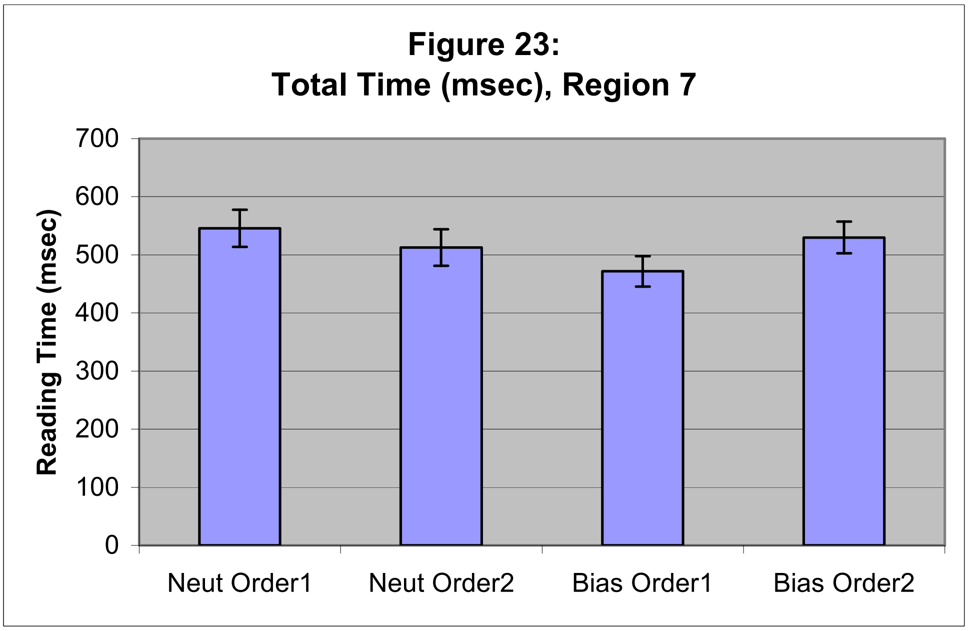


Figure 23:
Total Time (msec), Region 7

Discussion

Interpretations

Accuracy in the two Neutral verb conditions was not as high as might have been expected given their unambiguous nature, with both falling short of 80% correct. These accuracy levels may therefore be indicative of this type of construction being relatively difficult to comprehend correctly. For one thing, referential load is high (J. S. Sanford, A. J. Sanford, Filik and Molle, 2005; Warren and Gibson, 2002; Warren, 2001) , with three definite NPs as opposed to, say, the two definite NPs in experiment 2 where accuracy in the plausible conditions was around 90%. It is also possible that the grammatical control relation that requires the first NP of the main clause to be the unexpressed subject of the verb in the subordinate clause is not as strong as might be expected. This can be seen in real-life examples such as “After losing her job, Sandra’s life began to fall apart”. Here, the control relation would require ‘Sandra’s life’ to be the individual that lost the job, but the preferred interpretation makes ‘Sandra’ that individual, thereby violating the constraint. Any difficulties, however, were greatly eased by the use of a biasing verb, as evidenced by the near ceiling-level accuracy in the Biased Order 1 condition. The very low accuracy in the implausible Biased Order 2 condition certainly indicates a high proportion, in fact a majority proportion, of non-grammatical interpretations. Readers were clearly led to normalise their interpretations based on the more plausible scenario suggested by the connection between the verb and it’s most likely agent (e.g. *quizzed* and *lawyer*). As mentioned in the results section, the rates of normalisation were surprisingly low, and point again to the error (pointed out by Ferreira) in assuming that participants in psycholinguistic experiments will comprehend accurately even *most* of the time when pragmatic cues bias away from a strictly syntax-based interpretation.

Processing results

The evidence of online disruption in the implausible Biased Order 2 condition, caused by correct interpretation, as well as the high accuracy in the two Neutral conditions, demonstrates that readers were not interpreting solely according to semantics-based heuristics. On the question of our main interest – how comprehenders process those sentences which they later misinterpret – the results do not give us a definitive answer. We should say first that online disruption effects, while certainly visible, were not as prevalent or robust as they had been in previous studies. Only direct comparisons in regions 4 and 5 revealed evidence of correct online interpretation. So we might expect that reducing the power by conducting an analysis of incorrectly answered trials only would make any evidence of correct online interpretation very difficult to detect. As it was, with such high accuracy in the first three conditions, we had to restrict this analysis to only the two Order Two conditions, and only to the analysis of first fixation time in region 4 (the earliest evidence of correct interpretation in the full analysis). This analysis did not yield a significant difference between the plausible and the implausible condition ($F_1 < 1$). Was this because readers who made interpretation mistakes in the implausible condition never interpreted correctly in the first place, or because the reduced power obscured a weak effect? The former option would support the 'either syntax or semantics' account of interpretation, but again, we would be rash to endorse it on the current evidence. Instead, on balance, we maintain that it is safer to say the evidence supports the 'syntax first' account.

Anomaly detection effects

On the issue raised in the introduction to experiment 5 – the timing of anomaly detection effects – our results would seem to be in line with those findings in the literature of 'late' effects for semantic/pragmatic anomalies. While the eyetracking methodology has enabled us to observe effects somewhat earlier than those seen in the self-paced reading experiments, the anomaly effects in experiments 5 and 6 are not what we could call 'first-pass', i.e. occurring immediately on the anomalous region. Our present results therefore do not challenge any theory, such as that put forward by Boland (2004), which attributes immediate anomaly effects to violations of a syntactic nature, and are in line with the view that violations of a semantic/pragmatic nature appear slightly later in the processing record.

Chapter 5:
Contextual Fit, Shallow Processing and the Time Course of Interpretation

Introduction

A number of studies have demonstrated that readers (and listeners) will fail to detect a serious semantic anomaly when the fit of the anomalous word to the overall context is good. Barton and Sanford (1993) manipulated the fit of anomalous noun phrases with context in their famous 'Air crash' scenario. 'Fit' was understood as statistical fit, rather than a strict match in terms of the semantic content of words, and anomaly detection rates were observed to fall as fit was improved. For example, following context material about an air crash on the border of France and Spain, participants answered the question 'Where should the survivors be buried?'. If anomaly detection was successful participants would have been expected to point out the error in burying people who had survived, but in this case participants answered correctly only 59% of the time. Barton and Sanford argued that the semantic representation was left underspecified due to the good fit of the word 'survivors' with the 'air crash' scenario; readers had apparently not included the 'is alive' feature in their representation of 'survivor'. There was further support for this interpretation from the observation that whenever 'air crash' was changed to 'bicycle crash', thus reducing the good fit of the word 'survivors', anomaly detection rates increased considerably.

This general finding has been replicated in some recent studies by Daneman and colleagues (e.g. Hannon and Daneman, 2004; Daneman, Lennertz and Hannon, 2007). Using 'incidental anomalies' modelled on those used by Barton and Sanford, they reported low detection rates under good fit conditions. For example, after a short passage about a girl who had drunk too many cups of coffee, readers were presented with the sentence, 'Amanda was bouncing all over because she'd had too many tranquillizing sedatives in one day', and asked what they thought Amanda should do. Rather than pointing out that 'tranquillizing sedatives' would not explain why Amanda was 'bouncing all over', participants frequently made some suggestion about drinking less coffee. The semantic relatedness of the impostor word *sedatives* to the 'correct' word *stimulants* had clearly had an effect on detection and allowed a further demonstration of one of Erickson and Mattson's (1981) Moses Illusion findings, namely that the greater the semantic relatedness between the correct word and the impostor word, the less likely people are to notice the anomaly. Daneman, Lennertz and Hannon (2004) also manipulated the semantic coherence of the anomalous NP and found that accuracy was poorest with internally incoherent anomalous NP, such as *tranquillizing stimulants*. The good fit of *stimulants* with the 'too much coffee' scenario resulted in an underspecified representation of the text in the same way that *survivors* had with the 'air crash' scenario (although it was only less-skilled readers who were observed to have particular difficulty with this type of anomalous NP).

Another recent study examining the effect of context on anomaly detection was an ERP study by Nieuwland and Van Berkum (2005). Participants listened to a short passage describing, for example, a man checking in at an airport check-in desk. After a number of references to the man, the passage began referring to 'the suitcase' instead, as though the suitcase were the animate protagonist that had hitherto been described in the passage. This of course entailed a serious breach of semantics, as inanimate objects do not conduct conversations etc. If participants successfully detected the anomaly, the predicted ERP output would have been the N400 effect, a negative-going waveform typically associated with semantic difficulties (e.g. Kutas and Hillyard, 1980; Van Berkum, Hagoort and Brown, 1999). Instead, the authors reported a P600, an effect normally – but controversially – associated with syntactic anomalies (e.g. Hagoort, Brown and Groothusen, 1993, Osterhout and Nicol, 1999; Osterhout, 1997; Osterhout, Holcomb and Swinney, 1994). They interpreted the ERP data without assigning a precise meaning to the P600, and instead focussed on the simple fact that the anomaly was not reflected immediately in the ERP output, but was apparently detected after some delay (indexed by the P600). The participants, they argued, had been subject to a temporary semantic illusion (‘change deafness’) and really had not noticed the switch from the animate ‘man’ character to the inanimate ‘suitcase’. So while context was not explicitly manipulated in this study, the good fit of ‘suitcase’ with the ‘airport check-in’ scenario was one factor which may have encouraged shallow processing of the anomalous NP, albeit briefly.

However, it is not al all clear that the N400/P600 distinction reflects the timing of anomaly detection effects, or the temporal relations of syntax-based versus semantics-based interpretation processes. Recent reviews of the ERP anomaly literature (Kuperberg, 2007; A. J. Sanford, Bohan, Molle and Leuthold, in preparation) have made attempts to identify exactly what the two wave forms reflect with regard to their appearance in anomaly settings, and it will be worth touching briefly on their conclusions in order to justify the current investigations more fully.

The main point of interest is that both waveforms can be elicited by anomalies regarded as semantic, and this observation therefore does away with the traditional semantics/syntax distinction. Instead, the factor influencing whether or not an anomaly will elicit an N400 or a P600 appears to be the fit of the anomalous word to its context: An anomalous word that does not fit well with it's context will elicit the N400, while an anomaly that does fit well will elicit the P600. Evidence for this distinction comes from the fact that the N400 has been observed even in non-anomalous settings in which a word merely fits in poorly with its sentential context (Kutas and Hillyard, 1984). It has also been shown that an acceptable fit with sentential context but a poor fit with discourse context elicits an enhanced N400, for example, *John is very* ***fast*** following a sentence in which John was running slower than usual (Van Berkum, Hagoort and Brown, 1999). Outside of sentential and discourse contexts, the N400 to single words has been shown to be modulated by close semantic links between other single words (Bentin et al., 1985; Rugg, 1985) and by the semantic properties of an unseen linking word (Chwila and Kolk, 2002; Chwila, Kolk and Mulder 2000; Kreher, Holcomb and Kuperberg, 2006). Crucially, the N400 is modulated by semantic associations between words even when a sentence is completely plausible (Van Petten, 1993).

Turning to the P600 – traditionally associated with syntactic anomalies – there is evidence that a P600 can be observed in cases of thematic role violations of verbs – almost certainly a semantic anomaly (Hoeks, Stowe and Doedens, 2004; Kim and Osterhout, 2005; Kolk and Chwila, 2007; Kuperberg, 2007). It is unlikely to be the case that such effects reflect syntactic reanalysis, as, for example, in Nieuwland and Van Berkum's materials (2005), syntactic reanalysis could not make sense of the anomaly. Instead, it seems likely that on those occasions in which a (non-syntactic) anomaly has elicited a P600, it can be seen that the anomalous word has indeed fit well with its context (Nieuwland and Van Berkum, 2005; Kim and Osterhout, 2005). As a further means of substituting the standard distinction between the two waveforms for one based on fit with global context, Kuperberg (2007) describes a study by Sitnikova, Holcomb and Kuperberg (unpublished) in which participants watched a man rubbing his face as if in need of a shave, and then either attempting to shave with a rolling pin, or rolling out dough with a rolling pin. In the case of the former, a P600 is observed, suggesting that the act of shaving, while anomalous insofar as it is attempted with a rolling pin, fits well with the overall context of needing a shave. In the case of the dough-rolling continuation, nothing is anomalous, but the action fits poorly with the established context of shaving, and an N400 is observed. While further bolstering this new distinction, this study suggest the N400/P600 distinction is not even language specific.

The relevance of this to the current work is that, while fit with context is certainly represented by distinct neural responses, the current state of the ERP literature cannot help us much with the question of whether or not contextual fit influences, i.e. suspends or prevents, online, syntax-based interpretation. While the results of the Nieuwland and Van Berkum study led the authors to allow for the possibility of semantics-led interpretation, and the Kim and Osterhout (2005) study prompted similar 'semantics-in-control' interpretations, the current understanding of the N400/P600 distinction suggests that these interpretations of the ERP data may not be satisfactory and may need to be revised in terms of the anomaly's fit with its context. Using eyetracking should provide a reliable means of investigating the online effects of context (i.e. semantics) on the building of an interpretation based on syntax, as it should be a straightforward case of observing whether the anomaly detection occurs later following a well-fitting context than following a neutrally-fitting context.

There is one piece of evidence suggesting that contextual fit may influence online processing. Daneman et al. (2006) reported that older readers who had successfully detected internally coherent anomalies in their 'air crash' type materials (e.g. *surviving injured*) had longer first pass times on encountering the anomalies, while the more difficult *surviving dead*, with its arguably better fit with context, was not detected immediately. However, this finding did not extend to younger readers, who never detected anomalies immediately. The authors were unsure as to why the older readers had been able to detect earlier, and suggested they were linguistically more sophisticated; but an analysis of look-back measure showed that the older readers took significantly longer to recover from the anomalies. The picture, then, is not consistent. Also, the analyses used in the Daneman studies only covered first pass and look-back measures on the anomalous phrase itself, and therefore could not tell us exactly when detection occurred if it did not occur immediately. (This narrow range of analysis likely reflected limitations of the small material set.) Bohan and Sanford (in press) eyetracked a larger set of similar materials and reported that while there was no evidence of detection at the critical anomaly itself, there was an increased number of regressions out of the post-critical region when the anomaly had been successfully detected, and this perhaps sheds light on when non-immediate detection may have occurred in the Daneman studies. (This study, however, was not concerned with explicitly manipulating the fit of the anomalous word to the context.)

A fuller understanding of the effect of a well-fitting context on processing could be gained using the kind of measure so far reported in this thesis, as well as by testing a larger material set (the Daneman studies tested only three). With regard to the time course of syntactic versus heuristic processing, the question is whether this ‘goodness-of-fit’ mechanism is active at the earliest stages of processing. Nieuwland and Van Berkum suggest that their data are indeed consistent with a semantics-first approach to interpretation where contextual fit information is applied early. If statistical associations influence interpretation at a very early stage, then in the type of sentence prone to being normalised we might expect to see a later anomaly detection following a well-fitting context than following a neutral one. The present studies will manipulate the fit between context and anomaly to explicitly test for the early operation of a goodness-of-fit heuristic

Experiment 7

Introduction

In this experiment both plausibility and contextual fit were manipulated in order to observe the time course of anomaly detection under differing context conditions. Consider the following passage:

It was almost nine o'clock.

The menu served the meal but the kitchen was in chaos.

The second sentence is clearly anomalous, as 'menu' (an inanimate object) would not be capable of serving a meal. Under most reading conditions we would expect a very early, perhaps immediate, detection of this obvious breach of semantics. But consider the same anomaly preceded by a different context:

The service was slow in the restaurant.

The menu served the meal but the kitchen was in chaos.

'Menu' now fits well with the 'restaurant' context, compared with the semantically neutral context of the previous passage which merely stated the time of day. Under this new condition, would the good fit of the word 'menu' with the context of being in a restaurant influence the timing of the anomaly detection, or possibly even prevent it?. If the 'goodness of fit' interpretive process is active at the earliest stages of processing, it is conceivable that an underspecified representation (based on word associations) could be in control of the comprehension process, rather than one based on algorithmic grammatical interpretation. If the associations between context and anomaly were particularly strong, they could delay the instantiation of an algorithmically-derived interpretation, and we would predict a later detection in the good context condition relative to the neutral context condition. In other words, in our example, we might see detection occurring at 'served' following the neutral context, but not see evidence of detection until ‘the meal’ following the good (‘restaurant’) context. If the ‘goodness of fit’ heuristic does not operate early, i.e. is post-syntactic, then we would not predict any difference in the time course of anomaly detection under the different context conditions.

This experiment will continue in the vein of using eyetracking as the most sensitive and reliable measure of anomaly detection. As the primary analysis concern in this experiment is the timing of online interpretive processes, we will dispense with the type of comprehension question so far used in this thesis, i.e. we will not use questions that probe readers’ interpretations of critical anomalies. This should allow for the most natural reading conditions, free of heavy secondary task demands, and give the best chances of observing any fast heuristic processes which may be ‘relegated’ under more taxing comprehension constraints.

Method

Participants

32 participants were recruited from the University of Glasgow student population and were paid for their participation. All were native English speakers, had normal or corrected-to-normal vision, and had not been diagnosed with any reading disorders. They were naïve as to the purpose of the experiment.

Materials

The material set comprised 32 two-sentence passages. An example material is presented in Table 27 in each of its four conditions.

The critical sentence was always the second and began by describing a simple transitive event that could be either semantically plausible or anomalous. The anomaly was created via an animacy violation. Following this critical clause there was some further material, approximately 5 to 6 words, intended as spillover regions for analysis purposes. This further material could be, for example, a prepositional or adverbial clause and did not contain anything essential to the overall understanding of the passage.

Each critical sentence was preceded by a context sentence which could be either related (‘Good’) or unrelated (‘Neutral’) to the content of the second sentence, specifically the subject NP of the critical clause, filling the agent role of the main verb. In the example given in table 27, the context sentence describing a situation in a restaurant obviously ‘fits’ better with the NPs *menu* and *waiter* than does the context sentence that simply states the time of day.

Table 27: Example material in all four conditions, with questions.

Condition	Example passage
Good Context Plausible	The service was slow in the restaurant. The waiter served the meal but the kitchen was in chaos. *How was the service?* *Slow < > Fast*
Good Context Anomalous	The service was slow in the restaurant. The menu served the meal but the kitchen was in chaos. *How was the service?* *Slow < > Fast*
Neutral Context Plausible	It was almost nine o'clock. The waiter served the meal but the kitchen was in chaos. *What time was it?* *Nine o'clock < > ten o'clock*
Neutral Context Anomalous	It was almost nine o'clock. The menu served the meal but the kitchen was in chaos. *What time was it?* *Nine o'clock < > ten o'clock*

Half of the experimental items were followed by comprehension questions. None of the questions probed the critical clause. Half of the questions focused on the content of the context sentence and half probed the material following the critical clause. In each of these question types, half of the correct answers were presented on the left and half were presented on the right. The question types and answer positions were therefore balanced against answering strategies not based on comprehension.

In a break with our earlier studies, the comprehension questions were not a crucial element in the design. (As mentioned in the introduction, they did not directly probe the critical clause.) The studies so far presented have already established significant rates of shallow processing and so replication is not necessary at this point. It was also felt that as the most important aspect of this study is the reading time data, fewer and less challenging comprehension questions might allow for more natural reading.

The experimental items were interspersed with 76 filler items. Twenty-four of the fillers were experimental items for experiment 8; the remaining 52 were short passages modelled after the experimental items for experiments 7 and 8. All filler materials, barring the items from experiment 8, were intended to be semantically plausible. Materials were divided across 4 experimental lists using a latin square design. Each list was viewed by a total of 8 participants, and each participant saw all 32 items in one of their four conditions. In each experimental list there was a total of 40 semantically implausible materials.

A norming study was carried out to confirm the suitability of the materials on both the plausibility of the critical clause and the fit of the context sentence. 24 participants rated 36 passages in a questionnaire study. An example ratings question and accompanying questions is given below:

The service was slow in the restaurant.

The menu served the meal but the kitchen was in chaos.

Makes no sense at all 1 2 3 4 5 6 7 Makes complete sense

How relevant is MENU to being in a restaurant?

Not relevant at all 1 2 3 4 5 6 7 Highly relevant

(The questions would be similar for the Neutral contexts, e.g. “How relevant is MENU to is being 9 o’clock in the evening?”.)

The criteria for inclusion in the final material set was a score between 1 and 3.5 for Implausible items and Neutral contexts, and between 4.5 and 7 for Plausible items and Good contexts. Three items were excluded from the final set on plausibility grounds and 1 item on contextual fit grounds. (Later reading time results confirmed the effectiveness of these criteria.)

The mean rating for Plausible items was 6.273 and the mean rating for Anomalous items was 1.744. Analysis of variance indicated that this difference was significant ($F_1(1,23) = 295.461$, MSe = 1.666, $p < 0.001$; $F_2(1,31) = 1636.939$, MSe = 0.200, $p < 0.001$). The mean rating for Good contexts was 6.516, and for Neutral contexts it was 2.353. This difference was significant ($F_1(1,23) = 293.266$, MSe = 0.709, $p < 0.001$; $F_2(1,31) = 620.000$, MSe = 0.448, $p < 0.001$).

Apparatus

The apparatus was identical to that used in experiments 5 and 6.

Procedure

The procedure was identical to that carried out in experiments 5 and 6.

Analysis

The materials were divided into regions for analysis, as indicated below.

$_1$There was a new play on in the theatre.|

$_2$The spotlight|$_{\mathbf{3}}$ **recited**|$_4$ the speech|$_5$ that|$_6$ began|$_7$ the first act.|

The critical region was region 3, the verb of the opening clause in the second sentence, as this was the earliest point at which the anomaly could be processed and detected. The rest of the analysis (reading time measures, etc.) was identical to that outlined for experiments 5 and 6.

Results

Reading time results are reported by region. To preview, there is evidence of anomaly detection at the earliest point. Type of context did not affect the time course of anomaly detection, and effects of context are limited to (a) the magnitude of the initial disruption and (b) the time course of recovery from the disruption caused by the anomaly. By the second spillover region there is evidence that the disruption caused by the Good Context anomaly is beginning to die down, ahead of that caused by the Neutral Context anomaly.

Reading Time Results

For ease of reference, an example material is presented below (in the Good Context/Anomalous condition), divided into its analysis regions.

Good Context / Anomalous

$_1$There was a new play on in the theatre.|

$_2$The spotlight|$_{\mathbf{3}}$ **recited**|$_4$ the speech|$_5$ that|$_6$ began|$_7$ the first act.|

For each analysis, two ANOVAs were conducted: One by participants (F_1) and one by materials (or items; F_2). Mean reading time measures are presented in Table 28, below. One item removed from analysis due to a typographical error.

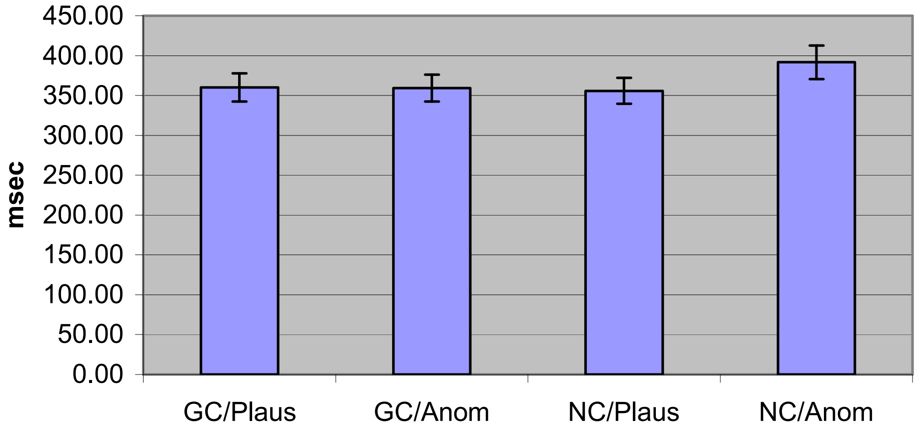
Table 28: Mean reading time measures for regions 3-7

	Region					
	2 (pre-critical)	**3 (critical)**	**4**	**5**	**6**	**7**
Measure	*The spotlight* Mean (SD)	*recited*	*the speech*	*that*	*began*	*the first act*
First Fixation (msec)						
Good/Plausible	254 (47)	252 (51)	254 (44)	242 (37)	241 (36)	258 (51)
Good/Implausible	264 (52)	265 (51)	267 (38)	231 (42)	242 (41)	278 (52)
Neutral/Plausible	265 (41)	258 (47)	255 (39)	231 (36)	243 (37)	276 (58)
Neutral/Implausible	268 (47)	266 (49)	267 (40)	250 (52)	238 (32)	273 (57)
First Pass (msec)						
Good/Plausible	360 (99)	295 (68)	347 (80)	371 (75)	309 (71)	425 (149)
Good/Implausible	359 (95)	327 (85	376 (78)	347 (57)	315 (67)	473 (155)
Neutral/Plausible	356 (92)	310 (74)	348 (72	242 (43)	312 (70)	459 (148)
Neutral/Implausible	392 (118)	363 (110)	393 (74)	264 (73	308 (67)	475 (165)
First Pass Reg. Out						
Good/Plausible	0.81 (3.21)	10.94 (21.3)	16.44 (15.38)	7.94 (13.01)	9.38 (13.78)	33.18 (29.47
Good/Implausible	1.75 (6.90)	13.47 (13.47)	32.28 (22.41)	14.78 (19.46)	14.09 (19.26)	29.94 (26.30)
Neutral/Plausible	0.94 (3.70)	9.50 (15.70)	13.41 (13.73)	6.97 (12)	9.69 (19.25)	30.81 (29.16)
Neutral/Implausible	2.09 (5.88)	16.59 (13.1)	27.88 (21.85)	14.78 (17.84)	11.13 (12.54)	35.63 (24.42)
Regression Path (msec)						
Good/Plausible	371 (125)	338 (96)	435 (113)	296 (87)	343 (83)	628 (319)
Good/Implausible	381 (108)	444 (188)	633 (243)	323 (112)	393 (120)	673 (315)
Neutral/Plausible	379 (120)	364 (141)	427 (139)	281 (66)	369 (119)	673 (345)
Neutral/Implausible	406 (130)	446 (157)	616 (202)	357 (157)	382 (130)	726 (324)
Total Time (msec)						
Good/Plausible	392 (145)	362 (117)	417 (111)	284 (73)	363 (90)	489 (193)
Good/Implausible	490 (163)	487 (161)	522 (157)	295 (102)	373 (95)	515 (184)
Neutral/Plausible	392 (122)	376 (133)	418 (111)	273 (68)	377 (98)	511 (169)
Neutral/Implausible	554 (186)	503 (147)	523 (125)	290 (95)	372 (97	532 (185)

Pre-critical region (region 2)

In first fixation there was no main effect of context ($F_1(1,31) = 1.335$, MSe = 1320.421, $p > 0.2$; $F_2(1, 30) = 2.723$, MSe = 615.753, $p > 0.1$), no effect of plausibility ($F_1(1,31) = 1.242$, MSe = 1021.256, $p > 0.2$; $F_2(1, 30) = 1.025$, MSe = 1283.941, $p > 0.3$), and no significant interaction (both *F*s < 1). In first pass analysis there was no significant effect of context ($F_1(1,31) = 2.160$, MSe = 2949.334, $p > 0.1$; $F_2(1, 30) = 1.558$, MSe = 4525.185, $p > 0.2$) or plausibility ($F_1(1,31) = 2.625$, MSe = 3739.427, $p > 0.1$; $F_2(1, 30) = 2.122$, MSe = 3716.856, $p > 0.1$). However, there was an interaction, significant only by items ($F_1(1,31) = 2.727$, MSe = 3928.449, $p > 0.1$; $F_2(1, 30) = 4.902$, MSe = 3521.324, $p < 0.05$) (see Figure 24) Comparisons of the two Good Context conditions indicated there was no difference in reading times (both *t*s < 1), while a comparison of the Neutral conditions revealed a difference that was marginal by participants and significant by items ($t_1(31) = 1.894$, $p = 0.068$; $t_2(30) = 2.712$, $p < 0.05$).

Figure 24:
First Pass Reading Time (msec), Parafoveal Region



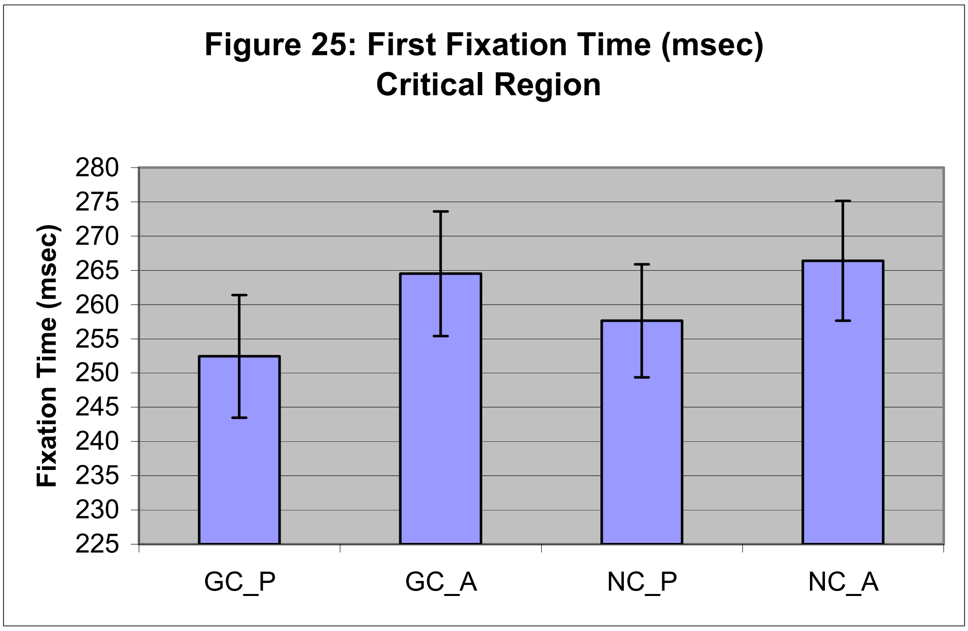
This effect therefore might indicate that, under Good Context conditions, an early fit-with-context heuristic is operative, delaying the detection of an anomalous phrase relative to the same phrase in a semantically neutral context.

The effect is not robust, however, so a firm conclusion along these lines is not warranted. We will return to this effect in chapter 6, where an analysis of this region by Reading Ability will shed more light on its origin. For now, suffice to say that the effect may be representative only of strategies used by less-skilled readers.

In first pass regressions out there was no significant effect of context (both *F*s < 1), no effect of plausibility ($F_1(1,31) = 1.642$, MSe = 21.361, $p > 0.2$; $F_2(1, 30) = 1.949$, MSe = 14.899, $p > 0.1$), and no interaction (both *F*s < 1). In first pass regressions out there was no significant effect of context (both *F*s < 1), no effect of plausibility ($F_1(1,31) = 1.642$, MSe = 21.361, $p > 0.2$; $F_2(1, 30) = 1.949$, MSe = 14.899, $p > 0.1$) and no interaction (both *F*s < 1). Likewise in regression path analysis, there was no effect of context ($F_1(1,31) = 1.605$, MSe = 5426.808, $p > 0.2$; $F_2(1, 30) = 1.128$, MSe = 6632.532, $p > 0.2$), no effect of plausibility ($F_1(1,31) = 1.651$, MSe = 6746.841, $p > 0.2$; $F_2(1, 30) = 1.665$, MSe = 5451.742, $p > 0.2$), and no interaction ($F_1 < 1$; $F_2(1, 30) = 1.640$, MSe = 5086.924, $p > 0.2$). Lastly, in total time there was a marginal effect of context, with longer reading times in the Neutral context conditions ($F_1(1,31) = 3.055$, MSe = 11011.292, $p = 0.09$; $F_2(1, 30) = 3.284$, MSe = 8680.092, $p = 0.08$). There was also a significant effect of plausibility, with longer total reading times in the anomalous conditions ($F_1(1,31) = 62.228$, MSe = 8462.885, $p < 0.001$; $F_2(1, 30) = 35.634$, MSe = 13931.923, $p < 0.001$). These effects were modulated by an interaction, significant by items only ($F_1(1,31) = 2.538$, MSe = 12473.554, $p > 0.1$; $F_2(1, 30) = 5.689$, MSe = 6527.878, $p < 0.05$).

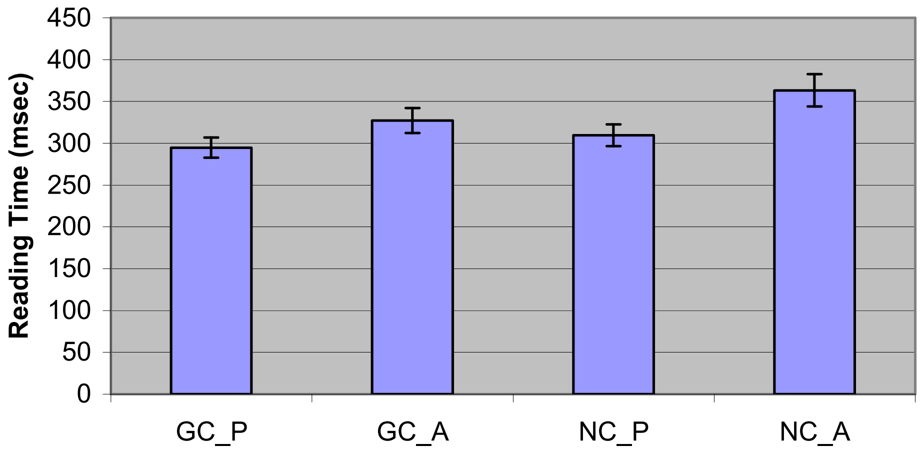
Region 3 (critical region)

In first fixation there was a main effect of Plausibility, significant by participants, with longer initial looking times in the Anomalous conditions ($F_1(1,31) = 4.300$, Mse = 808.276, $p < 0.05$; $F_2(1,30) = 3.067$, Mse = 1295.592, $p = 0.09$) (see Figure 25).



There was no significant effect of context and no significant Context*Plausibility interaction (all Fs < 1). First pass also showed a significant main effect of plausibility, with longer reading times in the Anomalous conditions ($F_1(1,31) =$ 13.254, Mse = 4493.472, $p < 0.005$; $F_2(1,30) = 18.602$, Mse = 2870.133, p <0.001). This analysis also showed an effect of Context, significant by participants only, with longer reading times in the Neutral Context conditions ($F_1(1,31) = 4.324$, Mse = 4794.409, $p < 0.05$; $F_2(1,30) = 2.502$, Mse = 4884.703, p >0.1). There was no significant interaction (F_1 <1; $F_2(1, 30) = 2.316$, Mse = 3980.031, $p > 0.1$) (see Figure 26).

Figure 26: First Pass Reading Time (msec)
Critical Region



Analysis of first pass regressions out showed a main effect of plausibility, marginal by participants and significant by items, with more regressions in the Anomalous conditions (F_1(1,31) = 3.762, Mse = 196.980, p = 0.062; F_2(1,30) = 7.300, Mse = 195.813, p<0.02). Neither the effect of context nor the interaction were significant (all Fs < 1). Regression path time showed a significant effect of plausibility (F_1(1,31) = 35.793, Mse = 7876.133, p<0.001; F_2(1,30) = 29.66, Mse = 9311.345, p<0.001), with longer regression path times in the anomalous conditions. Again, there were no other significant effects (all Fs < 1). Total time analysis showed the same plausibility main effect (F_1(1,31) = 53.994, Mse = 9374.097, p<0.001; F_2(1,30) = 51.086, Mse = 10040.099, p<0.001), and no other significant effects (all Fs < 1).

Both first fixation and first pass analyses indicate that the anomaly was detected at the earliest point. While there was no significant interaction, a numerical difference in the means of the two implausible conditions suggests that the disruption in comprehension caused by the anomaly was slightly greater following a neutral context (first pass: 32 msec in the Good Context Implausible condition vs. 53 msec in the Neutral Context Implausible condition).

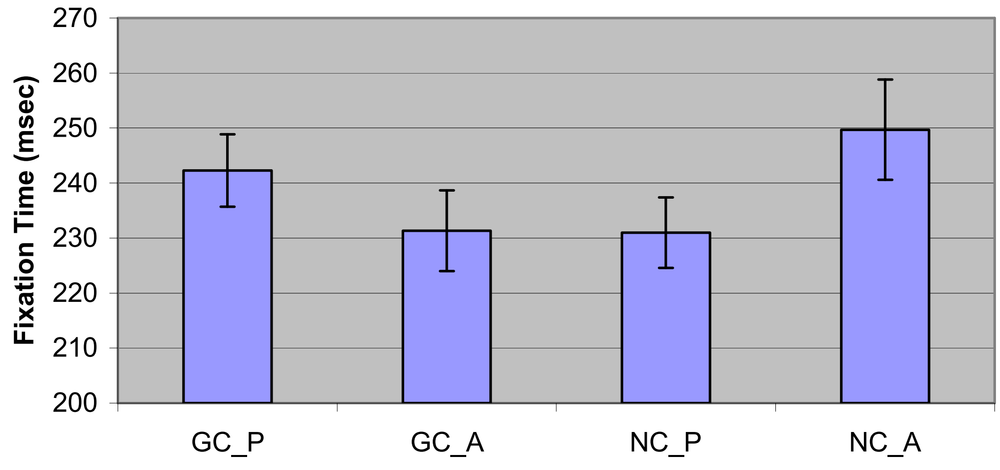
Region 4 (first spillover region)

First fixation showed only a plausibility main effect, with longer initial fixation times in the anomalous conditions ($F_1(1,31) = 6.965$, Mse = 672.007, $p < 0.02$; $F_2(1,30) = 5.673$, Mse = 706.566, $p < 0.03$; all other Fs < 1). First pass showed exactly the same pattern of results: a main effect of plausibility (($F_1(1,31) = 17.833$, Mse = 2423.435, $p < 0.001$; $F_2(1,30) = 15.262$, Mse = 2739.631, $p < 0.001$; all other Fs < 1). First pass regressions out showed the now typical plausibility effect ($F_1(1,31) = 30.105$, Mse = 244.168, $p<0.001$; $F_2(1,30) = 23.032$, Mse = 321.356, $p<0.001$). There was no significant effect of context ($F_1(1,31) = 1.663$, Mse = 266.080, $p > 0.2$; $F_2(1,30) = 3.367$, , Mse = 126.713, $p = 0.076$), though the items analysis suggests marginally more regressions in the Good Context conditions (32.19 in the Good Context/Anomalous condition vs. 27.42 in the Neutral Context/Anomalous condition). The interaction was not significant (both Fs < 1). Regression Path time showed the same significant plausibility effect ($F_1(1,31) = 35.152$, Mse = 34134.455, $p < 0.001$; $F_2(1,30) =$ 53.471, Mse = 21699.982, $p < 0.001$), and no other significant effects (all other Fs < 1). Total time likewise showed significantly more time spent reading this region in the Anomalous conditions, and no other significant effects ($F_1(1,31) = 31.892$, Mse = 11016.386, $p < 0.001$; $F_2(1,30) = 32.067$, Mse = 10998.173, $p < 0.001$; all other Fs < 1). This region thus contains effects driven by the spillover in processing difficulty caused by the anomalies.

Region 5 (second spillover region)

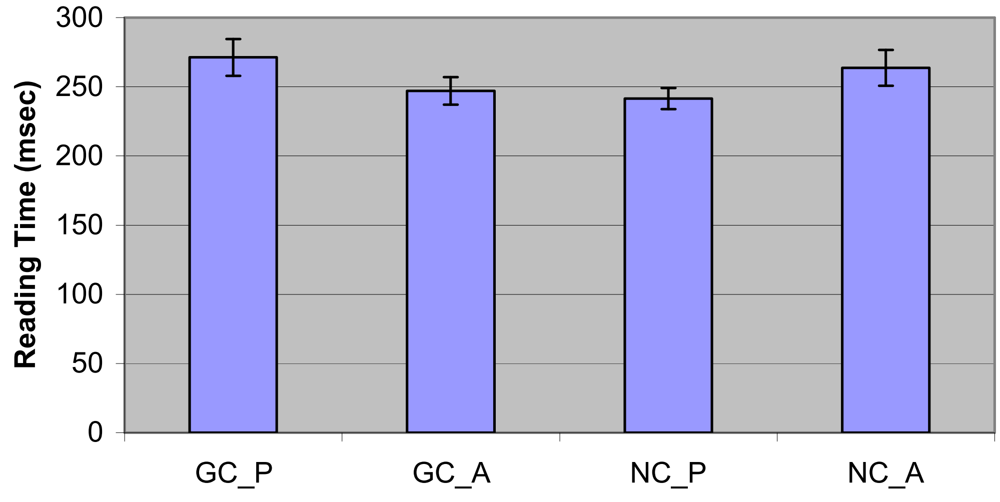
First fixation analysis revealed there were no significant main effects of either context or plausibility (all *F*s < 1). However, there was a significant context*plausibility interaction ($F_1(1,31) = 11.969$, Mse = 589.104, $p < 0.005$; $F_2(1,30) = 18.247$, Mse = 591.640, $p < 0.001$). Planned comparisons indicated that the difference between the two Good Context conditions was significant only by items ($t_1(31) = 1.329$, $p > 0.1$; $t_2(30) = 2.233$, $p < 0.05$), with longer fixation times in the Plausible condition, while the difference between the Neutral Context conditions was significant by both participants and items ($t_1(31) = 2.572$, $p < 0.05$, $t_2(30) = 3.607$, $p < 0.005$), with significantly longer fixation times in the Anomalous condition. Thus it would appear that by the second spillover region, in terms of initial looking time at least, the difficulty associated with the Good context anomaly is beginning to settle down while the difficulty with the Neutral context anomaly is still quite robust (see Figure 27).

Figure 27: First Fixation Time (msec)
Region 5



This impression is also borne out in first pass analysis. There were no significant effects of context ($F_1 < 1$; F_2 (1,30) = 1.075, Mse = 1117.347, $p > 0.3$) or plausibility (both *F*s <1), but there was a significant interaction (F_1(1,31) = 7.814, Mse = 2210.693, $p < 0.05$; F_2(1,30) = 11.785, Mse = 1547.931, $p < 0.005$). T-tests comparing the Good context conditions showed no difference by participants and a marginal difference by items, with longer reading times in the Plausible condition (t_1(31) = 1.598, $p > 0.1$; t_2(30) = 1.966, $p = 0.059$), while comparisons of the Neutral context conditions showed a marginal difference by participants and a significant difference by items, with longer times in the Anomalous condition (t_1(31) = 1.991, $p = 0.055$; t_2(30) = 3.446, $p < 0.005$) (see Figure 28).

Figure 28: First Pass Reading Time (msec)
Region 5



First pass regressions out show a main effect of plausibility with significantly more regressions in the Anomalous conditions (F_1(1,31) = 11.305, Mse =

152.010, $p < 0.005$; $F_2(1,30) = 9.923$, Mse = 161.661, $p < 0.005$). Neither the effect of context nor the interaction were significant (all Fs < 1). Regression path analysis showed the same plausibility main effect ($F_1(1,31) = 7.533$, Mse = 11349.64, $p < 0.05$; $F_2(1,30) = 12.167$, Mse = 8927.728, $p < 0.001$). There was no significant effect of context ($F_1 < 1$; $F_2(1,30) = 1.274$, Mse = 11467.780, $p >$ 0.2) and no significant interaction ($F_1(1,31) = 2.493$, Mse = 7784.581, $p > 0.1$; $F_2(1,30) = 1.538$, Mse = 1131.680, $p > 0.2$). In total time analysis there was no significant effect of context (both Fs < 1), or plausibility ($F_1(1,31) = 1.077$, Mse = 5632.322, $p > 0.3$; $F_2(1,30) = 1.970$, Mse = 3859.223, $p > 0.1$), and no significant interaction (both Fs < 1).

So while two early processing measures – first fixation and first pass – give the impression that the comprehension disruption is dying down for the Good Context anomaly compared to the Neutral Context anomaly, both first pass regressions out and regression path analyses indicate that both anomalies are continuing to cause difficulty and are prompting rereading of earlier material in order to attempt a resolution.

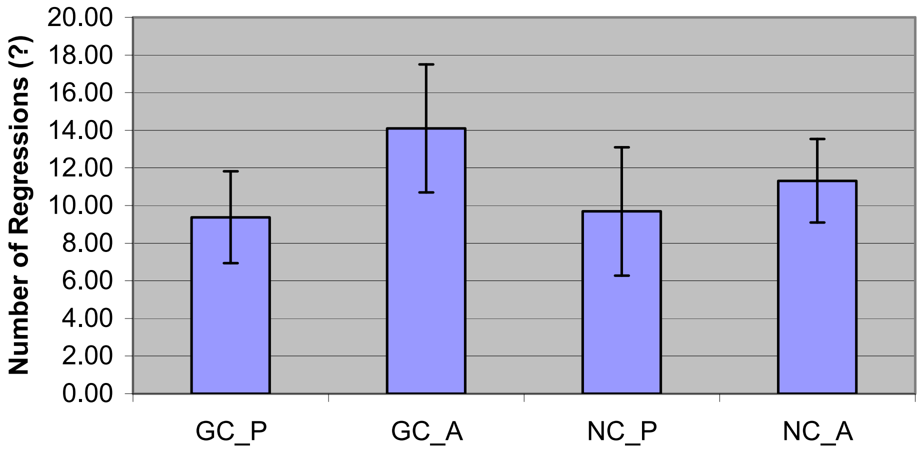
Region 6 (third spillover region)

First fixation analysis revealed no main effects of either context or plausibility (all Fs < 1) and no interaction (F_1 <1; $F_2(1, 30) = 1.015$, Mse = 753.566, $p > 0.3$). First pass likewise showed no significant effects at all (all Fs < 1). With first pass regressions out there was no significant effect of context (F_1<1; $F_2(1,30) = 1.471$, MSe = 92.567, $p > 0.2$), but there was a main effect of plausibility, significant only by items, suggesting a greater number of regressions in the Anomalous conditions ($F_1(1,31) = 1.548$, Mse = 207.945, $p > 0.2$; $F_2(1,30) = 4.823$, Mse =

162.766, $p < 0.05$). The interaction was not significant (both Fs < 1) (see Figure 29). Regression path analysis showed a similar weak effect of plausibility ($F_1(1,31) = 2.871$, Mse = 10993.555, $p = 0.1$; $F_2(1,30) = 4.895$, Mse = 9792.759, $p < 0.05$), no significant effect of context (both Fs < 1) and no interaction ($F_1<1$; $F_2(1,30) = 1.174$, Mse = 4940.09, $p > 0.2$). Total time analysis for this region showed no significant effects at all (all Fs < 1).

So by region six it seems there is an overall dying-down of the disruption caused by the anomalies, with residual disruption evident only as weak effects in the regression behaviour analyses.

Figure 29:
First Pass Regressions Out Region 6



Region 7 (wrap up)

In first fixation analysis neither the effect of context ($F_1(1,31) = 1.267$, Mse = 986.645, $p > 0.2$; $F_2(1,30) = 1.615$, Mse = 967.024, $p > 0.2$) nor the effect of plausibility ($F_1(1,31) = 2.179$, Mse = 1214.443, $p > 0.1$; $F_2(1,30) = 1.827$, Mse = 1184.370, $p > 0.1$) were significant. There was a marginally significant interaction ($F_1(1,31) = 3.822$, Mse = 1065.628, $p = 0.06$; $F_2(1,30) = 4.094$, Mse = 1504.624, $p = 0.052$), most likely indicative of the fact that the Good Context/Plausible condition contained the most easily comprehensible passages overall. First pass revealed no significant effect of context ($F_1(1,31) = 1.475$, Mse = 6539.149, $p > 0.2$; $F_2(1,30) = 1.344$, Mse = 7182.720, $p > 0.2$). But there was a main effect of plausibility, marginal by participants and significant by items, indicating longer reading times in the anomalous conditions ($F_1(1,31) = 4.031$, Mse = 8088.760, $p = 0.053$; $F_2(1,30) = 7.039$, Mse = 5139.456, $p < 0.02$). The interaction was not significant ($F_1(1,31) = 1.372$, Mse = 5934.764, $p > 0.2$; $F_2(1,30) = 2.906$, Mse = 5423.106, $p > 0.05$). First pass regressions out analysis showed no significant effects (context and plausibility main effects: all Fs < 1; interaction: ($F_1(1,31) = 2.102$, Mse = 287.184, $p > 0.1$; $F_2(1,30) = 2.022$, Mse = 378.4, $p > 0.1$). Regression path analysis showed a main effect of context, marginal by participants and non-significant by items suggesting longer regression path times in the Neutral Context conditions ($F_1(1,31) = 3.758$, Mse = 20509.760, $p = 0.062$; $F_2(1,30) = 1.623$, Mse = 45484.678, $p > 0.2$). There was also a main effect of plausibility, significant by participants and marginal by items, with longer regression path times in the Anomalous conditions ($F_1(1,31) = 5.141$, Mse = 14765.062, $p < 0.05$; $F_2(1,30) = 3.732$, Mse = 21631.982, $p = 0.063$). The interaction was not significant (both Fs<1). Finally, with total time, there was no significant effect of context ($F_1(1,31) = 1.487$, Mse = 8208.999, $p > 0.3$; $F_2(1,30) = 1.028$, Mse = 10155.29, $p > 0.3$). There was an effect of plausibility, significant only by items, with longer total times in the Anomalous conditions ($F_1(1,31) = 2.747$, Mse = 6622.491, $p > 0.1$; $F_2(1,30) = 3.784$, Mse = 6348.923, $p = 0.061$). There was no significant interaction (both Fs < 1).

All effects in this region are interpretable in terms of the overall 'comprehensibility' of the passages in each condition, e.g. the interaction in first fixation and the (non-significant) trends in the means for first pass and total time which showed shortest reading times in the Good Context/Plausible condition, and in regression path time which also showed numerically longest times in the Neutral Context/Anomalous condition.

Summary and discussion of eyetracking results

The detection of the anomaly is immediate in both Implausible conditions, as evidenced by the lengthened first fixation times at the critical region, and the more robust effect in first pass reading time in the same region. In terms of the most reliable effects, there is therefore no evidence that the semantic fit of the anomalous word to its context affects the time course of anomaly detection: a good fit did not slow detection relative to a neutral fit. The marginal effect in the pre-critical region will require further analysis before it can be interpreted, and, to look ahead to chapter 6, it does suggest that less-skilled readers may use a goodness-of-fit heuristic in the earliest stages of interpretation (although it is short-lived and they do detect anomalies online in both context conditions). Otherwise, early effects of the context manipulation are limited to the significant context effect (participants only) in the first pass analysis of the critical region. In this region there were longer overall reading times in the Neutral conditions, but there was also a numerical difference (not statistically significant) in the magnitude of the disruption caused by the anomaly: this disruption was numerically greater following a semantically neutral context. However, there was no significant interaction and so no firm conclusions are warranted. By the second spillover region the disruption was still influencing regression and rereading behaviour in both anomaly conditions, but was beginning to die down in measures of initial fixation and early reading time in the Good Context conditions. By the third spillover region, the disruption had settled down in first fixation and first pass for both Implausible conditions, but weak regression effects in these conditions showed some lingering general disruption.

A possible explanation for the recovery benefit in the Good Context/ Anomalous condition is the suggestion in Region 4 that readers were making more regressions in the Good Context conditions (F_2 marginal only). This increased rate of regressions at this stage may have aided the slightly earlier recovery in the Good Context/Anomalous condition relative to the Neutral Context/Anomalous condition. But this effect is only marginal in the items analysis, and so is by no means conclusive.

General Discussion

In relation to our hypothesis, the picture that emerges from experiment 7 seems clear. The question was whether context could delay the detection of an anomaly if the words making up the anomaly constituted a good ‘fit’ with the situation described in the context. Recall that shallow processing was held to occur in the Barton and Sanford (1993) study because of the good fit of the anomaly (‘survivors’) with the overall scenario (an air crash). If anomaly detection could be delayed following a well-fitting context, but not a neutral one, then it would be evidence for the early operation of semantic interpretation processes, ahead of syntactic interpretation. This evidence would in turn bolster a ‘semantics first’ account in the issue of when semantic interpretation occurs relative to syntactic interpretation.

The data indicate that anomaly detection was immediate following both well-fitting and neutral contexts. These results are therefore in line with earlier studies in this thesis that show immediate or very early anomaly detection, and argue, most likely, for a syntax-first account of interpretation. There is, however, a challenge to this picture in the effect – marginal overall but significant in the data for the less-skilled readers – suggesting that, in the pre-critical region, parafoveal anomaly detection had occurred in the Neutral context conditions but not the Good context conditions. This certainly would be consistent with a heuristics-first account and is therefore extremely interesting. It would also suggest that, in some cases at least, online use of heuristics is modulated by individual difference factors, and would perhaps be in line with Daneman et al. (2006) who reported delayed detection for older readers only, when fit with context was good (see chapters 6 and 7 for further details and discussion).

To return to the robust effects reported here, an alternative interpretation of the immediate anomaly effects, which needn't necessarily rely on syntactic processing, is the N-V-N heuristic. This putative heuristic would assign the thematic roles *agent-verb-theme* to a string such as *the menu served the meal*, and could, at 'meal', entail the generation a semantic error signal. However, given that the only evidence we have seen so far for the operation of the N-V-N strategy has been under conditions of heavy syntactic load – very unlike the materials in the current study – the operation of the N-V-N heuristic is not a compelling interpretation in this case.

Interpretive processes that judge statistical relations between words, or goodness-of-fit, are therefore likely to be post-syntactic and operate by overruling interpretations generated from the output of the parser. But, as just indicated, it may be the case that this does not hold across the spectrum of reading comprehension ability, and we leave open the possibility that less-skilled readers rely, in the very early stages of processing, on 'fit' heuristics, even though they very rapidly bring syntactic information to bear on interpretation. The study by Barton and Sanford, and others by Daneman and colleagues, demonstrate the power of this semantic heuristic with regard to interpretations, but it would seem that when forming an initial interpretation, at the earliest point we can observe, syntax is most reliably observed to be the primary information source drawn on by the comprehension system. (One further point we can draw from our very early anomaly detection effect, is that we have further evidence that anomaly detection effects that appear in first pass measures are not limited to violations of a syntactic nature.)

Because the studies presented here track anomaly detection along the processing stream, we were able, as a secondary concern, to investigate any differential effects of our two contexts after the point at which the anomalies are detected.

Type of context did have an effect at the critical region, in that this region was read for longer in the Neutral context conditions regardless of plausibility. The Good context probably allowed for a faster integration of the beginning of the critical sentence, although any advantage did not impinge on its interpretation. There also appeared to be a difference in the magnitude of the initial disruption caused by the anomaly, with a numerically greater difference between the two Neutral context conditions than between the two Good context conditions. As there was no significant interaction, no firm conclusions can be drawn here. The only other effect to note was a benefit in recovery from the anomaly following a Good context. By the third spillover region there was still disruption evident in both implausible conditions, although disruption in the early fixation/reading measures has died down. However, in the Good Context/Implausible condition, this dying down had begun earlier, in the second spillover region.

It isn't immediately obvious why a well-fitting context should aid recovery from such a severe anomaly. After all, the animacy violation is rigid and technically no resolution is possible. There is, however, a literature detailing the effects on anomaly disruption when the anomaly is preceded by a highly supportive context, often of a fictionalised, 'cartoon' nature (Nieuwland and Van Berkum (2006) progressively eliminated an N400 effect elicited by a pragmatic anomaly using a strongly supportive discourse context. Even immediate anomaly effects can be eliminated through the use of a sufficiently detailed back-story. This phenomenon cannot explain the current findings, as nothing in the context attributes animacy to the inanimate NPs, but the lessening of processing disruption we see here may represent the comprehension system attempting to respond to the well-fitting context by constructing a coherent semantic representation. However, any such representation would still be of a highly heuristic nature, receiving virtually no support from the text itself (see experiment 8 General Discussion for another possible explanation).

In conclusion then, the data presented here suggest that, in the case of animacy violations at least, any goodness-of-fit heuristic probably comes into play *after* an interpretation has been generated according to syntactic rules (though analysis of the timing of detection for skilled and less-skilled readers suggest that this may not hold across reading ability). In our analysis here, collapsing over reading ability, any effects of context type are limited to (perhaps) the magnitude of the initial anomaly-related disruption, and the time course of the recovery from this disruption.

Experiment 8

Introduction

The present study is a further attempt to observe any early operation of the goodness-of-fit heuristic, an interpretation process that seems able to account for a number of cases of shallow processing. Experiment 7 failed to provide any evidence of this heuristic operating at a pre-syntactic phase. Following both well-fitting and neutral contexts, anomaly detection was observed at the earliest possible point in the processing record.

The anomalies used in experiment 7 were robust animacy violations. Given the somewhat severe nature of an animacy violation, it is conceivable that a heuristic based on statistical relations between words may have been ‘trumped’ by a simple animacy check that immediately recognised the implausibility. There were also no other semantic cues to suggest an alternative interpretation to the syntactic one: menu’s do not serve meals, and neither do meals serve menus.

Although Nieuwland and van Berkum (2005) reported delayed anomaly detection following an animacy violation, the present investigation differs in it's methodology (Nieuwland and Van Berkum reported ERP data) and for completeness' sake experiment 8 will test anomalies of a more subtle, and difficult, nature.

Consider the following:

It was the middle of the night.
The policeman was chased by the burglar down the dark empty street.

The second sentence is reminiscent of the implausible items used in experiments 1-3: it is not impossible that a burglar should chase a policeman, but highly unlikely. Rather, it's much more likely that a policeman should chase a burglar, and misinterpretation ensues when a reader or listener interprets on the basis of this pragmatic cue rather than according to the specifications of the syntax. The implausible event is also stated in the passive voice, a factor which the work of Ferreira (2003), and experiments 2, 3 and 4 in this thesis, have demonstrated to enhance the probability of misinterpretation. In the above example, then, could we expect to see an early detection relative to the same anomaly when preceded by a better-fitting context such as:

The robbery had gone wrong.
The policeman was chased by the burglar down the dark empty street.

‘Policeman’ now fits well with the robbery context and this might lure a reader into a shallow processing of the implausible verb phrase, for a short time at least. Experiment 8 will proceed in a similar manner to experiment 7, and test for evidence of immediate effects on interpretation of the Goodness-of-fit heuristic.

Now that it has been established several times that implausible sentences are misinterpreted with a surprisingly high frequency, experiment 8 will dispense with plausible control conditions and simply compare the reading of implausible sentences under Good and Neutral context conditions.

Method

Participants

32 participants were recruited from the University of Glasgow student population and were paid for their participation. All were native English speakers, had normal or corrected-to-normal vision, and had not been diagnosed with any reading disorders. They were naïve as to the purpose of the experiment.

Materials

Twenty-four materials were constructed, similar to those used in experiment 7. Table 29 gives an example material in its two conditions.

Table 29: Example material in both conditions, with questions.

Condition	Example Passage
Good Context	The robbery had gone wrong. The policeman was chased by the burglar down the dark empty street. What had gone wrong? Robbery < > burglary
Neutral Context	It was the middle of the night. The policeman was chased by the burglar down the dark empty street. What time was it? Middle of the night < > middle of the morning

The critical sentence was again the second one, which began by describing a simple transitive event that could be either semantically plausible or anomalous. The anomaly was a pragmatic anomaly, created by reversing the most plausible order of the initial verb's two arguments, for example, *the* **policeman** *was chased by the* **burglar**, rather than the more plausible, *the* **burglar** *was chased by the* **policeman**. The critical clause was this time phrased in the passive voice, to maximise the chances of it's being processed at a shallow level. Following this critical clause there was some further material, approximately 5 to 6 words, intended as spillover regions for analysis purposes. This further material could be, for example, a prepositional or adverbial clause and did not contain anything essential to the overall understanding of the passage.

A total of 28 materials were normed on a 7-point scale for plausibility and contextual fit (N = 24). An example item, with questions, is given below:

The robbery had gone wrong.

The policeman was chased by the burglar down the dark empty street.

How relevant is POLICEMAN to a robbery?

Not relevant at all 1 2 3 4 5 6 7 Highly relevant

How likely is it that a burglar would chase a policeman?

Not likely at all 1 2 3 4 5 6 7 Highly likely

The criteria for inclusion in the final material set was a score of 1 – 3.5 for Neutral Contexts and Plausibility, and a score of 4.5 – 7 for Good Contexts. All items met the criteria for Plausibility but 4 items were excluded on contextual grounds. The mean rating for Neutral contexts was 2.086 and the mean rating for Good contexts was 6.173. Analysis of variance indicated that the difference between means was significant (F_1(1, 23) = 520.036, MSe 0.385, p < 0.001; F_2 (1,23) = 482.501, MSe = 0.489, p < 0.001). The mean plausibility rating, collapsed across the Context variable, was 1.701. An ANOVA comparing plausibility ratings across Context conditions yielded no significant differences (both *F*s < 1). Type of context was therefore not a significant influence on participants' plausibility ratings.

The 24 experimental items were interspersed with 84 filler items. Thirty-two of the fillers were experimental items for experiment 7; the remaining 52 were short passages modelled after the experimental items for experiments 7 and

8. All filler materials, barring the Implausible conditions from experiment 7, were intended to be semantically plausible. In each experimental list there was a total of 40 semantically implausible materials. Half of the filler materials were followed by a comprehension question.

Materials were divided across 4 experimental lists using a latin square design. As experiment 8 had only two conditions, their repetition across lists 2 and 4 resulted in twice as many participants seeing each condition as in experiment 7 (this situation had the benefit of increasing the statistical power for experiment 8). Each list was viewed by a total of 8 participants, and each participant saw all 24 items in one of their two conditions.

Half of the items were followed by a comprehension question. The questions did not probe the critical clause: half probed the content of the context sentence, half probed the content of the material following the critical clause. Half of the questions had the correct answer presented on the, and half had the correct answer presented on the right. The question types and answer positions were therefore balanced against answering strategies not based on comprehension.

Apparatus

The apparatus was identical to that used in experiments 5 and 6.

Procedure

The procedure was identical to that carried out in experiments 5 and 6.

Analysis

The materials were divided into regions for analysis, as indicated below.

$_1$The army had already begun attacking the city.|
$_2$The soldier|$_3$ was protected by|$_4$ **the child**|$_5$ during|$_6$ all the|$_7$ heavy|$_8$ shooting.|

The critical region was region 4, the second NP of the opening clause in the second sentence, as this was the earliest point at which the anomaly could be processed and detected. The rest of the analysis (reading time measures, etc.) was identical to that outlined for experiments 5 and 6.

Results

Reading time analyses will be presented for the pre-critical region, the critical region and the following regions. For ease of reference, an example material (in the Good Context/Implausible condition) is given below, with analysis regions indicated:

$_1$The army had already begun attacking the city.|
$_2$The soldier|$_3$ was protected by|$_4$ **the child**|$_5$ during|$_6$ all the|$_7$ heavy|$_8$ shooting.|

Two ANOVAs were performed for each reading time measure, one by Participants (F_1) and one by materials, or items (F_2). Mean reading time measures are presented in Table 30, below.

Table 30: Mean Reading Time Measures for Critical and Spillover Regions

	Region					
	3 (pre-critical)	**4 (critical)**	**5**	**6**	**7**	**8**
Measure	*was protected by*	*the child*	*during*	*all the*	*heavy*	*shooting*
	Mean (SD)					
First Fixation (msec)						
Good Context	-	250 (32)	248 (52)	254 (44)	258 (53)	250 (49)
Neutral Context	-	252 (32)	251 (36)	246 (35)	281 (53)	268 (66)
First Pass (msec)						
Good Context	397 (125)	349 (86)	266 (70)	322 (64)	291 (63)	278 (68)
Neutral Context	410 (128)	356 (76)	271 (53)	321 (70)	322 (78)	298 (81)
First Pass Reg. Out						
Good Context	9.50 (11.35)	15.16 (12.24)	8.69 (18.41)	11.97 (11.97)	15.78 (17.87)	50.09 (30.92)
Neutral Context	7.81 (8.57)	17.38 (16.65)	5.72 (9.09)	11.16 (12.78)	22.91 (22.38)	50.34 (34.07
Regression Path (msec)						
Good Context	455 (141)	432 (115)	342 (152)	401 (100)	376 (129)	459 (233)
Neutral Context	472 (144)	461 (140)	304 (85	375 (98)	456 (247)	454 (256)
Total Time (msec)						
Good Context	520 (170)	412 (97)	295 (95)	378 (85)	329 (90)	302 (80)
Neutral Context	529 (173)	443 (110)	290 (69)	383 (91)	366 (116)	321 (95)

Reading Time results by region

To preview the results, there is no evidence that contextual fit modulates the time course of anomaly detection, and thus no evidence that there is a contextual fit heuristic operative at the earliest point in interpretation. There were no significant effects (in measures of early processing) at the critical region (or earlier). The earliest difference between the two context conditions was at the third spillover region, when the Neutral Context condition tended to have relatively longer reading times and increased regression behaviour.

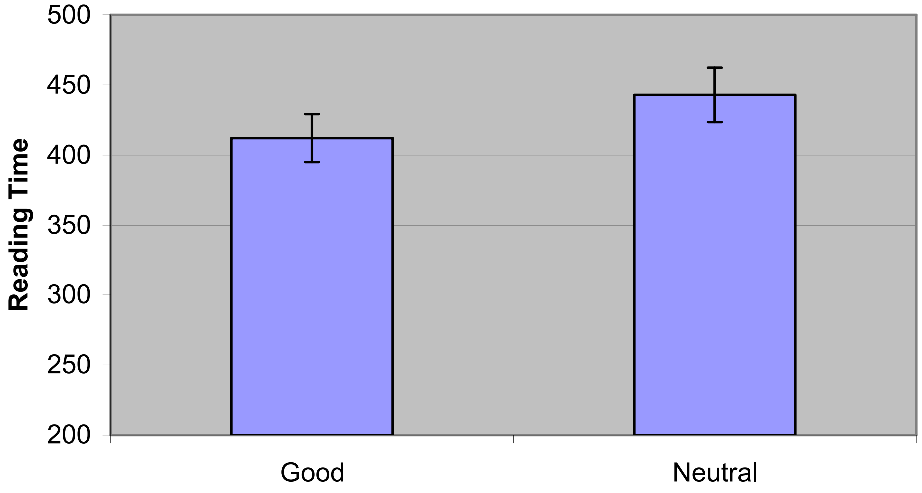
Region 3 (Pre-critical region)

There were no significant effects at all in the pre-critical region (F<1 or $p > 0.2$). We may note straight away that this is in contrast to experiment 7, where a marginal effect had been suggestive of early heuristic processing (and there were no significant effects in the data from either of the two reading ability groups; see chapter 6).

Region 4 (Critical Region)

There were no significant effects in first fixation time (both Fs < 1), first pass time (both Fs < 1), first pass regressions out ($F_1 < 1$; $F_2(1,23) = 1.715$, MSe = 43.739, $p > 0.2$), or regression path time ($F_1(1,31) = 2.173$, MSe = 6259.874, $p > 0.1$; $F_2(1,23) = 2.246$, MSe = 3740.068, $p > 0.1$). There was, however, a significant effect of context in the total time analysis, with longer total reading times in the Neutral Context condition ($F_1(1,31) = 5.030$, Mse = 3013.766, $p < 0.04$; $F_2(1, 23) = 5.900$, Mse = 1845.796, $p < 0.03$) (see Figure 30).

Figure 30:
Total Reading Time (msec) Critical Region

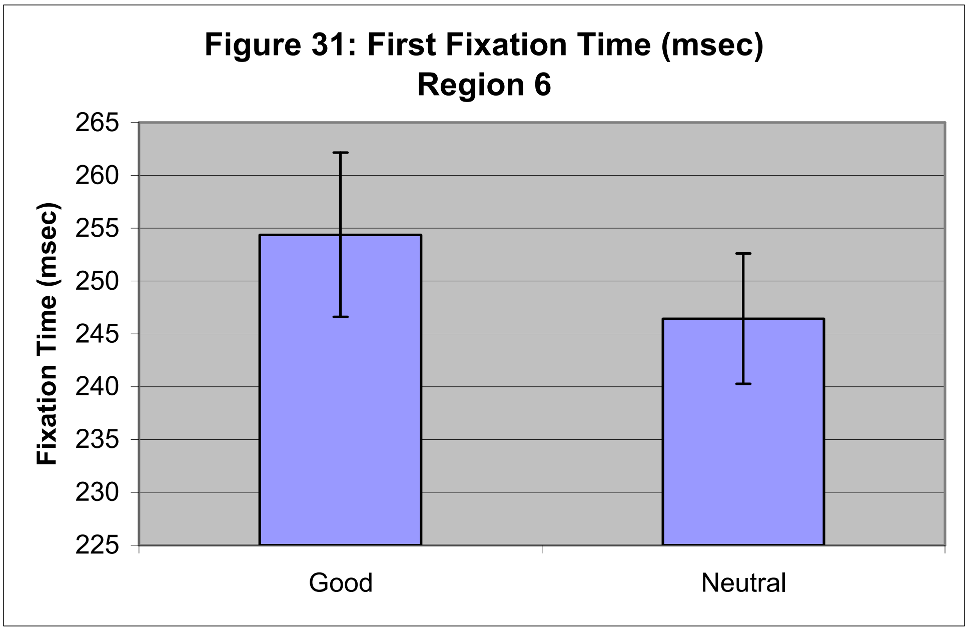


Region 5 (first spillover region)

There were no significant effects at all in this region (first fixation, first pass, first pass regressions out and total time: all *F*s < 1; regression path time: $F_1(1,31)$ 1.522, MSe = 15035.079, $p > 0.2$; $F_2(1,23) = 1.001$, MSe = 7297.159, $p > 0.3$).

Region 6 (second spillover region)

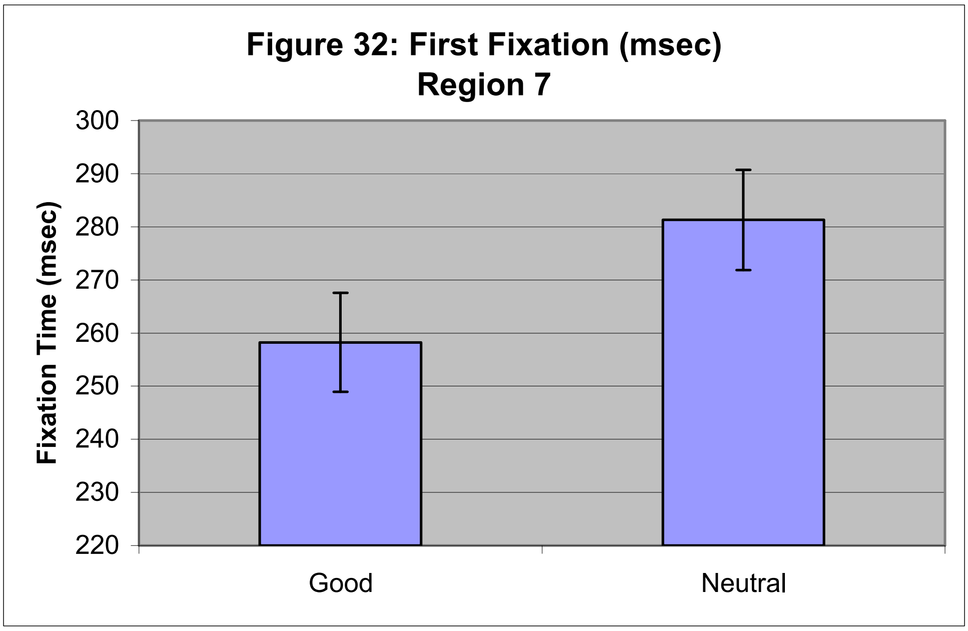
In region 6 there was an effect of context, significant by items only, indicating a slightly longer first fixation time in the Good Context condition ($F_1(1,31) = 1.839$, Mse = 548.256, $p > 0.1$; $F_2(1, 23) = 6.933$, Mse = 214.231, $p < 0.05$, see Figure 31). There were no significant effects in first pass (both *F*s < 1), first pass regressions out (both *F*s < 1), regression path ($F_1(1,31) = 2.569$, Mse = 4109.046, $p > 0.1$; $F_2(1, 23) = 2.071$, Mse = 2804.652, $p > 0.1$), or total time (both *F*s < 1).



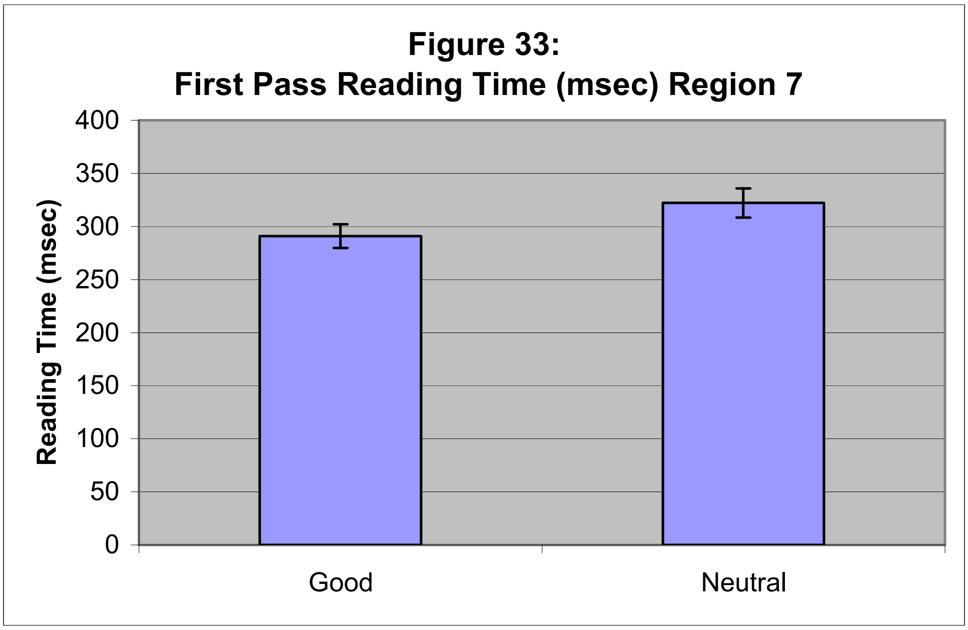
Given the lack of robust effects across the board in this region, it seems highly unlikely that the first fixation effect is genuine, i.e. it is unlikely that it indexes anomaly detection in the Good Context condition ahead of the Neutral Context condition.

Region 7 (third spillover region)

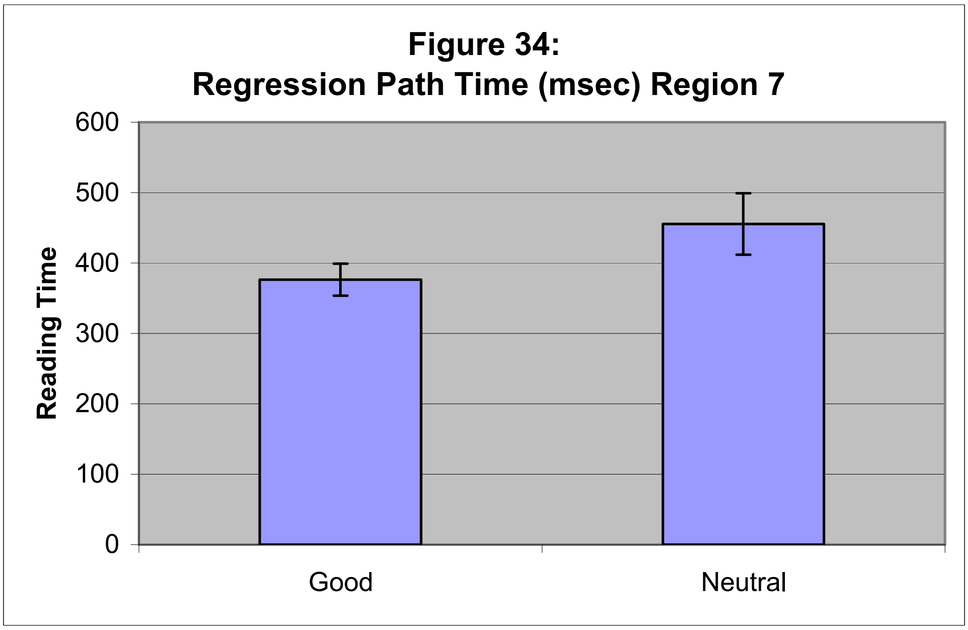
In first fixation analysis there was a significant effect of context, with longer initial fixation times in the Neutral Context condition $F_1(1,31) = 6.281$, Mse = 1351.145, $p < 0.02$; $F_2(1, 23) = 4.777$, Mse = 1179.377, $p < 0.05$) (see Figure 32).



The same effect was found in first pass $F_1(1,31) = 6.771$, Mse = 2335.417, $p < 0.02$; $F_2(1, 23) = 5.884$, Mse = 1690.695, $p < 0.03$) (Figure 33).



There was nothing significant in first pass regressions out (F_1(1,31) = 2.873, Mse = 282.734, $p > 0.1$; F_2(1, 23) = 1.117, Mse = 209.518, $p > 0.3$). Regression path analysis revealed an effect of context, significant by participants, with longer reading times in the Neutral context condition (F_1(1,31) = 4.795, Mse = 20908.165, $p < 0.05$; F_2(1, 23) = 2.453, Mse = 21095.768, $p > 0.1$) (Figure 34).



There was also an effect of context in total time, significant by participants and marginal by items, with longer total times in the Neutral condition ($F_1(1,31) = 8.510$, Mse = 2666.020, $p < 0.05$; $F_2(1, 23) = 3.374$, Mse = 3408.716, $p = 0.079$).

Given the results of our previous studies, which have shown immediate, or close-to-immediate anomaly detection under various reading conditions, it would be reasonable to interpret the effects in this region in terms of differential *recovery* from anomaly-related disruption, rather than an initial (and quite late) slowdown in the Neutral Context condition. In other words, we can assume that anomaly detection has occurred by this region in both conditions, with no differences in time course of detection. Support for this comes from the fact that in Experiment 2, a self-paced reading study using similar implausible materials, an effect of plausibility, i.e. anomaly detection, was observed on the second word after the critical word. Processes that are visible in self-paced reading must occur at the same time, or even earlier, in eyetracking. Also, there has so far been no evidence that implausible passives are treated differently, online, to implausible actives in the matter of anomaly detection; as the previous experiment failed to find any context-related difference in anomaly detection for implausible actives, it would be rash to insist that the effects we see here are detection effects rather than recovery effects. The longer reading times for the Neutral condition at this advanced point in the processing stream, combined with the increased total reading times in the same condition at the critical region, suggest instead that recovery from the anomaly is more difficult following a Neutral context than following a Good one. Without a Plausible condition to measure the Good Context/ Anomalous condition against, we cannot say that the disruption has died down altogether at this point; but it has certainly lessened relative to the Neutral Context/ Anomalous condition.

Region 8 (wrap-up)

There was a marginal effect (by participants) of context in first fixation (F_1(1,31) = 3.251, MSe = 1539.627, p = 0.081; F_2 < 1), with longer initial fixation times in the Neutral Context condition. There was nothing significant in first pass analysis (F_1(1,31) = 2.149, MSe = 2922.433, p > 0.1; F_2 < 1), first pass regressions out (both Fs < 1), regression path time (F_1(1,31) = 1.418, MSe = 13112.004, p > 0.2; F_2 < 1) or total time (F_1(1,23) = 1.798, MSe = 3222.370, p > 0.1; F_2 < 1)

Summary of eyetracking results

The results relating to the hypothesis are clear. The lack of a significant difference between the two context conditions, in measures indexing anomaly detection, indicates that contextual fit did not modulate the time course of anomaly detection. The effects occurring further downstream are most likely to reflect an earlier recovery from anomaly-related disruption following the Good context, a result seen to some extent in experiment 7.

General Discussion

The materials used in experiment 8 were intended to increase the chances of shallow processing relative to experiment 7, and so provide the best chances of observing the goodness-of-fit heuristic in operation. The anomalies were phrased in the passive voice and contained pragmatic cues to bias towards a non-syntactic interpretation. The question was whether, under these conditions favourable to shallow processing, the time course of online anomaly detection would be affected by the nature of the preceding context.

The results failed to provide any evidence that the time course of detection could be affected by context type. Although the absence of plausible control conditions masked the actual moment of detection in each condition, there were no differences between context conditions until an advanced stage, by which time detection had almost certainly taken place. Examination of the means for the various measures indicates that both conditions were virtually identical at the pre-critical, critical, first and second spillover regions; any invisible difference in the time course of detection would have required a substantial difference in the reading of (imaginary) plausible controls, and there would be no good reason to posit such a difference. The lack of a difference before the third spillover region would suggest that anomaly detection had taken place, either immediately as with experiment 7, or after some very small delay, in both conditions, and had caused an equal amount of disruption in each. As such, the results here support, or at least do not challenge, a processing model in which interpretations are first generated using syntactic information. Contrary to experiment 7, there were no marginal effects suggestive of semantic-first processing, either in the overall analysis or analysis by reading ability (see chapter 6).

The significant differences between conditions that emerge at the third spillover region suggest that recovery from the disruption caused by the anomaly is easier following a good context condition than a neutral context (this would certainly be in line with the results of experiment 7). This interpretation is bolstered by the difference in total reading time at the critical region, which saw more time in total spent on the region in the Neutral Context condition. The common finding from the two experiments, then, is that a good context – that is, a semantically supportive context – appears to aid recovery from the disruption caused by an anomaly, relative to a semantically neutral condition.

One possible explanation for this is the specificity effect described by Sanford and Garrod in their work on anaphoric reference (A J Sanford and Garrod, 1980). The authors reported that integration of subsequent material, including an anaphor, was easier following an antecedent that was specific as opposed to general. A sentence such as *the vehicle was overloaded* was read faster following the sentence *the lorry could not get up the hill*, than when the two co-referring NPs were switched to give *The vehicle could not get up the hill. The lorry was overloaded*. The authors suggested this effect was due to the richer scenario representation evoked by the sentence containing the specific antecedent (lorry), which enabled an easier integration, or mapping, of the following information into that representation. Further evidence comes from a study (A. J. Sanford, Garrod and Bell, 1979) in which a specific representation (*knife*) facilitated all NP anaphors better than a non-specific representation (*weapon*). It could be that the good contexts in the present experiments had a similar effect, encouraging the creation of a richer scenario at the outset, and enabling faster integration of the information in the critical sentences (though anomalous) as a result. However, as there is no more recent evidence in support of this view, it must be conjectural for now.

Chapter 6:

Individual Differences

Following the main part of the experimental sessions in experiments 4-8, our participants completed a test of reading comprehension ability. We were investigating whether we could replicate recent findings in the literature suggesting that less-skilled readers are more prone to shallow processing than skilled readers, and also whether there are any observable processing differences between skill groups that might account for any differences in interpretation skill. This latter was purely an investigative project – we made no predictions regarding the behaviour of either skilled or less-skilled readers.

The results of our individual differences analyses will be presented here in summary form only, as the analyses generally did not yield many interesting findings and a full presentation of the data would effectively treble the space needed to present all results from the previous five experiments. The focus will be on the interpretation accuracy data from experiments 4-6, and some differences observed between skill groups in the reading time data. To preview, there were some departures from the current (small) literature on individual differences in shallow processing, but nothing consistently different. Reading time analysis revealed one interesting finding (experiment 7) and also some evidence suggesting that some of our pragmatic anomalies were detected *parafoveally*. This interesting in light of the discussion of anomaly timing effects at the beginning of Chapter 4, with the studies reviewed tending to focus largely on delayed effects, or effects only on critical regions (i.e. not regions that took the possibility of parafoveal preview into account). However, these new findings must be considered preliminary and partial, and cannot by themselves generate a theory of how skilled and less-skilled readers process anomalous material. As such, our analyses generally reflect the findings of two studies already referred to (Daneman and Hannon, 2004; Daneman et al., 2007), in which reading time analyses were collapsed across reading skill as analyses including it as a factor yielded no significant results.

Nothing resulting from these analyses poses a serious challenge to the conclusions so far adduced in this thesis, and they are included for the sake of completeness. Any conclusions are therefore tentative and subject to revision by much-needed further work on the question of whether the tendency to normalise differs as a function of reading ability.

The reading ability test was the Reading Comprehension section from the Nelson Denny Reading Test (Form E: Brown, Bennett and Hanna, 1981). In experiments 4-8 participants completed this as part of the experimental session. According to the Nelson Denny scoring norms, a score of 25.25 represents the 50th percentile. Participants scoring 25 or less were classified as Less-skilled readers, and participants with a score greater than 25 were classified as Skilled readers.

All analyses of processing measures used participant means only (F_1 only) due to an unacceptable number of missing cells in the F_2 data, especially in the data from the Skilled reader groups. Reading groups were analysed separately and we were not concerned with direct comparisons: effects observed in the overall analyses (i.e. the analyses presented up to now), and in the separate analyses below, are often not very robust (e.g. experiment 6, 7) and would be less visible if investigated using less powerful tests involving between–subjects analysis.

Experiment 4:
Interpreting Elliptical Verb Phrases (EVPs): A Self-paced Reading Study.

Summary of findings

Less-skilled readers were no more prone to normalisation than skilled readers. Both skill groups correctly interpret the anomalies online, but less-skilled readers appear to detect them, and recover from the disruption they cause, slightly earlier than skilled readers.

Eighteen participants were classified as Skilled readers (Mean 29.9, SD 1.88), 14 were classified as Less-skilled readers (Mean 19.57, SD 3.55).

Accuracy Results

Mean accuracy results for both reading ability groups is presented in Table 31 and Figure 34.

Table 31: Mean Accuracy results for Skilled and Less-Skilled readers

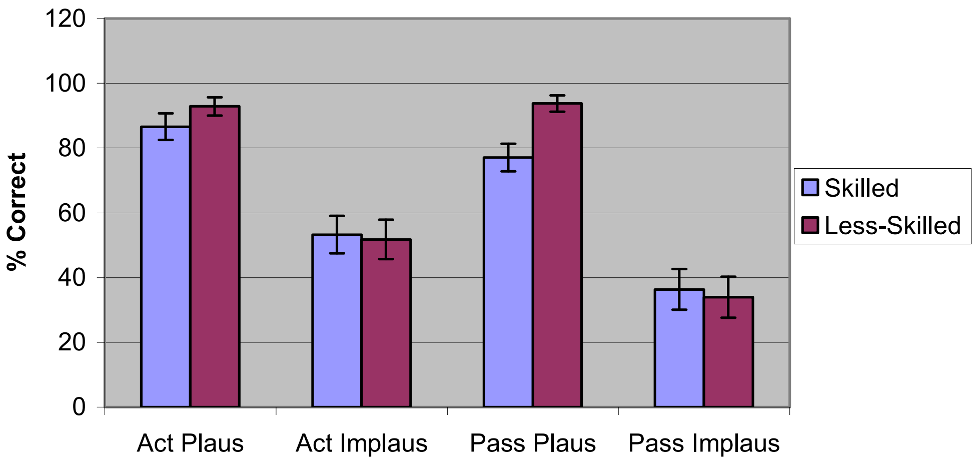
	Condition			
% Accuracy	**Active Plausible**	**Active Implausible**	**Passive Plausible**	**Passive Implausible**
Skilled Readers	86.57 (17.42)	53.24 (24.53)	77.08 (17.98)	36.34 (26.65)
Less-skilled Readers	92.86 (10.65)	51.79 (22.63)	93.75 (9.49)	33.93 (23.62)

Analysis of the Skilled readers accuracy yielded a main effect of Voice, with greater accuracy in the Active conditions ($F_1(1, 17) = 7.868$, MSe = 398.284, $p <$ 0.05; $F_2(1, 29) = 10.229$, MSe = 478.847, p 0.005). There was also a significant main effect of Plausibility with greater accuracy in the Plausible conditions ($F_1(1, 17) = 51.964$, MSe = 475.161, $p < 0.001$; $F_2(1, 29) = 87.435$, MSe = 526.341, $p < 0.001$). There was no significant interaction however ($F_1(1,17) =$ 1.041, MSe = 237.212, $p > 0.3$; $F_2 < 1$).

Analysis of the Less-Skilled readers accuracy yielded only a marginal main effect of Voice, with greater accuracy in the Active conditions ($F_1(1, 13) =$ 3.824, MSe = 263.398, $p = 0.0724$; $F_2(1, 29) = 2.992$, MSe = 503.053, $p =$ 0.0943). There was a significant main effect of Plausibility with greater accuracy in the Plausible conditions ($F_1(1, 13) = 68.920$, MSe = 516.946, $p < 0.001$; $F_2(1, 29) = 161.243$, MSe = 488.685, $p = 0.001$). There was a significant interaction ($F_1(1,13) = 8.913$, MSe = 138.054, $p < 0.05$; F_2 (1, 29) = 5.252, MSe = 524.605, $p < 0.05$). Planned Comparisons confirmed there was no significant difference between the two Plausible conditions (both Fs < 1) and that there was a significant difference between the two Implausible conditions ($F_1(1,13) = 8.097$, MSe = 275.629, $p < 0.05$; F_2 (1,29) = 5.043, MSe = 826.149, $p < 0.05$)

A three-way ANOVA including reading ability failed to yield a significant effect of Reading Ability ($F_1(1, 31) = 1.150$, MSe = 623.021, $p > 0.2$; $F_2(1, 29) = 2.741$, MSe = 446.051, $p > 0.1$). However, there was an interaction between reading skill and Plausibility, significant only by items ($F_1(1, 31) = 2.871$, MSe = 493.268, $p > 0.1$; $F_2(1, 29) = 5.002$, MSe = 537.829, $p < 0.05$). As Figure 35 shows, this interaction seems to be driven by a difference in the Passive Plausible condition, with the Less-skilled readers interpreting with greater accuracy. There also appears to be a similar (marginal) difference in the Active Plausible condition. These differences perhaps reflect a more cautious overall approach to the task among less-skilled readers, with more confident, skilled readers making more errors in the 'easiest' conditions (although there was no evidence of greater caution in a comparison of decision time – both *F*s < 1 – and analysis of reading times did not indicate that less-skilled readers were simply reading more slowly – all *F*s < 1).

Figure 35: Accuracy (%) by Reading Ability



Overall then, there are no major differences in the success of Skilled and Less-skilled readers in interpretation, and both groups are apparently equally prone to shallow processing and misinterpretation. This finding is contrary to findings by Daneman and colleagues (e.g. Hannon and Daneman, 2004) which showed that less-skilled readers were significantly more susceptible to shallow processing than skilled readers.

Reading time results

There was a difference in the timing of anomaly detection between the two skill groups, with evidence of anomaly detection appearing in the first spillover word for the less-skilled group (marginal effect of Plausibility (by participants) with longer reading times in the Implausible conditions ($F_1(1,13) = 4.368$, MSe = 11895.667, $p = 0.057$), reaching significance by the second spillover word ($F_1(1,13) = 5.151$, MSe = 9007.424, $p < 0.05$; $F_2(1,31) = 4.697$, MSe = 21539.420, $p < 0.05$). With the skilled readers there was no evidence of detection until the second spillover word (marginal effect of Plausibility, by items, with longer reading times in the Implausible conditions ($F_1(1, 17) = 1.480$, MSe = 13699.662, $p > 0.2$; $F_2(1, 31) = 3.434$, MSe = 15605.234, $p = 0.073$), not reaching significance until the third spillover word ($F_1(1, 17) = 5.009$, MSe = 1858.112, $p < 0.05$; $F_2(1, 31) = 2.612$, MSe = 7552.638, $p > 0.1$). Also, the disruption had disappeared for the less-skilled group by the fourth spillover region (both Fs < 1), but was still active for the skilled group ($F_1(1, 17) = 7.082$, MSe = 6537.795, $p < 0.05$; $F_2(1, 31) = 4.843$, MSe = 18276.580, $p < 0.05$).

Both skilled and less-skilled readers are susceptible to normalisation when interpreting implausible EVPs, and less-skilled readers are no more susceptible than skilled readers. Both skill groups detect anomalies online, but less-skilled readers detected them earlier than skilled readers and also seemed to recover from the disruption earlier. If we had seen a difference in the accuracy results, with more normalisation in the less-skilled groups, we might have out this earlier recovery down to the online substitution of the problematic interpretation by a more plausible heuristic one. As there are no such differences, however, this does not seem a very likely interpretation – the skilled readers clearly held the syntactic interpretation later into the processing stream, and yet still normalised to the same degree. We could perhaps allow that some readers may apply heuristic interpretations earlier than others, and the online/offline distinction we have been considering is not a genuine dilemma, but there is not enough evidence to justify firm conclusions.

Experiment 5:

Interpreting elliptical verb phrases (EVPs): An eyetracking study.

Summary of findings

Contrary to the results of the self-paced reading version of this study, less-skilled readers were more prone to normalising than skilled readers (see Table 32 and Figure 36). Both skill groups detect anomalies online, but this time it is the skilled readers who recover from the disruption first. A finding not seen in the overall analysis is that less-skilled readers appear to detect the anomaly parafoveally in the Active Implausible condition.

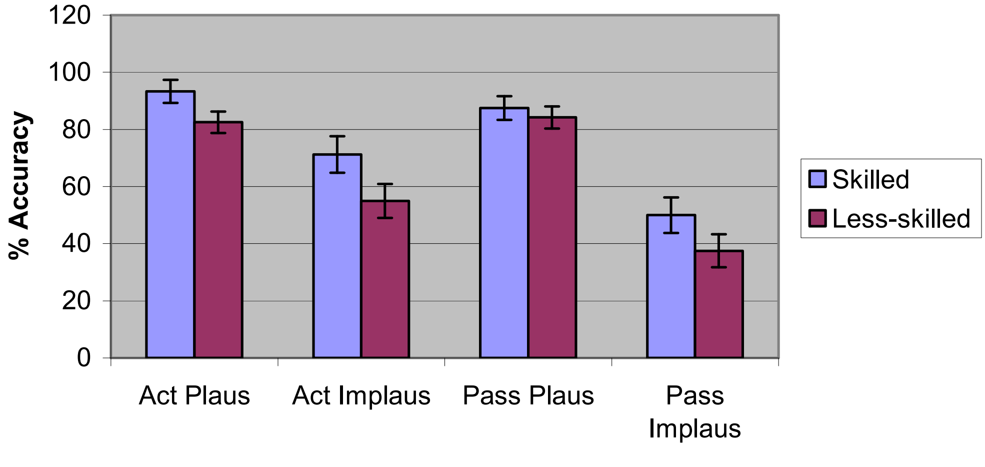
As four participants were unavailable to complete the test, analysis for this experiment includes only 28 participants. Thirteen participants were classified as Skilled readers (Mean 30.08, SD 2.14) and 15 as Less-skilled readers (Mean 19.53, SD 4.07).

Accuracy results

Table 32: Mean Interpretation Accuracy (%) for Skilled and Less-skilled Readers

	Condition							
% Accuracy	**Active Plausible**		**Active Implausible**		**Passive Plausible**		**Passive Implausible**	
Skilled Readers	93.3	(9.7)	71.2	(17.2)	87.5	(10.2)	50.0	(23.9)
Less-skilled Readers	82.5	(17.6)	55.0	(27.1)	84.2	(18.0)	37.5	(21.1)

Figure 36:
Interpretation Accuracy (%) by Reading Ability



In this version of the experiment, a mixed ANOVA indicated there was a significant effect of the Reading Ability (between subjects) factor, with the skilled reading group interpreting with greater accuracy than the less-skilled group (F_1 (2, 26) = 5.625, MSe = 565.844, $p < 0.05$). So, while strangely in contrast to the results of experiment 4, which used exactly the same materials, these results are more in keeping with the Daneman studies which reported greater levels of normalisation from less-skilled readers compared to skilled readers.

Reading Time Results

As with the previous experiment, there was evidence that both groups had detected the anomalies online. Contrary to the previous study, the skilled readers recovered from the anomaly-related disruption ahead of the less-skilled readers, displaying no difficulty after the first spillover region.

As the skilled readers performed better on the interpretation task, we can venture that recovery from anomaly-related disruption does not index the late, online application of a heuristic interpretation. If it did then we might expect to see the group that recovered earliest producing more normalised interpretations. Timing of recovery, while possibly related to reading ability, is not therefore related to the tendency to produce normalised interpretations.

In the less-skilled readers' data, there was evidence of parafoveal detection of the active anomaly, but not the passive anomaly. In the first pass analysis of the pre-critical region there was a significant Voice*Plausibility interaction (F_1(1,14) = 7.707, MSe = 2925.067, $p < 0.05$). Planned comparisons confirmed there was a significant difference between the means of the two Active conditions ($t_1(14) = 2.340$, $p < 0.05$) and no significant difference between the means of the two Passive conditions ($t_1 < 1$). There is therefore evidence suggesting that less-skilled readers detected the Active anomaly prior to directly fixating the anomalous region itself, and ahead of the skilled readers. This does open the possibility that, at this point at least, the less-skilled readers' interpretations may have been subject to the early operation of a plausibility heuristic in the Passive case. If so, the fact that they then went on to detect the passive anomalies indicates that this was not their sole interpretation strategy, and that syntax soon exerted its influence. (And, as discussed in the next chapter, this would lead to a strangely non-parsimonious account of online interpretation.) It must also be noted that this effect, if real, would probably not have been predicted by parafoveal processing theories. If readers were fixating the end of the pre-critical region it would be quite surprising if they were able to detect the ellipsis in advance, which is after all signalled by two whole words (*had too*).

So for the less-skilled readers, some plausibility-based processing may have been active early in the interpretation of the Passive conditions – an effect which might have been predicted by the very low interpretation accuracy in the Passive Implausible condition. However, we can note that, even for the less-skilled readers, the correct syntax-based interpretation was made quickly afterwards; and if this difference between the actives and passives does reflect non-syntactic processing, we have no evidence that it was performed because the syntactic processing was too challenging – a point relevant when considering the conditions for shallow processing.

Experiment 6:

Interpreting Implausible Text: A Further Eyetracking Study

Summary of findings

Less-skilled readers were no more likely to normalise than skilled readers (see Table 33 and Figure 37). Analyses of reading time data for both groups, while not indicating any deviations from the overall analysis, were characterised by a scarcity of significant effects. This is most likely a statistical power issue.

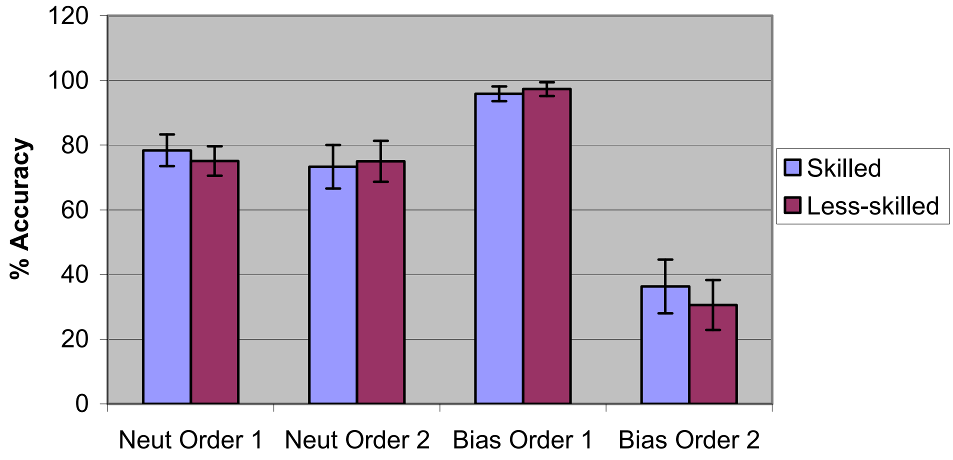
As four participants were unavailable to complete the test, analysis for this experiment includes only 28 participants. Thirteen participants were classified as Skilled readers (Mean 30.08, SD 2.14) and 15 as Less-skilled readers (Mean 19.53, SD 4.07).

Table 33: Accuracy Results (%) For Both Reading Groups

			Condition					
% Accuracy	**Neutral Order 1**		**Neutral Order 2**		**Biased Order 1**		**Biased Order 2**	
Skilled Readers	78.4	(19.1)	73.3	(24.8)	95.9	(8.7)	36.3	(31.3)
Less-skilled Readers	75.1	(16.1)	75.0	(24.0)	97.3	(7.8)	30.6	(28.3)

A mixed ANOVA with reading ability as a between-subjects factor failed to yield a significant effect of reading ability ($F_1 < 1$) – skilled readers were thus no more successful in the interpretation task than less-skilled readers. So we do not have a consistent picture from experiments 4-6 regarding whether or not less-skilled readers are more prone to normalising than skilled readers.

Figure 37:
Interpretation Accuracy (%) by Reading Skill



Reading Time Results

With the less-skilled readers there were very few significant results. As with the previous experiment, there was some evidence in the pre-critical region suggesting that the anomaly had been detected prior to fixation. The effect was marginal, however, so we cannot conclude with certainty that the detection took place this early (first pass regressions out, marginal interaction (F_1(1, 14) = 3.356, MSe = 38.460, $p < 0.088$). The critical region contained the same (confound-related) main effect of the Order variable as in the overall analysis, with longer first pass times in the Order 2 conditions (F_1(1,14) = 7.211, MSe = 5660.917, $p < 0.02$). This effect of order appeared to spillover into the first spillover region, appearing as a marginal effect in the Regression Path analysis (F_1(1,14) = 4.161, MSe = 23575.767, $p = 0.061$). There is only a suggestion in the means that anomaly detection occurred in the second spillover region (first fixation time, region 5) but there was no significant interaction ($F_1(1,14) = 1.338$, MSe = 513.352, $p > 0.2$). The only significant effect was an interaction in region 7 – the wrap-up region – in the regression path analysis, which was identical to the wrap-up effects seen in the overall analysis ($F_1(1,14) = 4.759$, MSe = 15282.136, $p = 0.05$). There were marginal interactions of the same nature in first fixation ($F_1(1,14) = 3.746$, MSe = 599.210, $p = 0.073$) and first pass regressions out ($F_1(1,14) = 4.272$, MSe = 274.445, $p = 0.058$). For all other measures in all regions, $F_1 < 3$ and $p > 0.1$.

For the skilled readers, apart from the standard wrap-up effects, which did not differ from the overall analysis or the analysis of the less-skilled readers, and marginal effects of the Order variable in region 5 (first pass regressions out: $F_1(1,12) = 3.191$, MSe = 150.452, $p = 0.099$; regression path: $F_1(1,12) = 3.605$, MSe = 4583.641, $p = 0.082$), there were no significant effects. In the critical region, there was a numerical difference between the means of the two Biased conditions (Biased order 1: 271 msec; Biased Order 2: 303 msec) which may reflect an early anomaly detection, but there was no significant interaction. Apart from the wrap-up effects and marginal effects of Order, all other *F*s < 3 and *p*s > 0.1.

The general lack of significant effects cannot be taken as evidence that either group failed to detect the anomaly online (and therefore engage in online, syntax-based interpretation). The effects in the overall analysis were not very robust, and we must assume that the lack of clear effects here relates to the lower power in these analyses. Overall, then, these results do not offer a clear picture of the processing styles of either skill group.

Experiment 7:

Normalisation and Goodness-of -Fit

No accuracy data was collected in either experiment 7 or experiment 8. We will look only at the reading time data. One participant was unable to complete the reading comprehension test, and analysis was carried out on reading time data from 31 participants. Fifteen were classified as skilled readers (Mean 30.0, SD 2.507), 16 were classified as less-skilled (Mean 21.19, SD 3.85).

Summary of findings:

Both skill groups detect the anomaly online. Contrary to the overall analysis in chapter 5, analyses of the pre-critical region suggests that the skilled readers detected both anomalies as early as the pre-critical region, while the less-skilled readers apparently detected only the Neutral Context anomaly. Skilled readers recover from the anomaly slightly earlier than less-skilled readers.

Reading Time Results

Recall that in the main analysis of experiment 7 there was a marginal effect in the pre-critical region (first pass), possibly indicating that the anomaly had been detected in the Neutral context but not the Good context conditions. This result, if robust and genuine, would have argued for the early operation of a Goodness-of-fit heuristic. In the first pass analysis for the less-skilled readers there was a significant context*plausibility interaction ($F_1(1,15) = 4.984$, MSe = 5323.066, $p < 0.05$) (see Figure 38). Comparisons indicated that there was no significant difference between the two Good context conditions ($t_1 < 1$), and that there was only a marginal difference between the means of the two Neutral context conditions (t_1 (15) = 2.022, p = 0.061). While the direct comparison of the Good context conditions did not yield a significant result, this is potentially evidence for a semantics-first strategy with less-skilled readers.

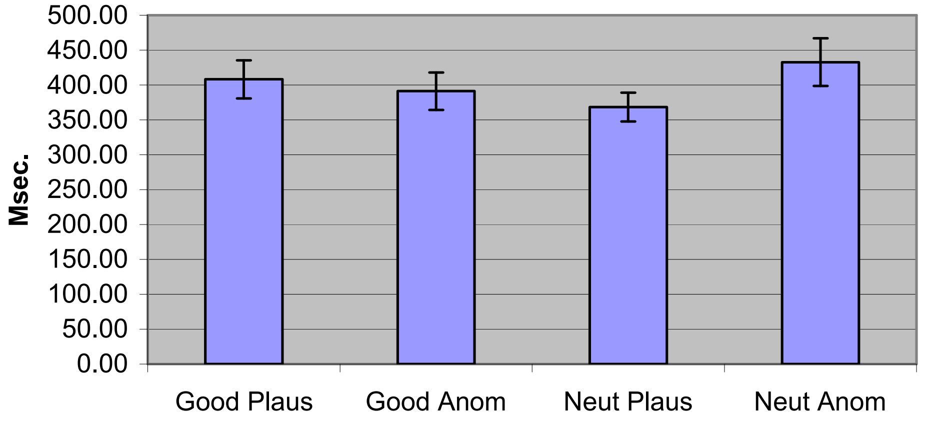


Figure 38:
First Pass Reading Time
Pre-critical Region (Less-skilled Readers)

Following this effect, disruption was seen in the critical region for both anomaly types, in first pass and regression path analyses, which yielded significant effects of the plausibility variable (first pass: $F_1(1, 15) = 4.763$, MSe = 4858.316, $p < 0.05$; regression path: $F_1(1, 15) = 9.340$, MSe = 10742.263, $p < 0.05$). In the following region both anomalies were causing significant disruption in both regression measures (first pass regressions: $F_1(1, 15) = 11.123$, MSe = 271.967, $p < 0.05$; regression path: $F_1(1, 15) = 14.363$, MSe = 38268.591, $p < 0.005$). In the second spillover region, a significant interaction in first fixation time ($F_1(1, 15) = 10.605$, MSe = 557.226, $p < 0.05$) was driven by a marginal difference between the two neutral conditions ($t_1(15) = 2.048$, $p = 0.058$). In first pass analysis, a main effect of context indicated that the Good context conditions were being read for longer ($F_1(1, 15) = 5.072$, MSe = 2762.807, $p < 0.05$), but a marginal interaction, whose means followed the same pattern as those in the first fixation interaction, suggested the Neutral conditions were the only ones reflecting a plausibility effect ($F_1(1, 15) = 3.957$, MSe = 3219.599, $p = 0.065$). In the third spillover region (region 6) effects in regression behaviour indicated that the implausible conditions were still causing disruption.

Evidence for anomaly detection among the skilled readers was also observed in the pre-critical region, with a significant main effect of plausibility in the first fixation measure, indicating a parafoveal detection of anomaly in both conditions ($F_1(1, 15) = 19.986$, MSe = 377.710, $p < 0.05$). So unlike the less-skilled readers, there is nothing to suggest that early interpretation was not syntax-driven. Disruption continued into the critical region itself (first fixation, first pass, regression path; all Plausibility *p*s < 0.02; first pass regressions out: $p = 0.084$), and then into the first spillover region, where there was also a marginal effect of context suggesting more regressions in the Good Context conditions ($F_1(1, 15) = 3.756$, MSe = 203.231, $p = 0.073$). Region 5 contained further plausibility effects but the means suggested a slightly greater difficulty with the Neutral conditions, and a significant effect of context in regression path analysis indicated that skilled readers had longer times in the neutral conditions ($F_1(1, 15) = 6.107$, MSe = 7776.576, $p < 0.05$). By region 6 there was no evidence at all of plausibility-related disruption (all *F*s < 1).

The main finding of interest, then, is the parafoveal effect in the less-skilled readers' data. While it remains interesting and relevant that detection of pragmatic anomalies can occur prior to direct fixation, the fact that less-skilled readers appeared to detect only the Neutral Context anomalies at this stage offers support to the idea that they were using the statistical fit of context words to guide interpretation. The robust plausibility effects that were observed quickly afterwards (i.e. the critical region) demonstrate that this was certainly only a preliminary strategy. However, the finding does prohibit a firm conclusion on whether or not readers generally use this kind of information early in interpretation, and suggests that a final answer will lie in further work with readers of differing reading ability.

Other than the parafoveal effects, the only difference of interest between the two skill groups is in the timing of recovery, with the skilled readers appearing to recover from the anomaly one region earlier than the less-skilled readers.

Experiment 8

Normalisation and Goodness-of-Fit: A Further Study

One participant was unable to complete the reading comprehension test, and analysis was carried out on reading time data from 31 participants. Fifteen were classified as skilled readers (Mean 30.0, SD 2.507), 16 were classified as less-skilled (Mean 21.19, SD 3.85).

There were no significant results at all in the data from the skilled readers (all *F*s < 3, all *p*s > 0.1), so we will look only at results from the less-skilled readers. The only deviation from the overall analysis was a difference in first fixation on the second spillover region, with longer fixation times in the Good Context condition ($F_1(1,15) = 6.521$, MSe = 695.615, $p < 0.03$). Thereafter, any significant differences involved, as with the overall analysis, longer reading times/more regressions in the Neutral Context condition. There were no significant effects in the pre-critical region for either skill group.

The lack of significant effects in the skilled readers data is the only substantial difference so far observed between skill groups. However, it is highly unlikely to relate to our main concern which is anomaly detection. Recall, from the previous chapter, that the effects observed in experiment 8 are best interpreted as differential recovery effects, with observed differences being caused by the influence of Good and Neutral contexts on recovery from anomaly-related disruption. In interpreting these data from the skilled groups, we could only conclude that Context does not play a significant role in the recovery from anomaly disruption, as it certainly appears to for less-skilled readers. The other alternative is that context does affect recovery for skilled readers, but only weakly, and the lower power of our analysis obscures these effects.

Conclusion

We have seen that the results from our reading ability analyses have not been consistent. In two studies out of three we failed to observe any influence of reading ability on the tendency to produce normalised interpretations. Oddly, though, skilled readers were the more successful interpreters in experiment 5, but not in experiment 4, despite both experiments using the same materials. The only conclusion we can draw from this, in combination with the few reports in the literature, is that less-skilled readers will not *necessarily* generate normalised interpretations to a greater extent than skilled readers.

In terms of differential processing styles, there are some results worth discussing. Firstly, splitting the reading time data by comprehension ability revealed some effects that were obscured in the overall analyses, namely, parafoveal detection of anomalies. In experiment 5, less-skilled readers appeared to detect the Active condition anomaly in the pre-critical region. However, given that the critical region itself – the ellipsis – consisted of two words, there is good reason to wonder if this effect is a genuine preview effect. If it is genuine, we must allow that an earlier detection of the Active condition anomaly may be consistent with an early non-syntactic processing of the Passive anomaly, an processing affect perhaps predictable in light of the poor accuracy with passive anomalies. In experiment 6 there is only a (non-significant) suggestion of parafoveal anomaly detection, again among the less-skilled readers. In experiment 7 there is the intriguing possibility that the less-skilled readers are using contextual fit to guide interpretation, and consequently detect only the Neutral context anomaly in the pre-critical region (the skilled readers detect both anomalies in this region).

Overall, these analyses do not challenge the conclusion that pragmatic anomalies are detected online, and very early. In fact, these parafoveal effects, while generally weak, suggest that eyetracking studies on plausibility should be looking for detection ahead of the ‘critical’ region. The parafoveal effects for the less-skilled readers in experiment 7 perhaps present the only real challenge in this thesis to the theory that readers will always make an initial interpretation based on syntax. It must be noted, however, that this result was not seen in the data for the skilled readers, nor was it seen in experiment 8. As stated at the beginning of this chapter there is nothing consistently observed that would allow us to seriously model online processing of implausible material for either skilled or less-skilled readers. As a final comment on these possible preview effects, we must submit that it is very surprising to discover them in the data from the less-skilled readers – it surely goes against intuition that readers who were capable of exploiting pragmatic information in non-fixated material should be classified as less-skilled, and more powerful studies would likely be needed before we could accept findings that apparently show skilled readers being out-performed by less-skilled readers.

In experiment 4, the less-skilled readers detect the anomaly earliest and recover from it earliest. In experiment 5 (same materials), it is the skilled readers who recover fastest, and similarly in experiment 7. Experiment 8, with its focus on recovery effects, appears to show that a well-fitting context aids recovery relative to a neutral context, but only with less-skilled readers (the exact timing of recovery with skilled readers is obscured due to the design not featuring plausible control conditions). Overall, we can say that both skill groups detect anomalies online, with neither group clearly excelling the other in the timing of detection or recovery. Ultimately, we must allow that these findings are drawn from lower-power analyses, and further studies are required to resolve any inconsistencies and build confidence, if indeed there are genuine difference to be found.

Chapter 7:

Conclusion

This thesis began with the statement of five points for investigation. The following discussion will address them in approximately the same point-by-point format and order in which they were originally outlined. It will begin with a brief summary of the broad topic under consideration, and after that each main conclusion point will be dealt with in its own section:

1. The present studies' contribution to the evidence for normalisation and shallow processing.
2. Conclusions relating to the time course of syntax-based and heuristic sentence interpretation.
3. Conclusions relating to the conditions for normalisation.
4. The relevance to these findings to auditory comprehension.
5. Findings related to the goodness-of-fit heuristic.
6. Conclusions relating to individual differences.
7. Suggestions for future work.

Overview: Shallow processing and normalisation

Normalisation is only one instance of shallow processing but it has received considerable attention since it was first reported and can be considered a central area in the overall topic. Ferreira's important paper (2003) on Good Enough processing studied normalised interpretations, and reviews of shallow processing (e.g. Sanford and Sturt, 2002; Ferreira, Ferraro and Bailey, 2002) frequently cite evidence of underspecified interpretations as being a fundamental element of the phenomenon. Early evidence (Fillenbaum, 1971, 1974) demonstrated that readers would normalise unusual parts of a story upon recall in ways that brought them more in line with schematic knowledge. Well-known semantic illusions such as the Moses Illusion (Erikson and Mattson, 1981) and the air crash scenario (Barton and Sanford, 1993) demonstrate that comprehenders would not construct a representation of a sentence that fully reflected its semantic content. Normalised interpretations such as those reported by Garnham and Oakhill (1987) and Ferreira (2003) demonstrate that when semantic cues are biased towards a particular interpretation, comprehenders will frequently choose that interpretation even thought it is not licensed by the grammar, and they will do this with sentences that are syntactically unambiguous. These observations and others like them have given rise to the consideration of a role for heuristic processing in language comprehension. The many instances of a failure to fully, i.e. algorithmically, utilise all available information to arrive at a correct interpretation suggests we may need to posit an architectural component of the language system, which, as in other cognitive domains, reaches interpretations via fast, resource-efficient, and largely reliable heuristics (Ferreira, 2003).

This thesis set out to replicate some important instances of normalisation and to attempt to observe this phenomenon in some new settings. The main question of interest related to the time course of heuristic and algorithmic processing. Granted that heuristics may be responsible for certain mistaken interpretations, what is their relationship to grammatically-generated interpretation? For example, is it the case that syntax would always be used to generate an interpretation, and at the earliest stages, or would we see cases in which semantics-based, heuristic processes were ‘in control’ of interpretation? We also attempted to narrow the list of factors that could account for normalisation, and examined the impact of syntactic complexity and memory constraints. The following sections outline the conclusions reached for each of the topics listed above, and make some suggestions for future work in this area.

1. New evidence for shallow processing (Pragmatic Normalisation)

The first stated aim of this thesis was to replicate and extend findings, both recent and early, of shallow processing in the interpretation of non-ambiguous sentences. The initial impetus came from studies by Ferreira and Stacey (unpublished manuscript; experiment 1) and Ferreira (2003; experiment 1), whose participants exhibited a significant tendency to judge implausible sentences as being plausible when they were cast in the passive voice compared with the active voice, and to make errors on a thematic role judgement task – under the same passive/implausible conditions – indicative of pragmatic normalisation. An earlier study by Garnham and Oakhill (1987) also reported high rates of misinterpretation with implausible elided verb phrases, again indicative of readers failing to interpret using fully specified semantic representations.

The first point to make is that shallow, non-syntactic interpretation is clearly a robust phenomenon – surprisingly so in some cases. With regard to experiments 1-3, which extended the Ferreira studies, the main conclusion is that under certain conditions comprehenders do exhibit a tendency to misinterpret non-ambiguous sentences. While the incorrect interpretations did not form the majority of responses, incorrect responses were significantly increased when the grammatically licensed meaning was implausible. (The second conclusion is that the special difficulty with implausible passive sentences is a replicable phenomenon, but not as robust as some original findings had suggested, and it is not clear exactly what accounts for this particular difficulty – more on this in section 3.) There were initial difficulties in replicating any misinterpretation effect (experiment 1). Experiments 2 and 3, however, reported interpretation accuracy levels as low as 72% and 59%, respectively. Accuracy was significantly lowered when a sentence's correct interpretation was implausible compared with when it was plausible, with up to 40% of interpretations reflecting plausibility rather than being fully informed by grammar. In experiment 4, the extension of the Garnham and Oakhill (1987) study with implausible EVPs, misinterpretation rates actually exceeded those reported in the original study. When an elided verb phrase had an implausible meaning (e.g. that a nurse had been examined by a doctor), readers' interpretation accuracy fell as low as 35%. This was substantially lower than Garnham and Oakhill's 67%, and may have been due to the differences in methodology (segment-based vs. word-by-word presentation of materials). Experiment 6 examined interpretations under more natural reading conditions and still reported surprisingly high rates of plausibility-based misinterpretation. It was assumed that with the free reading conditions afforded by the eyetracking methodology, with participants able to regress and reread at will, interpretations would be consistently accurate relative to the more demanding word-by-word methods of experiments 2-4. Yet accuracy in the implausible condition was only 34%.

Clearly, Ferreira's (2003) point about the assumption of comprehension in psycholinguistic experiments must be taken very seriously indeed (i.e. the assumption that participants will always derive the correct interpretations from the materials they are presented with, as long as they are syntactically unambiguous. Interpretations may justly have been expected to suffer under taxing moving-window methods, but the amazingly low accuracy in experiment 6 gives the lie to the idea that free reading ensures high rates of successful comprehension. Even when given the opportunity to regress and reread, comprehenders may ultimately base a substantial proportion of their interpretation on non-syntactic sources of information, which in many cases here proved to be highly misleading. It certainly counters the idea that a lack of syntactic ambiguity entails correct interpretation. Interpretations were highly successful only in the absence of plausibility cues that conflicted with the correct, syntactic interpretation. While this may describe the majority of materials used in psycholinguistic experiments that assume or rely on comprehension, the prevalence here of faulty interpretations should make plausibility a serious design consideration.

The results in this thesis are entirely consonant with the Good Enough approach to language comprehension, and we can safely venture that in a substantial number of cases, plausibility-based interpretations are judged to be good enough. In fact, as we shall see in the next section, they are more properly judged 'better than' interpretations based on more reliable sources of information.

2. The time course of syntactic vs. heuristic processing

This was the central concern of the thesis, with every experiment except experiment 1 employing measures of online processing time. It must be noted straight away that due to recurrent sparse data issues, the ideal type of analyses was not possible: we were unable to analyse processing data for a sufficiently powerful set of incorrectly answered trials. This analysis could have provided robust evidence either of correct online interpretations coupled with eventual incorrect ones, or an absence of correct (online) interpretation altogether. In terms of models of interpretation, the latter option would have argued for an '*either* syntax *or* heuristics' model, while the former would have argued for a model in which the correct interpretation is computed and then overridden by semantic considerations – a 'syntax-first' model.

However, on the basis of the evidence available it seems reasonable to conclude in favour of the 'syntax first' model. This is contrary to e.g. Nieuwland and Van Berkum's (2005) suggestion of a semantics-first account of their ERP data, and supports standard syntax-first accounts of interpretation (e.g. Frazier and Rayner, 1982; Ferreira and Clifton, 1986; Sturt 2003) – although these syntax-first accounts would not have predicted the kind of misinterpretations reported in this thesis and elsewhere. The robust nature of the observed online anomaly effects, coupled with substantial rates of misinterpretation (reported in experiments 2-6) argue for exactly that account of normalisation. Anomaly detection effects were always visible online (regardless of voice) and timings ranged from immediate (possibly parafoveal) to early, with any later detections, such as those in experiments 2-4, likely being an artefact of the methodology.

In light of the clear online effects caused by the full and correct computation of meaning, and at a very early stage, we could posit a model in which interpretive heuristics act as checking mechanisms, with the secondary role of measuring the interpretive output of the parser and having the power to impose a veto on it. There doesn't seem to be any evidence here for an alternative model in which heuristics only operate if the parser is struggling under time constraints, e.g. a parallel model with a first-past-the-post output, as evidence of correct interpretation appears in the processing record at such an early stage. Another alternative would be that heuristics operate immeasurably quickly and precede the computation of an interpretation based on syntax. But in that that case, a heuristic interpretation would be generated first, followed extremely quickly by a syntactic one, and subsequently chosen in favour of the syntactic interpretation at a frequency dependant on, for example, plausibility. But this is hardly the most parsimonious account and obviously poses severe methodological problems. Our contention remains that the evidence here is best interpreted as supporting the syntax-first, heuristics-second account. There was one result that challenged this account: the parafoveal detection effect for the less-skilled readers in experiment 7. Here, it appeared that these readers had made very early detections, but only when the anomalous phrase was preceded by a semantically non-supportive context. As discussed in chapter 5, we could interpret this finding to mean that less-skilled readers were using the semantic relationships between words to guide their earliest interpretations, with correct interpretation being slightly delayed when those relationships constituted a 'good fit'. However, the same effect was not observed with the skilled readers, nor was it observed in any analysis of experiment 8. While intriguing, the most we can say is that if semantics can exert the primary influence in interpretation, they do not do so with all readers and we do not see it with all anomaly types. (To do it justice in this discussion, this effect could suggest a model in which both heuristic and syntactic interpretations are computed online, and early – even poor readers had correctly interpreted the Good Context anomaly by the critical region). As it stands, this finding is certainly a clear avenue for further research, but cannot by itself offer a strong alternative to the syntax-first account.

The most interesting question remains why the syntactic interpretation should be overridden, given its status as a more or less infallible guide. Ferreira (2003) characterised syntactic representations as 'fragile', and the findings here would seem to support that view, with the contribution of suggesting a likely temporal relation to heuristic representations. In thinking about the relation of algorithmic to heuristic processing, it is very tempting to focus on the algorithmic processing as being time and resource-heavy, unfriendly to the natural-world situations in which cognition really happens. Heuristics, by contrast, are time-efficient and resource-light: 'fast and frugal'. The further temptation might be to assume that the difficulties with algorithmic interpretation lie at the beginning of the process, with the work of actually constructing a syntactic representation and building a semantic one onto it. But the fact is, there is no evidence in the current studies to suggest that algorithmic interpretation is too difficult to perform, even given highly complex linguistic constructions and non-optimal reading conditions. The reliable online disruption caused by our anomalies shows that algorithmic processing happens early, if not immediately, and is not influenced by the semantic content of the constructions (with one possible exception in experiment 7), or the voice in which they are cast. There were no consistent processing differences, remember, between the active and passive implausible sentences, relative to their plausible baselines, despite interpretive differences emerging later (again, there was one possible, though unlikely, exception in the experiment 5 data for the less-skilled readers).

The issue with algorithmic representations seems to be one of confidence rather than resources. If we allow for argument's sake that syntax is always computed, and a meaning representation based on it is always constructed, we can think of the emergence of a non-syntactic interpretation as emerging due to the comprehension system having a lack of confidence in the syntactic one. And on this argument, we would have to say that confidence decreases as syntactic complexity increases, or as the opportunity to regress is removed. Both these observations would suggest that the problematic resources in question are not computational, but memory-based. What is missing in the cases where misinterpretation rates are highest is the opportunity to *confirm* initial interpretations, due either to high memory load or denial of the chance to reread (or both). In short, the failings of syntax-based meaning representations do not lie in the building work, but in their retention. Syntax is reliable, and reliably used, but is not the 'be all and end all' of comprehension. It is one source of information among several and while it may have temporal primacy it will be overruled if other constraints prevail.

This discussion of the role of memory leads us into the next section, a consideration of the conditions for normalisation.

3. Conditions for shallow processing

The main work of the thesis in this regard was replication – a vital approach, given that concentrated work in the field of Good Enough processing can be considered still in its infancy (and much of it has been conducted using only off-line measures). The studies presented here offer both a challenge to recent findings and firmer conclusions on the role of memory, as well as some tentative observations on the influence of task type.

The direct investigation into the role of memory in normalisation, via manipulations of syntactic load and type of methodology, enabled clear conclusions. Experiment 1, with virtually no memory constraints, yielded correct interpretations and no evidence of shallow processing. Experiment 2, with it's more challenging moving window format (no opportunity to reread text) indicated significant rates of normalised interpretations. A comparison with experiment 3 indicated that a higher syntactic load entailed significantly poorer accuracy. And a direct comparison of experiment 4 (moving window, higher memory strain) and experiment 5 (free reading, lower memory strain) indicated that interpretation accuracy was improved under free reading conditions. As discussed above, the job of memory in these cases seems to centre on confirming and building confidence in syntax-based interpretations that may have jarred with pragmatic constraints.

Although direct comparisons were not made, it seems probable that task demands influenced the likelihood that misinterpretation would occur. Experiment 1 can be considered the most successful in terms of interpretations. The task was the only one that explicitly required an evaluation of the overall plausibility. (It was also the only one in which the sentences were still available at the time of performing the task, but, as discussed in chapter 2, this is unlikely to have had a drastic impact.) Experiments 2 and 3, which featured the thematic role judgement task, yielded poorer accuracy, but not as poor as the accuracy in experiments 4 and 6 whose comprehension questions were of a less abstract, more ‘traditional’ nature. Of course, a direct comparison across so many experiments isn’t reasonable, as differences in accuracy may be down to differences in the properties of the materials themselves. But, with that in mind, we might still tentatively conclude that a blanket plausibility judgement is easier than a forced-choice judgement that only tests interpretation, because it encourages a participant to monitor for plausibility throughout. The other tasks, while probing the content of the implausible material, do not explicitly require this judgement on it and so at least allow the possibility that some unsuspected implausibility might slip by unnoticed, or be overridden at the global level in the absence of any imperative to identify it. The relatively higher accuracy in the thematic role judgement task, with its unusual and rather more abstract nature, may have ‘raised the bar’ for what the comprehension system considered good enough and given syntax the final say in a greater number of cases. Again, these ideas are tentative, but do suggest that task type may exert a considerable influence on which sources of information win out at the global level of interpretation.

Experiments 1-3 also tested the proposal that an implausible sentence is more likely to be normalised if cast in the passive voice. Experiment 1 tested plausibility judgements on a set of materials similar to those tested by Ferreira and Stacey, in a questionnaire study judged to be a fair analogue of Ferreira and Stacey’s reading task, and found no differences at all between the interpretations of implausible actives and passives. A clear-cut main effect of plausibility indicated that readers judged implausible sentences as implausible relative to plausible ones, regardless of the voice in which they were phrased. Experiment 2 tested similar materials in a self-paced reading study judged to be a fair analogue of Ferreira's auditory experiments. Results indicated that readers were indeed guilty of normalised interpretations, but only a trend was evident in terms of the proposed active-passive distinction. Only in experiment 3, in which the materials were embedded in a syntactically demanding sentence frame, did we observe a significant difference in the interpretation of implausible actives vs. implausible passives, relative to their implausible baselines. Hence, we can conclude that readers are certainly prone to mistakes arising from global interpretations based on plausibility when reading conditions are non-optimal; but heuristics, in this case based on canonicity of thematic role assignment order, come into play only under conditions of particularly heavy syntactic and memory load.

The finding in experiment 4 that participants fared worse when interpreting implausible elliptical verb phrases (EVPs) with passive antecedent sentences than active antecedent sentences argues again for passive constructions being particularly prone to shallow processing, but perhaps suggests a simpler account than Ferreira's, based simply on more rapid decay of the passive syntactic representation than the active one. The N-V-N strategy, which could explain why comprehenders were poorer at interpreting simple implausible passives (experiment 3), does not have the same explanatory power in experiment 4 as the interpretations it would generate were not semantically compelling.

4. Written vs. spoken language comprehension

An important consideration regarding the Ferreira studies is the fact that the studies here have been reading studies, while Ferreira's active/passive studies involved auditory presentation. Given the primacy of spoken language, any demonstration using auditory methodology would arguably have better claims to validity than studies investigating reading comprehension. It is still reasonable, however, to assume that the findings reported here would hold with auditory comprehension. The best comparison lies in the self-paced reading studies, which presented words one at a time in moving window format, with each word disappearing before the next one could be read, and no opportunity to reread a whole section of text. This surely offers a good visual analogue to spoken language, in which words likewise arrive in strict sequence (although not segmented into single word units) with no opportunity – usually – to hear them repeated. It may be that future work will increasingly employ auditory presentation with ERP recordings as a measure of online processing time, but until the ambiguities in this methodology have been fully resolved, e.g. the controversy over the interpretation of the N400 and P600, and their relation (if any) to different linguistic domains, self-paced reading and eyetracking can provide reliable data and demand serious attention.

5. Normalisation and Contextual fit

The fit of an anomalous word to its context has been established as a factor in whether or not the anomaly will be detected. Recall that in Barton and Sanford's study (1993), participants detected the survivor anomaly on 59% of the time when it featured in the context of an air crash. Experiments 7 & 8 tested anomalies that had a good fit to the context to see whether the goodness of fit would interfere with (i.e. delay or prevent) successful interpretation based on the grammar. The results were generally consistent, with one departure. Anomalies that were both severe and more subtle were successfully detected early online. In experiment 7, skilled readers detected Good Context and Neutral Context anomalies as early as the pre-critical region; less-skilled readers, on the other hand, appeared to detect the Neutral Context anomaly slightly earlier than the Good Context anomaly, consistent with the hypothesis that they were using semantics to guide early interpretations. So we must conclude that goodness-of-fit heuristic processing, if it is active at the earliest stages of interpretation, is only used by less-skilled readers and only with certain anomaly types (in this case, strong animacy violations). Otherwise, this type of heuristic processing is apparently not exempt from the tendency we have already discussed, namely, to construct a meaning based on the grammar at the earliest stage. Interpretations that take fit information into account are therefore most likely be applied at a post-syntactic stage, though individual differences in reading ability may be a salient factor in the timing of their application.

6. Individual differences

Overall, these analyses were inconclusive. Looking firstly at the interpretation data, we see that greater reading skill does not necessarily entail more successful interpretation than poorer reading skill. In two studies out of three, reading ability scores were not reflected in a split of the interpretation data. In experiment 5, skilled readers were more successful, but in experiment 4, which used exactly the same materials, there was no difference between the skill groups, nor was there any difference in experiment 6. Taking the three studies together, this picture does differ from the Hannon and Daneman (2004) study, but only by being inconsistent (their findings would agree with our experiment 5).

While it may seem intuitively surprising that there is no clear reading ability difference, we should bear in mind that what seems to be emerging is a model of normalisation in which misinterpretations increase as *confidence* in the syntactic interpretation decreases. Perhaps the Nelson Denny test is not suitable for differentiating levels of confidence. In other words, perhaps the confidence level at which the 'pass mark' is set for allowing implausible, syntactically licensed interpretations to win out may not correlate with reading ability as measured by the Nelson Denny test. Clearly there is scope for further work, and further testing may benefit by using further reading ability measures such as reading span, in conjunction with the reading comprehension test used here.

Turning to online processing measures, there are similar inconsistencies. But a point worth stressing immediately is that less-skilled readers do detect anomalies online, and in some cases even earlier than skilled readers. However, there is no clear picture emerging of different processing styles in the matter of anomaly detection or recovery, and certainly nothing pointing to inadequate processing in either skill group that might account for differences between them in interpretation success. As we have already discussed, there is some evidence that less-skilled readers make greater use of parafoveal preview information than skilled readers. There is also the fact that the only finding suggestive of early heuristic processing was located only in the data from less-skilled readers. As such, the question of whether goodness-of-fit processing is pre- or post-syntax must be left open for now, with an answer dependant on further individual differences studies. It must also be allowed that, because these parafoveal-on-foveal effects come almost entirely from less-skilled readers, who might be expected to make *less* use of the available information, a sceptical approach may be the best one to take until they are confirmed by more powerful studies. Skilled readers perhaps recover from anomalies a little earlier than less-skilled readers, but not always. (The one clear difference was in experiment 8, in which skilled readers apparently did not recover from an anomaly any faster or slower depending upon the semantic content of the preceding context.) Again, further work would be necessary to uncover any real differences between the two groups in terms of tendency to normalise or the online handling of anomalies.

Future Work

The immediate task would be to work at getting definitive confirmation of the interpretation outlined in section 2, above. While the data here is highly suggestive, only analysing the processing data from a powerful set of normalised trials will confirm that early syntactic interpretations are instantiated but overruled by semantic heuristics. If this is confirmed, we would attempt to pin down the temporal operation of the heuristics themselves: are they applied late online, or offline?. It may be as simple as taking trials in which anomalies have been correctly detected (as indexed by online disruption) and analysing decision times on the interpretation questions. If heuristic interpretations are applied offline, we could expect to see longer decision times on incorrectly answered questions, as presumably it would take longer to apply them than not to apply them. Again, this would require a more powerful data set than is currently available. If, on the other hand, the evidence showed that normalised interpretations are always coupled with a lack of online disruption, and that therefore the ‘either/or’ model was more likely, we would investigate further the conditions responsible for syntactic interpretation not taking place.

We would continue work on the contextual fit heuristic - taking an individual differences approach - as our analyses of experiment 7 are the only ones that really allow for the possibility that (less-skilled) readers recruit this heuristic very early and ahead of syntax. A larger data set would be necessary for more confident conclusions, but there could obviously be implications here for modelling heuristic vs. syntactic interpretation, with the use and timing of heuristics ultimately being explainable, at least in part, by reading ability.

We would investigate more fully the influence of task demands on normalisation, as discussed in section 3. This would involve testing a material set such as that used in experiments 1-3, and manipulating task type, using plausibility judgements, thematic role judgements, naming tasks, and perhaps others. We would also run further memory tests in an attempt to robustly locate the memory hypothesis in between-group interpretation differences.

In closing, this thesis has offered evidence for a model of sentence interpretation in which syntax is the primary information source used to generate an interpretation, and in which strong semantic/pragmatic cues may overrule to provide an interpretation more in line with world knowledge. In doing so it has replicated some normalisation effects, as well as contributing some new ones, and made a contribution to our understanding of the conditions for shallow processing. Finally, it has outlined some experimental work that could provide suitable further testing of the conclusions reached in this research.

References

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

Bates, E., Devescovi, A., & DAmico, S. (1999). Processing complex sentences: A cross-linguistic study. *Language and Cognitive Processes, 14*, 69–123.

Bentin, S., McCarthy, G., & Wood, C. C. (1985). Event-related potentials, lexical decision and semantic priming. *Electroencephalography and Clinical Neurophysiology, 60*, 343-355.

Bohan, J., & Sanford, A. J. (in press). Semantic anomalies at the borderline of consciousness: An eyetracking investigation. *The Quarterly Journal of Experimental Psychology*.

Boland, J. E. (2004). Linking eye movements to sentence comprehension in reading and listening. In M. Carreiras & C. Clifton (Eds.), *The on-line study of sentence comprehension: Eyetracking, ERPs, and beyond*. New York: Psychology Press.

Boland, J. E., & Blodgett, A. (2002). *Eye movements as a measure of syntactic and semantic incongruity in unambiguous sentences*. Unpublished manuscript, University of Michigan at Ann Arbor.

Bornkessel, I.D., Fiebach, C.J. & Friederici, A.D. (2004). On the cost of syntactic ambiguity in human language comprehension: An individual differences approach. *Cognitive Brain Research, 21*, 11-21.

Braze, D., Shankweiler, D., Ni, W., & Palumbo, L., C. (2002). Reader's eye movements distinguish anomalies of form and content. *Journal of Psycholinguistic Research*, 31*(1)*, 25-44.

Bredart, S., & Modolo, K.(1988).Moses strikes again: Focalization effects on a semantic illusion. *Acta Psychologica*, *67*, 135–144.

Brown, J. I., Bennett, J. M., & Hanna, G. (1981). *Nelson-Denny Reading Test*. Chicago: Riverside.

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

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*, 205-238.

Chwilla, D. J., & Kolk, H. H. (2002). Three-step priming in lexical decision. *Memory and Cognition, 30*, 217-225.

Chwilla, D. J., Kolk, H. H., & Mulder, G. (2000). Mediated priming in the lexical decision task.: evidence from event-related potentials and reaction time. *Journal of Memory and Language, 42*, 314-341.

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*, 177–204.

Daneman, M., Lennertz, T., & Hannon, B. (2007). Shallow semantic processing of text: Evidence from eye movements. *Language and Cognitive Processes, 22*, 83-105.

Dowty, D. (1991). Thematic proto-rules and argument selection. *Language, 67*, 547–619.

Duffy, S. A., Henderson, J. M., & Morris, R. K. (1989). Semantic facilitation of lexical access during sentence processing. *Journal of Experimental Psychology: Learning, Memory, Cognition, 15*, 791–801.

Eastwick, E. C., & Phillips, C. Variability in semantic cue effectiveness on syntactic ambiguity resolution: Inducing low-span performance in high-span readers. Architectures and Mechanisms for Language Processing IV. University of Edinburgh, Scotland. September 1999.

Erickson, T. A., & Mattson, M. E. (1981). From words to meanings: A semantic illusion. *Journal of Verbal Learning and Verbal Behavior, 20*, 540–552.

Ferreira, F., & Swets, B. (2005). The production and comprehension of resumptive pronouns in relative clause "island" contexts. In A. Cutler (Ed.), *Twenty-first Century Psycholinguistics: Four Cornerstones* (pp. 263-278). Mahway, NJ: Lawrence Erlbaum Associates.

Ferreira, F. (2003). The misinterpretation of noncanonical sentences. *Cognitive Psychology, 47*, 164-203.

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

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

Ferreira, F. & Stacey, J. L.. The misinterpretation of passive sentences. Unpublished manuscript, Michigan State University.

Fillenbaum, S. (1971). Processing and recall of compatible and incompatible question and answer pairs. *Language and Speech, 14*, 256–265.

Fillenbaum, S. (1974). Pragmatic normalization: Further results for some conjunctive and disjunctive sentences. *Journal of Experimental Psychology, 102*, 574–578.

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

Frisson, S. P., & Pickering, M. J. (1999). The processing of metonymy: Evidence from eye movements. *Journal of Experimental Psychology: Learning, Memory and Cognition, 25*, 1366-1383.

Garnham, A., & Oakhill, J. (1987). Interpreting elliptical verb phrases. *Quarterly Journal of Experimental Psychology, 39A*, 611–625.

Gibson, E. & Thomas, J. (1999). Memory limitations and structural forgetting: The perception of complex ungrammatical sentences as grammatical. *Language and Cognitive Processes, 14*, 225-248.

Gigerenzer, G. (2000). *Adaptive thinking: Rationality in the real world*. New York: Oxford University Press.

Gigerenzer, G., Todd, P.M., & ABC Research Group (1999). *Simple heuristics that make us smart*. New York: Oxford University Press.

Hagoort, P., Brown, C.. & Groothusen, J. (1993). The syntactic positive shift as an ERP measure of syntactic processing. *Language and Cognitive Processes, 8*, 439-483.

Hannon, B., & Daneman, M. (2004). Shallow semantic processing of text: An individual-differences account. *Discourse Processes, 37*, 187–204.

Henderson, J. M., Falk, R., Minut, S., Dyer, F. C., & Mahadevan, S. (2001). Gaze control for face learning and recognition by humans and machines. In T. F. Shipley, & P. J. Kellman (Eds.), *From fragments to objects: Segmentation and grouping in vision* (pp. 463–481). Amsterdam: Elsevier.

Hobbs, J. R., & Schieber, S. M. (1987). An algorithm for generating quantifier scopings. *Computational Linguistics, 13*, 47–63.

Hoeks, J.C.J., Stowe, L. A. & Doedens, G. (2004). Seeing words in context: the interaction of lexical and sentence level information during reading. *Cognitive Brain Research, 11*, 199-212.

Kamas, E. M., Reder, L. M., & Ayers, M. S. (1996). Partial matching in the Moses illusion: Response bias not sensitivity. *Memory and Cognition, 24*, 687-699.

Kennedy, A., Murray, W. S., & Boissiere, C. (2004). Parafoveal pragmatics revisited. *European Journal of Cognitive Psychology, 16*, 128-153.

Kennedy, A., & Pynte, J. (2005). Parafoveal-on-foveal effects in normal reading. *Vision Research, 45*, 153-168.

Kim, A. & Osterhout, L. (2005). The independence of combinatory semantic processing: Evidence from event-related potentials. *Journal of Memory and Language, 52*, 205-225.

Kolk, H. H., & Chwilla, D. J. (2007). Late positivities in unusual situations: A commentary to (a) Kuperberg, Kreher, Sitnikova, Caplan and Holcomb, and (b) Kremmerer, Weber-fox, Price, Zdanczyk and Way. *Brain and Language, 100*, 257-262.

Kreher, D. A., Holcomb, P. J., & Kuperberg, G. (2006). An electrophysiological investigation of indirect semantic priming. *Psychophysiology, 43*, 550-563.

Kuperberg, G. (2007). Neural mechanisms of language comprehension: Challenges to syntax. *Brain Research, 1146*, 23-49.

Kutas, M., and Hillyard, S. A. (1984). Brain potentials during reading reflect word expectancy and semantic association. *Nature, 307*, 161-163

Kutas, M., and Hillyard, S. A., (1980). Reading senseless sentences: Brain potentials reflect semantic incongruity. *Science, 207*, 203-205.

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

Michael, M. & Gordon, P. C. (2003). Differential processing of sentential information: Effects on recovery from the Garden Path. Poster presented at the Annual Architectures and Mechanisms for Language Processing Conference, 2003, University of Glasgow, Scotland.

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

Natsopoulos, D. (1985) A verbal illusion in two languages. *Journal of Psycholinguistic Research, 14*, 385–397

Ni, W., Fodor, J., Crain, S., & Shankweiler, D. (1998). Anomaly detection: Eye movement patterns. *Journal of Psycholinguistic Research*, *27(5)*, 515-539.

Nieuwland, M. S., & Van Berkum, J. J. A. (2005). Testing the limits of the semantic illusion phenomenon: ERPs reveal temporary semantic change deafness in discourse comprehension. *Cognitive Brain Research*, *24*, 691-701.

Nieuwland, M. S., & Van Berkum, J. J. A. (2006). When peanuts fall in love: N400 evidence for the power of discourse. *Journal of Cognitive Neuroscience, 18*, 1098-1111.

Osterhout, L., & Nicol, J. (1999). On the distinctiveness, independence, and time course of the brain responses to syntactic and semantic anomalies. *Language and Cognitive Processes, 14*, 283-317.

Osterhout, L. (1997). On the brain response to syntactic anomalies: manipulations of word class and word position reveal individual differences. *Brain and Language, 59*, 494-522.

Osterhout, L., Holcomb, P.J., & Swinney, D.A. (1994). Brain potentials elicited by garden-path sentences: evidence of the application of verb information during parsing. *Journal of Experimental Psychology, LMC, 20*, 786-803.

Pearlmutter, N. J., Garnsey, S. M., & Bock, K. (1999). Agreement processes in sentence comprehension. *Journal of Memory and language, 41*, 427-456.

Rayner, K., White, S. J., Kambe, G., Miller, B., & Liversedge, S. P. (2003). On the processing of meaning from parafoveal vision during eye fixations in reading. In J. Hyona, R. Radach, & H, Deubel (Eds.), *The Mind's Eye: Cognitive and applied aspects of eye movement research* (pp 213-234). Amsterdam, Elsevier Science.

Rayner, K., Warren, T., Juhasz, B. J., & Liversedge, S. (2004). The effect of plausibility on eye movements in reading. *Journal of Experimental Psychology: Learning, memory and Cognition, 30*, 1290-1301.

Reder, L. M., & Kusbit, G. W. (1991). Locus of the Moses Illusion: Imperfect encoding, retrieval or match? *Journal of Memory and Language, 30*, 385–406.

Rugg, M. D. (1985). The effects of semantic priming and word repetition on event-related potentials. Psychophysiology, 22, 642-647.

Sanford, A. J., Bohan, J., Molle, J., & Leuthold, H. (manuscript in preparation). ERP characteristics of situation-based anomalies at the borderline of awareness: A comparison of reported and missed anomalies.

Sanford, A. J., & Graesser, A. C. (2006). Shallow Processing and Underspecification. *Discourse Processes, 42*, pp. 99-108.

Sanford, J. S., Sanford, A. J., Filik, R., & Molle, J. (2005). Depth of lexical-semantic processing and sentential load. *Journal of Memory and Language, 53*, 378-396.

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

Sanford, A. J., & Garrod, S. C. (1980). Memory and attention in text comprehension: The problem of reference. In R. S. Nickerson (ed), *Attention and Performance, 8*, Hillsdale, NH: Laurence Erlbaum Associates.

Sanford, A. J., Garrod, S. C., & Bell, E. (1979). Aspects of memory dynamics in text comprehension. In M. Gruneberg, P. E. Morris & R. N. Sykes, *Practical Aspects of Memory*, London: Academic Press.

Simons, D.J., & Levin, D.T. (1997). Change blindness. *Trends in Cognitive Sciences, 1*, 261–267.

Sitnikova, T., Holcomb, P., & Kuperberg, G. (unpublished). Two neurocognitive mechanisms of semantic integration during the comprehension of visual real-world events.

Sturt, P. (2003). Semantic Inertia. Poster presented at the Annual CUNY conference on Sentence Processing, 2003.

Sturt, P. (2007). Semantic re-interpretation and garden path recovery. *Cognition, 105*, 477-488.

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.

Townsend, D. J., & Bever, T. G. (2001). *Sentence comprehension: The integration of habits and rules*. Cambridge, MA: MIT Press.

Trask, R. L. (1993). *A dictionary of grammatical terms in linguistics*. New York: Routledge.

Van Berkum, J. J. A., Hagoort, P., & Brown, C. M. (1999). Semantic integration in sentences and discourse: Evidence from the N400. *Journal of Cognitive Neuroscience, 11*, 657-671.

Van Petten, C. (1993). A comparison of lexical and sentence-level context effects in event-related potentials. Special issue: Event-related brain potentials in the study of language. *Language and Cognitive Processes, 8*, 485-531.

Warren, T. (2001). *Understanding the role of referential processing in sentence complexity*. Unpublished doctoral thesis, Massachusetts Institute of Technology, September 2001.

Warren, T., Gibson, E. (2002). The influence of referential processing on sentence complexity. *Cognition, 85*, 79-112.

Wason, P. and Reich, S.S. (1979) A verbal illusion. *Quarterly Journal of Experimental Psychology*, *31*, 591–597.

Appendix A:

Experimental Materials for Experimental 1

1.
Active Plausible
The bird ate the worm very slowly
Active Implausible
The worm ate the bird very slowly
Passive Plausible
The worm was eaten by the bird very slowly
Passive Implausible
The bird was eaten by the worm very slowly

2.
Active Plausible
The soldier protected the child in the battle
Active Implausible
The child protected the soldier in the battle
Passive Plausible
The child was protected by the soldier in the battle
Passive Implausible
The soldier was protected by the child in the battle

3.
Active Plausible
The lawyer sued the builder for one million pounds
Active Implausible
The builder sued the lawyer for one million pounds
Passive Plausible
The builder was sued by the lawyer for one million pounds
Passive Implausible
The lawyer was sued by the builder for one million pounds

4.
Active Plausible
The teacher quizzed the pupil on arithmetic
Active Implausible
The pupil quizzed the teacher on arithmetic
Passive Plausible
The pupil was quizzed by the teacher on arithmetic
Passive Implausible
The teacher was quizzed by the pupil on arithmetic

5.
Active Plausible
The policeman pursued the thief for over an hour
Active Implausible
The thief pursued the policeman for over an hour
Passive Plausible
The thief was pursued by the policeman for over an hour
Passive Implausible
The policeman was pursued by the thief for over an hour

6.
Active Plausible
The waitress served the customer at lunchtime
Active Implausible
The customer served the waitress at lunchtime
Passive Plausible
The customer was served by the waitress at lunchtime
Passive Implausible
The waitress was served by the customer at lunchtime

7.
Active Plausible
The detective questioned the suspect at the crime scene
Active Implausible
The suspect questioned the detective at the crime scene
Passive Plausible
The suspect was questioned by the detective at the crime scene
Passive Implausible
The detective was questioned by the suspect at the crime scene

8.
Active Plausible
The doctor treated the patient in the surgery
Active Implausible
The patient treated the doctor in the surgery
Passive Plausible
The patient was treated by the doctor in the surgery
Passive Implausible
The doctor was treated by the patient in the surgery

9.
Active Plausible
The politician deceived the voter before the election
Active Implausible
The voter deceived the politician before the election
Passive Plausible
The voter was deceived by the politician before the election
Passive Implausible
The politician was deceived by the voter before the election

10.
Active Plausible
The hiker killed the mosquito on the mountain
Active Implausible
The mosquito killed the hiker on the mountain
Passive Plausible
The mosquito was killed by the hiker on the mountain
Passive Implausible
The hiker was killed by the mosquito on the mountain

11.
Active Plausible
The ghost terrified the woman in the haunted house
Active Implausible
The woman terrified the ghost in the haunted house
Passive Plausible
The woman was terrified by the ghost in the haunted house
Passive Implausible
The ghost was terrified by the woman in the haunted house

12.
Active Plausible
The horse kicked the jockey in the stable
Active Implausible
The jockey kicked the horse in the stable
Passive Plausible
The jockey was kicked by the horse in the stable
Passive Implausible
The horse was kicked by the jockey in the stable

13.
Active Plausible
The mugger attacked the pensioner in the busy street
Active Implausible
The pensioner attacked the mugger in the busy street
Passive Plausible
The pensioner was attacked by the mugger in the busy street
Passive Implausible
The mugger was attacked by the pensioner in the busy street

14.
Active Plausible
The widow forgave the murderer at the funeral
Active Implausible
The murderer forgave the widow at the funeral
Passive Plausible
The murderer was forgiven by the widow at the funeral
Passive Implausible
The widow was forgiven by the murderer at the funeral

15.
Active Plausible
The mother abandoned the child in the supermarket
Active Implausible
The child abandoned the mother in the supermarket
Passive Plausible
The child was abandoned by the mother in the supermarket
Passive Implausible
The mother was abandoned by the child in the supermarket

16.
Active Plausible
The teacher praised the pupil for a job well done
Active Implausible
The pupil praised the teacher for a job well done
Passive Plausible
The pupil was praised by the teacher for a job well done
Passive Implausible
The teacher was praised by the pupil for a job well done

17.
Active Plausible
The accountant advised the client on some difficult financial issues
Active Implausible
The client advised the accountant on some difficult financial issues
Passive Plausible
The client was advised by the accountant on some difficult financial issues
Passive Implausible
The accountant was advised by the client on some difficult financial issues

18.
Active Plausible
The boss bullied the trainee every single day
Active Implausible
The trainee bullied the boss every single day
Passive Plausible
The trainee was bullied by the boss every single day
Passive Implausible
The boss was bullied by the trainee every single day

19.
Active Plausible
The boxer punched the referee during the third round
Active Implausible
The referee punched the boxer during the third round
Passive Plausible
The referee was punched by the boxer during the third round
Passive Implausible
The boxer was punched by the referee during the third round

20.
Active Plausible
The spectator encouraged the runner towards the end of the race
Active Implausible
The runner encouraged the spectator towards the end of the race
Passive Plausible
The runner was encouraged by the spectator towards the end of the race
Passive Implausible
The spectator was encouraged by the runner towards the end of the race

21.
Active Plausible
The master whipped the slave in the entrance hall
Active Implausible
The slave whipped the master in the entrance hall
Passive Plausible
The slave was whipped by the master in the entrance hall
Passive Implausible
The master was whipped by the slave in the entrance hall

22.
Active Plausible
The king rewarded the farmer for his loyalty
Active Implausible
The farmer rewarded the king for his loyalty
Passive Plausible
The farmer was rewarded by the king for his loyalty
Passive Implausible
The king was rewarded by the farmer for his loyalty

23.
Active Plausible
The policeman interrogated the robber in the interview room
Active Implausible
The robber interrogated the policeman in the interview room
Passive Plausible
The robber was interrogated by the policeman in the interview room
Passive Implausible
The policeman was interrogated by the robber in the interview room

24.
Active Plausible
The guard released the prisoner on New Year’s Day
Active Implausible
The prisoner released the guard on New Year’s Day
Passive Plausible
The prisoner was released by the guard on New Year’s Day
Passive Implausible
The guard was released by the prisoner on New Year’s Day

25.
Active Plausible
The tailor measured the businessman in the fitting room
Active Implausible
The businessman measured the tailor in the fitting room
Passive Plausible
The businessman was measured by the tailor in the fitting room
Passive Implausible
The tailor was measured by the businessman in the fitting room

26.
Active Plausible
The judge summoned the defendant to appear before the court
Active Implausible
The defendant summoned the judge to appear before the court
Passive Plausible
The defendant was summoned by the judge to appear before the court
Passive Implausible
The judge was summoned by the defendant to appear before the court

27.
Active Plausible
The trainer coached the athlete for four hours every day
Active Implausible
The athlete coached the trainer for four hours every day
Passive Plausible
The athlete was coached by the trainer for four hours every day
Passive Implausible
The trainer was coached by the athlete for four hours every day

28.
Active Plausible
The conman tricked the investor about the sum of money involved
Active Implausible
The investor tricked the conman about the sum of money involved
Passive Plausible
The investor was tricked by the conman about the sum of money involved
Passive Implausible
The conman was tricked by the investor about the sum of money involved

29
Active Plausible
The vandal victimised the neighbour over a period of several months
Active Implausible
The neighbour victimised the vandal over a period of several months
Passive Plausible
The neighbour was victimised by the vandal over a period of several months
Passive Implausible
The vandal was victimised by the neighbour over a period of several months

30.
Active Plausible
The father punished the teenager after the greenhouse window was smashed
Active Implausible
The teenager punished the father after the greenhouse window was smashed
Passive Plausible
The teenager was punished by the father after the greenhouse window was smashed
Passive Implausible
The father was punished by the teenager after the greenhouse window was smashed.

31.
Active Plausible
The professor helped the student with the difficult essay question
Active Implausible
The student helped the professor with the difficult essay question
Passive Plausible
The student was helped by the professor with the difficult essay question
Passive Implausible
The professor was helped by the student with the difficult essay question

32.
Active Plausible
The zookeeper warned the visitor to keep away from the cages
Active Implausible
The visitor warned the zookeeper to keep away from the cages
Passive Plausible
The visitor was warned by the zookeeper to keep away from the cages
Passive Implausible
The zookeeper was warned by the visitor to keep away from the cages

33.
Active Plausible
The pilot asked the stewardess to begin serving dinner
Active Implausible
The stewardess asked the pilot to begin serving dinner
Passive Plausible
The stewardess was asked by the pilot to begin serving dinner
Passive Implausible
The pilot was asked by the stewardess to begin serving dinner

34.
Active Plausible
The bull chased the woman from one end of the field to the other
Active Implausible
The woman chased the bull from one end of the field to the other
Passive Plausible
The woman was chased by the bull from one end of the field to the other

Passive Implausible
The bull was chased by the woman from one end of the field to the other

35.
Active Plausible
The manager sacked the worker following the accident
Active Implausible
The worker sacked the manager following the accident
Passive Plausible
The worker was sacked by the manager following the accident
Passive Implausible
The manager was sacked by the worker following the accident

36
Active Plausible
The firefighter rescued the survivor from the blaze
Active Implausible
The survivor rescued the firefighter from the blaze
Passive Plausible
The survivor was rescued by the firefighter from the blaze
Passive Implausible
The firefighter was rescued by the survivor from the blaze

37
Active Plausible
The comedian entertained the audience in the new comedy club
Active Implausible
The audience entertained the comedian in the new comedy club
Passive Plausible
The audience was entertained by the comedian in the new comedy club
Passive Implausible
The comedian was entertained by the audience in the new comedy club

Appendix B:

Experimental Materials for Experiment 2

1.

Active Plausible
The soldier protected the child in the battle
Active Implausible
The child protected the soldier in the battle
Passive Plausible
The child was protected by the soldier in the battle
Passive Implausible
The soldier was protected by the child in the battle

2.
Active Plausible
The policeman pursued the thief for over an hour
Active Implausible
The thief pursued the policeman for over an hour
Passive Plausible
The thief was pursued by the policeman for over an hour
Passive Implausible
The policeman was pursued by the thief for over an hour

3.
Active Plausible
The waitress served the customer at lunchtime
Active Implausible
The customer served the waitress at lunchtime
Passive Plausible
The customer was served by the waitress at lunchtime
Passive Implausible
The waitress was served by the customer at lunchtime

4.
Active Plausible
The detective questioned the suspect at the crime scene
Active Implausible
The suspect questioned the detective at the crime scene
Passive Plausible
The suspect was questioned by the detective at the crime scene
Passive Implausible
The detective was questioned by the suspect at the crime scene

5.
Active Plausible
The doctor treated the patient in the surgery

Active Implausible
The patient treated the doctor in the surgery
Passive Plausible
The patient was treated by the doctor in the surgery
Passive Implausible
The doctor was treated by the patient in the surgery

6.
Active Plausible
The hiker killed the mosquito on the mountain
Active Implausible
The mosquito killed the hiker on the mountain
Passive Plausible
The mosquito was killed by the hiker on the mountain
Passive Implausible
The hiker was killed by the mosquito on the mountain

7.
Active Plausible
The ghost terrified the woman in the haunted house
Active Implausible
The woman terrified the ghost in the haunted house
Passive Plausible
The woman was terrified by the ghost in the haunted house
Passive Implausible
The ghost was terrified by the woman in the haunted house

8.
Active Plausible
The teacher praised the pupil for a job well done
Active Implausible
The pupil praised the teacher for a job well done
Passive Plausible
The pupil was praised by the teacher for a job well done
Passive Implausible
The teacher was praised by the pupil for a job well done

9.
Active Plausible
The accountant advised the client on some difficult financial issues
Active Implausible
The client advised the accountant on some difficult financial issues
Passive Plausible
The client was advised by the accountant on some difficult financial issues
Passive Implausible
The accountant was advised by the client on some difficult financial issues

10.
Active Plausible
The boss bullied the trainee every single day

Active Implausible
The trainee bullied the boss every single day
Passive Plausible
The trainee was bullied by the boss every single day
Passive Implausible
The boss was bullied by the trainee every single day

11.
Active Plausible
The boxer punched the referee during the third
Active Implausible
The referee punched the boxer during the third round
Passive Plausible
The referee was punched by the boxer during the third round
Passive Implausible
The boxer was punched by the referee during the third round

12.
Active Plausible
The spectator encouraged the runner towards the end of the race
Active Implausible
The runner encouraged the spectator towards the end of the race
Passive Plausible
The runner was encouraged by the spectator towards the end of the race
Passive Implausible
The spectator was encouraged by the runner towards the end of the race

13.
Active Plausible
The master whipped the slave in the entrance hall
Active Implausible
The slave whipped the master in the entrance hall
Passive Plausible
The slave was whipped by the master in the entrance hall
Passive Implausible
The master was whipped by the slave in the entrance hall

14.
Active Plausible
The policeman interrogated the robber in the interview room
Active Implausible
The robber interrogated the policeman in the interview room
Passive Plausible
The robber was interrogated by the policeman in the interview room
Passive Implausible
The policeman was interrogated by the robber in the interview room

15.
Active Plausible
The guard released the prisoner on New Year’s Day

Active Implausible
The prisoner released the guard on New Year's Day
Passive Plausible
The prisoner was released by the guard on New Year's Day
Passive Implausible
The guard was released by the prisoner on New Year's Day

16.
Active Plausible
The tailor measured the businessman in the fitting room
Active Implausible
The businessman measured the tailor in the fitting room
Passive Plausible
The businessman was measured by the tailor in the fitting room
Passive Implausible
The tailor was measured by the businessman in the fitting room

17.
Active Plausible
The judge summoned the defendant to appear before the court
Active Implausible
The defendant summoned the judge to appear before the court
Passive Plausible
The defendant was summoned by the judge to appear before the court
Passive Implausible
The judge was summoned by the defendant to appear before the court

18.
Active Plausible
The father punished the teenager after the greenhouse window was smashed
Active Implausible
The teenager punished the father after the greenhouse window was smashed
Passive Plausible
The teenager was punished by the father after the greenhouse window was smashed
Passive Implausible
The father was punished by the teenager after the greenhouse window was smashed

19.
Active Plausible
The professor helped the student with the difficult essay question
Active Implausible
The student helped the professor with the difficult essay question
Passive Plausible
The student was helped by the professor with the difficult essay question
Passive Implausible
The professor was helped by the student with the difficult essay question

20.

Active Plausible
The zookeeper warned the visitor to keep away from the cages
Active Implausible
The visitor warned the zookeeper to keep away from the cages
Passive Plausible
The visitor was warned by the zookeeper to keep away from the cages
Passive Implausible
The zookeeper was warned by the visitor to keep away from the cages

21.
Active Plausible
The pilot asked the stewardess to begin serving dinner
Active Implausible
The stewardess asked the pilot to begin serving dinner
Passive Plausible
The stewardess was asked by the pilot to begin serving dinner
Passive Implausible
The pilot was asked by the stewardess to begin serving dinner

22.
Active Plausible
The manager sacked the worker following the accident
Active Implausible
The worker sacked the manager following the accident
Passive Plausible
The worker was sacked by the manager following the accident
Passive Implausible
The manager was sacked by the worker following the accident

23.
Active Plausible
The politician deceived the voter before the election
Active Implausible
The voter deceived the politician before the election
Passive Plausible
The voter was deceived by the politician before the election
Passive Implausible
The politician was deceived by the voter before the election

24.
Active Plausible
The king rewarded the farmer for his loyalty
Active Implausible
The farmer rewarded the king for his loyalty
Passive Plausible
The farmer was rewarded by the king for his loyalty
Passive Implausible
The king was rewarded by the farmer for his loyalty

Appendix C:

Experimental Materials for Experiment 3

1.
Active Plausible
The commission said that the reports stating that the soldier protected the child in the battle would be taken into consideration
Active Implausible
The commission said that the reports stating that the child protected the soldier in the battle would be taken into consideration
Passive Plausible
The commission said that the reports stating that the child was protected by the soldier in the battle would be taken into consideration
Passive Implausible
The commission said that the reports stating that the soldier was protected by the child in the battle would be taken into consideration

2.
Active Plausible
The jury heard that the testimony revealing that the policeman pursued the thief for over an hour should not influence their decision
Active Implausible
The jury heard that the testimony revealing that the thief pursued the policeman for over an hour should not influence their decision
Passive Plausible
The jury heard that the testimony revealing that the thief was pursued by the policeman for over an hour should not influence their decision
Passive Implausible
The jury heard that the testimony revealing that the policeman was pursued by the thief for over an hour should not influence their decision

3.
Active Plausible
The restaurant owner thought that the receipt showing that the waitress served the customer at lunchtime should be given to the chef
Active Implausible
The restaurant owner thought that the receipt showing that the customer served the waitress at lunchtime should be given to the chef
Passive Plausible
The restaurant owner thought that the receipt showing that the customer was served by the waitress at lunchtime should be given to the chef
Passive Implausible
The restaurant owner thought that the receipt showing that the waitress was served by the customer at lunchtime should be given to the chef

4.

Active Plausible

The witness complained that the statement revealing that the detective questioned the suspect at the crime scene had gone missing

Active Implausible

The witness complained that the statement revealing that the suspect questioned the detective at the crime scene had gone missing

Passive Plausible

The witness complained that the statement revealing that the suspect was questioned by the detective at the crime scene had gone missing

Passive Implausible

The witness complained that the statement revealing that the detective was questioned by the suspect at the crime scene had gone missing

5.

Active Plausible

The nurse forgot that the letter saying that the doctor treated the patient in the surgery had already been posted out

Active Implausible

The nurse forgot that the letter saying that the patient treated the doctor in the surgery had already been posted out

Passive Plausible

The nurse forgot that the letter saying that the patient was treated by the doctor in the surgery had already been posted out

Passive Implausible

The nurse forgot that the letter saying that the doctor was treated by the patient in the surgery had already been posted out

6.

Active Plausible

The camper knew that the article telling that the hiker killed the mosquito on the mountain would come in useful

Active Implausible

The camper knew that the article telling that the mosquito killed the hiker on the mountain would come in useful

Passive Plausible

The camper knew that the article telling that the mosquito was killed by the hiker on the mountain would come in useful

Passive Implausible

The camper knew that the article telling that the hiker was killed by the mosquito on the mountain would come in useful

7.

Active Plausible

The owner said that the story recounting that the ghost terrified the woman in the haunted house had attracted many tourists

Active Implausible

The owner said that the story recounting that the woman terrified the ghost in the haunted house had attracted many tourists

Passive Plausible

The owner said that the story recounting that the woman was terrified by the ghost in the haunted house had attracted many tourists
Passive Implausible
The owner said that the story recounting that the ghost was terrified by the woman in the haunted house had attracted many tourists

8.
Active Plausible
The class heard that the article stating that the teacher praised the pupil for a job well done would be published soon
Active Implausible
The class heard that the article stating that the pupil praised the teacher for a job well done would be published soon
Passive Plausible
The class heard that the article stating that the pupil was praised by the teacher for a job well done would be published soon
Passive Implausible
The class heard that the article stating that the teacher was praised by the pupil for a job well done would be published soon

9.
Active Plausible
The secretary knew that the letter indicating that the accountant advised the client on some difficult financial issues was on the office desk
Active Implausible
The secretary knew that the letter indicating that the client advised the accountant on some difficult financial issues was on the office desk
Passive Plausible
The secretary knew that the letter indicating that the client was advised by the accountant on some difficult financial issues was on the office desk
Passive Implausible
The secretary knew that the letter indicating that the accountant was advised by the client on some difficult financial issues was on the office desk

10.
Active Plausible
The company announced that the memo stating that the boss bullied the trainee every single day had been covered up
Active Implausible
The company announced that the memo stating that the trainee bullied the boss every single day had been covered up
Passive Plausible
The company announced that the memo stating that the trainee was bullied by the boss every single day had been covered up
Passive Implausible
The company announced that the memo stating that the boss was bullied by the trainee every single day had been covered up

11.
Active Plausible

The commentator thought that the video showing that the boxer punched the referee during the third round was in his bag
Active Implausible
The commentator thought that the video showing that the referee punched the boxer during the third round was in his bag
Passive Plausible
The commentator thought that the video showing that the referee was punched by the boxer during the third round was in his bag
Passive Implausible
The commentator thought that the video showing that the boxer was punched by the referee during the third round was in his bag

12.
Active Plausible
The editor thought that the picture showing that the spectator encouraged the runner towards the end of the race was front page material
Active Implausible
The editor thought that the picture showing that the runner encouraged the spectator towards the end of the race was front page material
Passive Plausible
The editor thought that the picture showing that the runner was encouraged by the spectator towards the end of the race was front page material
Passive Implausible
The editor thought that the picture showing that the spectator was encouraged by the runner towards the end of the race was front page material

13.
Active Plausible
The historian admitted that the document indicating that the master whipped the slave in the entrance hall was written by him
Active Implausible
The historian admitted that the document indicating that the slave whipped the master in the entrance hall was written by him
Passive Plausible
The historian admitted that the document indicating that the slave was whipped by the master in the entrance hall was written by him
Passive Implausible
The historian admitted that the document indicating that the master was whipped by the slave in the entrance hall was written by him

14.
Active Plausible
The enquiry heard that the tape proving that the policeman interrogated the robber in the interview room would settle the matter
Active Implausible
The enquiry heard that the tape proving that the robber interrogated the policeman in the interview room was to be submitted as evidence
Passive Plausible
The enquiry heard that the tape proving that the robber was interrogated by the policeman in the interview room would settle the matter

Passive Implausible
The enquiry heard that the tape proving that the policeman was interrogated by the robber in the interview room would settle the matter

15.
Active Plausible
The family knew that the story claiming that the guard released the prisoner on New Year's Day was probably unreliable
Active Implausible
The family knew that the story claiming that the prisoner released the guard on New Year's Day was probably unreliable
Passive Plausible
The family knew that the story claiming that the prisoner was released by the guard on New Year's Day was probably unreliable
Passive Implausible
The family knew that the story claiming that the guard was released by the prisoner on New Year's Day was probably unreliable

16.
Active Plausible
The assistant knew that the note saying that the tailor measured the businessman in the fitting room was in her drawer
Active Implausible
The assistant knew that the note saying that the businessman measured the tailor in the fitting room was in her drawer
Passive Plausible
The assistant knew that the note saying that the businessman was measured by the tailor in the fitting room was in her drawer
Passive Implausible
The assistant knew that the note saying that the tailor was measured by the businessman in the fitting room was in her drawer

17.
Active Plausible
The BBC stated that the allegations denying that the judge summoned the defendant to appear before the court had now been withdrawn
Active Implausible
The BBC stated that the allegations denying that the defendant summoned the judge to appear before the court had now been withdrawn
Passive Plausible
The BBC stated that the allegations denying that the defendant was summoned by the judge to appear before the court had now been withdrawn
Passive Implausible
The BBC stated that the allegations denying that the judge was summoned by the defendant to appear before the court had now been withdrawn

18.
Active Plausible
The mother decided that the photograph showing that the father punished the teenager after the greenhouse window was smashed should be put away

Active Implausible
The mother decided that the photograph showing that the teenager punished the father after the greenhouse window was smashed should be put away
Passive Plausible
The mother decided that the photograph showing that the teenager was punished by the father after the greenhouse window was smashed should be put away
Passive Implausible
The mother decided that the photograph showing that the father was punished by the teenager after the greenhouse window was smashed should be put away

19.
Active Plausible
The results showed that the rumour suggesting that the professor helped the student with the difficult essay question was probably true
Active Implausible
The results showed that the rumour suggesting that the student helped the professor with the difficult essay question was probably true
Passive Plausible
The results showed that the rumour suggesting that the student was helped by the professor with the difficult essay question was probably true
Passive Implausible
The results showed that the rumour suggesting that the professor was helped by the student with the difficult essay question was probably true

20.
Active Plausible
The vet remembered that the report confirming that the zookeeper warned the visitor to keep away from the cages was in his possession
Active Implausible
The vet remembered that the report confirming that the visitor warned the zookeeper to keep away from the cages was in his possession
Passive Plausible
The vet remembered that the report confirming that the visitor was warned by the zookeeper to keep away from the cages was in his possession
Passive Implausible
The vet remembered that the report confirming that the zookeeper was warned by the visitor to keep away from the cages was in his possession

21.
Active Plausible
The passenger saw that the note showing that the pilot asked the stewardess to begin serving dinner was on the floor
Active Implausible
The passenger saw that the note showing that the stewardess asked the pilot to begin serving dinner was on the floor
Passive Plausible
The passenger saw that the note showing that the stewardess was asked by the pilot to begin serving dinner was on the floor
Passive Implausible
The passenger saw that the note showing that the pilot was asked by the

stewardess to begin serving dinner was on the floor

22.
Active Plausible
The trainee heard that the news revealing that the manager sacked the worker following the accident had spread through the company
Active Implausible
The trainee heard that the news revealing that the worker sacked the manager following the accident had spread through the company
Passive Plausible
The trainee heard that the news revealing that the worker was sacked by the manager following the accident had spread through the company
Passive Implausible
The trainee heard that the news revealing that the manager was sacked by the worker following the accident had spread through the company

23.
Active Plausible
The journalist said that the story claiming that the politician deceived the voter before the election was totally unfounded
Active Implausible
The journalist said that the story claiming that the voter deceived the politician before the election was totally unfounded
Passive Plausible
The journalist said that the story claiming that the voter was deceived by the politician before the election was totally unfounded
Passive Implausible
The journalist said that the story claiming that the politician was deceived by the voter before the election was totally unfounded

24.
Active Plausible
The records revealed that the story suggesting that the king rewarded the farmer for his loyalty was actually a myth
Active Implausible
The records revealed that the story suggesting that the farmer rewarded the king for his loyalty was actually a myth
Passive Plausible
The records revealed that the story suggesting that the farmer was rewarded by the king for his loyalty was actually a myth
Passive Implausible
The records revealed that the story suggesting that the king was rewarded by the farmer for his loyalty was actually a myth

Appendix D:

Experimental Materials for Experiments 4 and 5

(Slash marks ('/') indicate eyetracking region boundaries in Experiment 5)

1.

Active Plausible
The assistant/ had been serving/ the woman/ at the customer service desk./ The manager/ had too/ and the/ problem was/ resolved./
Active Implausible
The assistant/ had been serving/ the woman/ at the customer service desk./ The customer/ had too/ and the/ problem was/ resolved./
Passive Plausible
The woman/ had been served by/ the assistant/ at the customer service desk./ The customer/ had too/ and the/ problem was/ resolved./
Passive Implausible
The woman/ had been served by/ the assistant/ at the customer service desk./ The manager/ had too/ and the/ problem was/ resolved./

2.
Active Plausible
The mugger/ had been frightening/ the old woman/ in the park yesterday morning./ The thug/ had too/ according/ to the news/ report./
Active Implausible
The mugger/ had been frightening/ the old woman/ in the park yesterday morning./ The jogger/ had too/ according/ to the news/ report./
Passive Implausible
The old woman/ had been frightened by/ the mugger/ in the park yesterday morning./ The jogger/ had too/ according/ to the news/ report./
Passive Implausible
The old woman/ had been frightened by/ the mugger/ in the park yesterday morning./ The thug/ had too/ according/ to the news/ report./

3.
Active Plausible
The obsessed fan/ had been following/ the beautiful model/ for several weeks./ The stalker/ had too/ and the/ experience was/ terrifying./
Active Implausible
The obsessed fan/ had been following/ the beautiful model/ for several weeks./ The actress/ had too/ and the/ experience was/ terrifying./
Passive Implausible
The beautiful model/ had been followed by/ the obsessed fan/ for several weeks./ The actress/ had too/ and the/ experience was/ terrifying./
Passive Implausible
The beautiful model/ had been followed by/ the obsessed fan/ for several weeks./ The stalker/ had too/ and the/ experience was/ terrifying./

4.

Active Plausible

The psychologist/ had been counselling/ the victim/ after the serious railway accident./ The psychiatrist/ had too/ according/ to the medical/ report./

Active Implausible

The psychologist/ had been counselling/ the victim/after the serious railway accident./ The witness/ had too/ according/ to the medical/ report./

Passive Implausible

The victim/ had been counselled by/ the psychologist/ after the serious railway accident./ The witness/ had too/ according/ to the medical/ report./

Passive Implausible

The victim/ had been counselled by/ the psychologist/ after the serious railway accident./ The psychiatrist/ had too/ according/ to the medical/ report./

5.

Active Plausible

The officer/ had been interrogating/ the criminal/ at the local police station./ The detective/ had too/ and the/ investigation looked/ promising./

Active Implausible

The officer/ had been interrogating/ the criminal/ at the local police station./ The robber/ had too/ and the/ investigation looked/ promising./

Passive Implausible

The criminal/ had been interrogated by/ the officer/ at the local police station./ The robber/ had too/ and the/ investigation looked/ promising./

Passive Implausible

The criminal/ had been interrogated by/ the officer/ at the local police station./ The detective/ had too/ and the/ investigation looked/ promising./

6.

Active Plausible

The artist/ had been painting/ the queen/ for the new official portraits./ The painter/ had too/ and the/ exhibition was/ popular./

Active Implausible

The artist/ had been painting/ the queen/ for the new official portraits./ The princess/ had too/ and the/ exhibition was/ popular./

Passive Implausible

The queen/ had been painted by/ the artist/ for the new official portraits./ The princess/ had too/ and the/ exhibition was/ popular./

Passive Implausible

The queen/ had been painted by/ the artist/ for the new official portraits./ The painter/ had too/ and the/ exhibition was/ popular./

7.

Active Plausible

The waiter/ had been serving/ the politician/ in the newly opened restaurant./ The waitress/ had too/ and the/ food was/ delicious./

Active Implausible

The waiter/ had been serving/ the politician/ in the newly opened restaurant./ The actor/ had too/ and the/ food was/ delicious./

Passive Implausible

The politician/ had been served by/ the waiter/ in the newly opened restaurant./ The actor/ had too/ and the/ food was/ delicious./
Passive Implausible
The politician/ had been served by/ the waiter/ in the newly opened restaurant./ The waitress/ had too/ and the/ food was/ delicious./

8.
Active Plausible
The guard/ had been restraining/ the prisoner/ following the rioting in C-Block./ The warden/ had too/ and the/ incident passed/ quickly./
Active Implausible
The guard/ had been restraining/ the prisoner/ following the rioting in C-Block./ The ringleader/ had too/ and the/ incident passed/ quickly./
Passive Implausible
The prisoner/ had been restrained by/ the guard/ following the rioting in C-Block./ The ringleader/ had too/ and the/ incident passed/ quickly./
Passive Implausible
The prisoner/ had been restrained by/ the guard/ following the rioting in C-Block./ The warden/ had too/ and the/ incident passed/ quickly./

9.
Active Plausible
The lawyer/ had been advising/ the witness/ on answering any difficult questions./ The judge/ had too/ as the/ trial was/ important./
Active Implausible
The lawyer/ had been advising/ the witness/ on answering any difficult questions./ The defendant/ had too/ as the/ trial was/ important./
Passive Implausible
The witness/ had been advised by/ the lawyer/ on answering any difficult questions./ The defendant/ had too/ as the/ trial was/ important./
Passive Implausible
The witness/ had been advised by/ the lawyer/ on answering any difficult questions./ The judge/ had too/ as the/ trial was/ important./

10.
Active Plausible
The builder/ had been telling/ the landlady/ that the materials were expensive./ The joiner/ had too/ and provided/ a detailed/ invoice./
Active Implausible
The builder/ had been telling/ the landlady/ that the materials were expensive./ The tenant/ had too/ and provided/ a detailed/ invoice./
Passive Implausible
The landlady/ had been told by/ the builder/ that the materials were expensive./ The tenant/ had too/ and provided/ a detailed/ invoice./
Passive Implausible
The landlady/ had been told by/ the builder/ that the materials were expensive./ The joiner/ had too/ and provided/ a detailed/ invoice./

11.
Active Plausible

The midwife/ had been caring for/ the mother/ during the rather difficult birth./ The doctor/ had too/ and everything/ turned out/ okay./
Active Implausible
The midwife/ had been caring for/ the mother/ during the rather difficult birth./ The baby/ had too/ and everything/ turned out/ okay./
Passive Implausible
The mother/ had been cared for by/ the midwife/ during the rather difficult birth./ The baby/ had too/ and everything/ turned out/ okay./
Passive Implausible
The mother/ had been cared for by/ the midwife/ during the rather difficult birth./ The doctor/ had too/ and everything/ turned out/ okay./

12.
Active Plausible
The journalist/ had been questioning/ the historian/ at the museum opening event./ The reporter/ had too/ and the/ interview was/ televised./
Active Implausible
The journalist/ had been questioning/ the historian/ at the museum opening event./ The caretaker/ had too/ and the/ interview was/ televised./
Passive Implausible
The historian/ had been questioned by/ the journalist/ at the museum opening event./ The caretaker/ had too/ and the/ interview was/ televised./
Passive Implausible
The historian/ had been questioned by/ the journalist/ at the museum opening event./ The reporter/ had too/ and the/ interview was/ televised./

13.
Active Plausible
The lecturer/ had been confusing/ the student/ during the basic science demonstrations./ The professor/ had too/ at the/ university open/ day./
Active Implausible
The lecturer/ had been confusing/ the student/ during the basic science demonstrations./ The schoolboy/ had too/ at the/ university open/ day./
Passive Implausible
The student/ had been confused by/ the lecturer/ during the basic science demonstrations./ The schoolboy/ had too/ at the/ university open/ day./
Passive Implausible
The student/ had been confused by/ the lecturer/ during the basic science demonstrations./ The professor/ had too/ at the/ university open/ day./

14.
Active Plausible
The headmaster/ had been questioning/ the pupil/ during the school inspection day./ The inspector/ had too/ and many/ questions were/ asked./
Active Implausible
The headmaster/ had been questioning/ the pupil/ during the school inspection day./ The janitor/ had too/ and many/ questions were/ asked./
Passive Implausible
The pupil/ had been questioned by/ the headmaster/ during the school inspection day./ The janitor/ had too/ and many/ questions were/ asked./

Passive Implausible
The pupil/ had been questioned by/ the headmaster/ during the school inspection day./ The inspector/ had too/ and many/ questions were/ asked./

15.
Active Plausible
The chemist/ had been advising/ the woman/ on treating minor skin problems./ The doctor/ had too/ and the/ advice was/ helpful./
Active Implausible
The chemist/ had been advising/ the woman/ on treating minor skin problems./ The celebrity/ had too/ and the/ advice was/ helpful./
Passive Implausible
The woman/ had been advised by/ the chemist/ on treating minor skin problems./ The celebrity/ had too/ and the/ advice was/ helpful./
Passive Implausible
The woman/ had been advised by/ the chemist/ on treating minor skin problems./ The doctor/ had too/ and the/ advice was/ helpful./

16.
Active Plausible
The fireman/ had been instructing/ the shop owner/ on safety and security issues./ The policeman/ had too/ at the/ local community/ centre./
Active Implausible
The fireman/ had been instructing/ the shop owner/ on safety and security issues./ The homeowner/ had too/ at the/ local community/ centre./
Passive Implausible
The shop owner/ had been instructed by/ the fireman/ on safety and security issues./ The homeowner/ had too/ at the/ local community/ centre./
Passive Implausible
The shop owner/ had been instructed by/ the fireman/ on safety and security issues./ The policeman/ had too/ at the/ local community/ centre./

17.
Active Plausible
The lawyer/ had been criticising/ the drug dealer/ for his immoral money making./ The judge/ had too/ according to/ the trial/ transcripts./
Active Implausible
The lawyer/ had been criticising/ the drug dealer/ for his immoral money making./ The prostitute/ had too/ according to/ the trial/ transcripts./
Passive Implausible
The drug dealer/ had been criticised by/ the lawyer/ for his immoral money making./ The prostitute/ had too/ according to/ the trial/ transcripts./
Passive Implausible
The drug dealer/ had been criticised by/ the lawyer/ for his immoral money making./ The judge/ had too/ according to/ the trial/ transcripts./

18.
Active Plausible
The beautician/ had been treating/ the bride/ before the big wedding ceremony./ The hairdresser/ had too/ and everyone/ was very/ excited./

Active Implausible
The beautician/ had been treating/ the bride/ before the big wedding ceremony./ The bridesmaid/ had too/ and everyone/ was very/ excited./
Passive Implausible
The bride/ had been treated by/ the beautician/ before the big wedding ceremony./ The bridesmaid/ had too/ and everyone/ was very/ excited./
Passive Implausible
The bride/ had been treated by/ the beautician/ before the big wedding ceremony./ The hairdresser/ had too/ and everyone/ was very/ excited./

19.
Active Plausible
The king/ had been rewarding/ the general/ for bravery during the battle./ The queen/ had too/ in a/ very grand/ ceremony./
Active Implausible
The king/ had been rewarding/ the general/ for bravery during the battle./ The soldier/ had too/ in a/ very grand/ ceremony./
Passive Implausible
The general/ had been rewarded by/ the king/ for bravery during the battle./ The soldier/ had too/ in a/ very grand/ ceremony./
Passive Implausible
The general/ had been rewarded by/ the king/ for bravery during the battle./ The queen/ had too/ in a/ very grand/ ceremony./

20.
Active Plausible
The activist/ had been heckling/ the prime minister/ during his speech on pollution./ The protestor/ had too/ according to/ the news/ report./
Active Implausible
The activist/ had been heckling/ the prime minister/ during his speech on pollution./ The president/ had too/ according to/ the news/ report./
Passive Implausible
The prime minister/ had been heckled by/ the activist/ during his speech on pollution./ The president/ had too/ according to/ the news/ report./
Passive Implausible
The prime minister/ had been heckled by/ the activist/ during his speech on pollution./ The protestor/ had too/ according to/ the news/ report./

21.
Active Plausible
The clown/ had been entertaining/ the child/ as the parade passed by./ The juggler/ had too/ and it/ was quite/ spectacular./
Active Implausible
The clown/ had been entertaining/ the child/ as the parade passed by./ The parent/ had too/ and it/ was quite/ spectacular./
Passive Implausible
The child/ had been entertained by/ the clown/ as the parade passed by./ The parent/ had too/ and it/ was quite/ spectacular./
Passive Implausible
The child/ had been entertained by/ the clown/ as the parade passed

by./ The juggler/ had too/ and it/ was quite/ spectacular./

22.
Active Plausible
The lifeguard/ had been rescuing/ the surfer/ during the terrible thunder storm./ The coastguard/ had too/ because the/ conditions were/ treacherous./
Active Implausible
The lifeguard/ had been rescuing/ the surfer/ during the terrible thunder storm./ The swimmer/ had too/ because the/ conditions were/ treacherous./
Passive Implausible
The surfer/ had been rescued by/ the lifeguard/ during the terrible thunder storm./ The swimmer/ had too/ because the/ conditions were/ treacherous./
Passive Implausible
The surfer/ had been rescued by/ the lifeguard/ during the terrible thunder storm./ The coastguard/ had too/ because the/ conditions were/ treacherous./

23.
Active Plausible
The postman/ had been waking/ the baby/ when he delivered each morning./ The milkman/ had too,/ the residents/ committee was/ told./
Active Implausible
The postman/ had been waking/ the baby/ when he delivered each morning./ The neighbour/ had too,/ the residents/ committee was/ told./
Passive Implausible
The baby/ had been woken by/ the postman/ when he delivered each morning./ The neighbour/ had too,/ the residents/ committee/ was told./
Passive Implausible
The baby/ had been woken by/ the postman/ when he delivered each morning./ The milkman/ had too,/ the residents/ committee was/ told./

24.
Active Plausible
The treasurer/ had been telling/ the club member/ about the new membership rules./ The chairman/ had too/ during the/ annual business/ meeting./
Active Implausible
The treasurer/ had been telling/ the club member/ about the new membership rules./ The visitor/ had too/ during the/ annual business/ meeting./
Passive Implausible
The club member/ had been told by/ the treasurer/ about the new membership rules./ The visitor/ had too/ during the/ annual business/ meeting./
Passive Implausible
The club member/ had been told by/ the treasurer/ about the new membership rules./ The chairman/ had too/ during the/ annual business/ meeting./

25.
Active Plausible
The supporter/ had been booing at/ the footballer/ after the poor first half./ The spectator/ had too/ and everyone/ was very/ frustrated./
Active Implausible
The supporter/ had been booing at/ the footballer/ after the poor first

half./ The referee/ had too/ and everyone/ was very/ frustrated./
Passive Implausible
The footballer/ had been booed at by/ the supporter/ after the poor first half./ The referee/ had too/ and everyone/ was very/ frustrated./
Passive Implausible
The footballer/ had been booed at by/ the supporter/ after the poor first half./ The spectator/ had too/ and everyone/ was very/ frustrated./

26.
Active Plausible
The stewardess/ had been telling/ the passengers/ to prepare for some turbulence./ The pilot/ had too/ during a/ long intercom/ announcement./
Active Implausible
The stewardess/ had been telling/ the passengers/ to prepare for some turbulence./ The child/ had too/ during a/ long intercom/ announcement./
Passive Implausible
The passengers/ had been told by/ the stewardess/ to prepare for some turbulence./ The child/ had too/ during a/ long intercom/ announcement./
Passive Implausible
The passengers/ had been told by/ the stewardess/ to prepare for some turbulence./ The pilot/ had too/ during a/ long intercom/ announcement./

27.
Active Plausible
The council/ had been cautioning/ the pub owner/ about the recent drunken behaviour./ The police/ had too/ and things/ had quietened/ down./
Active Implausible
The council/ had been cautioning/ the pub owner/ about the recent drunken behaviour./ The drinker/ had too/ and things/ had quietened/ down./
Passive Implausible
The pub owner/ had been cautioned by/ the council/ about the recent drunken behaviour./ The drinker/ had too/ and things/ had quietened/ down./
Passive Implausible
The pub owner/ had been cautioned by/ the council/ about the recent drunken behaviour./ The police/ had too/ and things/ had quietened/ down./

28.
Active Plausible
The vicar/ had been blessing/ the bride/ at the beautiful wedding ceremony./ The bishop/ had too/ and people/ were quite/ emotional./
Active Implausible
The vicar/ had been blessing/ the bride/ at the beautiful wedding ceremony./ The groom/ had too/ and people/ were quite/ emotional./
Passive Implausible
The bride/ had been blessed by/ the vicar/ at the beautiful wedding ceremony./ The groom/ had too/ and people/ were quite/ emotional./
Passive Implausible
The bride/ had been blessed by/ the vicar/ at the beautiful wedding ceremony./ The bishop/ had too/ and people/ were quite/ emotional./

29.
Active Plausible
The tutor/ had been disciplining/ the postgraduate/ for the poor research report./ The professor/ had too,/ at the/ weekly laboratory/ meeting./
Active Implausible
The tutor/ had been disciplining/ the postgraduate/ for the poor research report./ The undergraduate/ had too/ at the/ weekly laboratory/ meeting./
Passive Implausible
The postgraduate/ had been disciplined by/ the tutor/ for the poor research report./ The undergraduate/ had too/ at the/ weekly laboratory/ meeting./
Passive Implausible
The postgraduate/ had been disciplined by/ the tutor/ for the poor research report./ The professor/ had too/ at the/ weekly laboratory/ meeting./

30.
Active Plausible
The duke/ had been reprimanding/ the butler/ after the dinner party disaster./ The duchess/ had too/ but the/ damage was/ done./
Active Implausible
The duke/ had been reprimanding/ the butler/ after the dinner party disaster./ The servant/ had too/ but the/ damage was/ done./
Passive Implausible
The butler/ had been reprimanded by/ the duke/ after the dinner party./ The servant/ had too/ but the/ damage was/ done./
Passive Implausible
The butler/ had been reprimanded by/ the duke/ after the dinner party./ The duchess/ had too/ but the/ damage was/ done./

31.
Active Plausible
The traffic warden/ had been fining/ the taxi driver/ for ignoring the parking laws./ The policeman/ had too/ and the/ fines were/ substantial./
Active Implausible
The traffic warden/ had been fining/ the taxi driver/ for ignoring the parking laws./ The chauffeur/ had too/ and the/ fines were/ substantial./
Passive Implausible
The taxi driver/ had been fined by/ the traffic warden/ for ignoring the parking laws./
The chauffeur/ had too/ and the/ fines were/ substantial./
Passive Implausible
The taxi driver/ had been fined by/ the traffic warden/ for ignoring the parking laws./
The policeman/ had too/ and the/ fines were/ substantial./

32.
Active Plausible
The tour guide/ had been telling/ the tourist/ that the refurbishment was complete./ The owner/ had too/ during the/ new guided/ tour./
Active Implausible
The tour guide/ had been telling/ the tourist/ that the refurbishment was

complete./ The visitor/ had too/ during the/ new guided/ tour./

Passive Implausible

The tourist/ had been told by/ the tour guide/ that the refurbishment was complete./ The visitor/ had too/ during the/ new guided/ tour./

Passive Implausible

The tourist/ had been told by/ the tour guide/ that the refurbishment was complete./ The owner/ had too/ during the/ new guided/ tour./

Appendix E:

Experimental Materials for Experiment 6

(Slash marks ('/') denote eyetracking region boundaries)

1.
Neutral Order 1
It was getting dark in the local park./ After frightening the
old lady,/ the child/ spotted/ the mugger/ and ran/ off quickly./
Neutral Order 2
It was getting dark in the local park./ After frightening the
old lady,/ the mugger/ spotted/ the child/ and ran/ off quickly./
Biased Order 1
It was getting dark in the local park./ After kissing the
old lady,/ the child/ spotted/ the mugger/ and ran/ off quickly./
Biased Order 2
It was getting dark in the local park./ After kissing the
old lady,/ the mugger/ spotted/ the child/ and ran/ off quickly./

2.
Neutral Order 1
The circus was visiting the primary school./ After seeing the
schoolboy,/ the clown/ spoke to/ the headmaster/ in the/ assembly hall./
Neutral Order 2
The circus was visiting the primary school./ After seeing the
schoolboy,/ the headmaster/ spoke to/ the clown/ in the/ assembly hall./
Biased Order 1
The circus was visiting the primary school./ After entertaining the
schoolboy,/ the clown/ spoke to/ the headmaster/ in the/ assembly hall./
Biased Order 2
The circus was visiting the primary school./ After entertaining the
schoolboy,/ the headmaster/ spoke to/ the clown/ in the/ assembly hall./

3.
Neutral Order 1
It was a typically frantic day in the office./ After speaking to the
employee,/ the manager/ emailed/ the trainee/ with some/ new information./
Neutral Order 2
It was a typically frantic day in the office./ After speaking to the
employee,/ the trainee/ emailed/ the manager/ with some/ new information./
Biased Order 1
It was a typically frantic day in the office./ After sacking the
employee,/ the manager/ emailed/ the trainee/ with some/ new information./
Biased Order 2
It was a typically frantic day in the office./ After sacking the
employee,/ the trainee/ emailed/ the manager/ with some/ new information./

4.

Neutral Order 1
The courtroom gallery was completely full./ After listening to the defendant,/ the lawyer/ noticed/ the journalist/ writing in/ his notebook./
Neutral Order 2
The courtroom gallery was completely full./ After listening to the defendant,/ the journalist/ noticed/ the lawyer/ writing in/ his notebook./
Biased Order 1
The courtroom gallery was completely full./ After quizzing the defendant,/ the lawyer/ noticed/ the journalist/ writing in/ his notebook./
Biased Order 2
The courtroom gallery was completely full./ After quizzing the defendant,/ the journalist/ noticed/ the lawyer/ writing in/ his notebook./

5.
Neutral Order 1
There had been some complaints in the restaurant./ After checking the food,/ the chef/ called/ the waiter/ in a/ loud voice./
Neutral Order 2
There had been some complaints in the restaurant./ After checking the food,/ the waiter/ called/ the chef/ in a/ loud voice./
Biased Order 1
There had been some complaints in the restaurant./ After cooking the food,/ the chef/ called/ the waiter/ in a/ loud voice./
Biased Order 2
There had been some complaints in the restaurant./ After cooking the food,/ the waiter/ called/ the chef/ in a/ loud voice./

6.
Neutral Order 1
The philosophy course was always popular./ After hearing the lecture,/ the professor/ spoke to/ the student/ about its/ main themes./
Neutral Order 2
The philosophy course was always popular./ After hearing the lecture,/ the student/ spoke to/ the professor/ about its/ main themes./
Biased Order 1
The philosophy course was always popular./ After giving the lecture,/ the professor/ spoke to/ the student/ about its/ main themes./
Biased Order 2
The philosophy course was always popular./ After giving the lecture,/ the student/ spoke to/ the professor/ about its/ main themes./

7.
Neutral Order 1
The front page deadline was looming./ After reading the article,/ the journalist/ passed it to/ the editor/ for a/ final check./
Neutral Order 2
The front page deadline was looming./ After reading the article,/ the editor/ passed it to/ the journalist/ for a/ final check./
Biased Order 1
The front page deadline was looming./ After writing the

article,/ the journalist/ passed it to/ the editor/ for a/ final check./
Biased Order 2
The front page deadline was looming./ After writing the
article,/ the editor/ passed it to/ the journalist/ for a/ final check./

8.
Neutral Order 1
People can get very upset in hospitals./ After calming the
patient,/ the doctor/ called for/ the relative/ to come/ and help./
Neutral Order 2
People can get very upset in hospitals./ After calming the
patient,/ the relative/ called for/ the doctor/ to come/ and help./
Biased Order 1
People can get very upset in hospitals./ After treating the
patient,/ the doctor/ called for/ the relative/ to come/ and help./
Biased Order 2
People can get very upset in hospitals./ After treating the
patient,/ the relative/ called for/ the doctor/ to come/ and help./

9.
Neutral Order 1
The construction work was finally finished./ After seeing the
house,/ the client/ contacted/ the agent/ about some/ legal matters./
Neutral Order 2
The construction work was finally finished./ After seeing the
house,/ the agent/ contacted/ the client/ about some/ legal matters./
Biased Order 1
The construction work was finally finished./ After buying the
house,/ the client/ contacted/ the agent/ about some/ legal matters./
Biased Order 2
The construction work was finally finished./ After buying the
house,/ the agent/ contacted/ the client/ about some/ legal matters./

10.
Neutral Order 1
Stomach operations can go on for hours./ While watching the
operation,/ the surgeon/ chatted to/ the nurse/ about a/ recent holiday./
Neutral Order 2
Stomach operations can go on for hours./ While watching the
operation,/ the nurse/ chatted to/ the surgeon/ about a/ recent holiday./
Biased Order 1
Stomach operations can go on for hours./ While performing the
operation,/ the surgeon/ chatted to/ the nurse/ about a/ recent holiday./
Biased Order 2
Stomach operations can go on for hours./ While performing the
operation,/ the nurse/ chatted to/ the surgeon/ about a/ recent holiday./

11.
Neutral Order 1
Things were very tense down at the police station./ While speaking to the

suspect,/ the detective/ ignored/ the solicitor/ and was/ very aggressive./
Neutral Order 2
Things were very tense down at the police station./ While speaking to the suspect,/ the solicitor/ ignored/ the detective/ and was/ very aggressive./
Biased Order 1
Things were very tense down at the police station./ While interrogating the suspect,/ the detective/ ignored/ the solicitor/ and was/ very aggressive./
Biased Order 2
Things were very tense down at the police station./ While interrogating the suspect/ the solicitor/ ignored the/ detective/ and was/ very aggressive./

12.
Neutral Order 1
All the athletes were enjoying the Olympics./ After beating his opponent,/ the boxer/ watched/ the sprinter/ win yet/ another medal./
Neutral Order 2
All the athletes were enjoying the Olympics./ After beating his opponent,/ the sprinter/ watched/ the boxer/ win yet/ another medal./
Biased Order 1
All the athletes were enjoying the Olympics./ After knocking out his opponent,/ the boxer/ watched/ the sprinter/ win yet/ another medal./
Biased Order 2
All the athletes were enjoying the Olympics./ After knocking out his opponent,/ the sprinter/ watched/ the boxer/ win yet/ another medal./

13.
Neutral Order 1
It’s hard to predict if a book will be popular./ After reading the book,/ the writer/ sent it to/ the publisher/ for an/ experienced opinion./
Neutral Order 2
It’s hard to predict if a book will be popular./ After reading the book,/ the publisher/ sent it to/ the writer/ for an/ experienced/ opinion./
Biased Order 1
It’s hard to predict if a book will be popular./ After writing the book,/ the writer/ sent it to/ the publisher/ for an/ experienced opinion./
Biased Order 2
It’s hard to predict if a book will be popular./ After writing the book,/ the publisher/ sent it to/ the writer/ for an/ experienced opinion./

14.
Neutral Order 1
Most people prefer local shops to big supermarkets./ After praising the sausages,/ the butcher/ chatted to/ the customer/ for half/ an hour./
Neutral Order 2
Most people prefer local shops to big supermarkets./ After praising the sausages,/ the customer/ chatted to/ the butcher/ for half/ an hour./
Biased Order 1
Most people prefer local shops to big supermarkets./ After preparing the sausages,/ the butcher/ chatted to/ the customer/ for half/ an hour./
Biased Order 2

Most people prefer local shops to big supermarkets./ After preparing the sausages,/ the customer/ chatted to/ the butcher/ for half/ an hour./

15.
Neutral Order 1
There was water all over the floor./ After examining the
leak,/ the plumber/ explained the problem to/ the woman/ with the/ wet feet./
Neutral Order 2
There was water all over the floor./ After examining the
leak,/ the woman/ explained the problem to/ the plumber/ with the/ wet feet./
Biased Order 1
There was water all over the floor./ After fixing the
leak,/ the plumber/ explained/ the problem to/ the woman/ with the/ wet feet./
Biased Order 2
There was water all over the floor./ After fixing the
leak,/ the woman/ explained the problem to/ the plumber/ with the/ wet feet./

16.
Neutral Order 1
The safari trip was very exciting./ After seeing the
lion,/ the hunter/ described it to/ the photographer/ back at/ the camp./
Neutral Order 2
The safari trip was very exciting./ After seeing the
lion,/ the photographer/ described it to/ the hunter/ back at/ the camp./
Biased Order 1
The safari trip was very exciting./ After killing the
lion,/ the hunter/ described it to/ the photographer/ back at/ the camp./
Biased Order 2
The safari trip was very exciting./ After killing the
lion,/ the photographer/ described it to/ the hunter/ back at/ the camp./

17.
Neutral Order 1
Almost everyone needs a computer these days./ Before switching on the
PC,/ the customer/ admired it with/ the salesman/ in the/ electronics shop./
Neutral Order 2
Almost everyone needs a computer these days./ Before switching on the
PC,/ the salesman/ admired it with/ the customer/ in the/ electronics shop./
Biased Order 1
Almost everyone needs a computer these days./ Before buying the
PC,/ the customer/ admired it with/ the salesman/ in the/ electronics shop./
Biased Order 2
Almost everyone needs a computer these days./ Before buying the
PC,/ the salesman/ admired it with/ the customer/ in the/ electronics shop./

18.
Neutral Order 1
The internet can be very useful in education./ Before attending the
class,/ the teacher/ emailed/ the pupil/ about the/ essay questions./
Neutral Order 2

The internet can be very useful in education./ Before attending the class,/ the pupil/ emailed/ the teacher/ about the/ essay questions./

Biased Order 1

The internet can be very useful in education./ Before teaching the class,/ the teacher/ emailed/ the pupil/ about the/ essay questions./

Biased Order 2

The internet can be very useful in education./ Before teaching the class,/ the pupil/ emailed/ the teacher/ about the/ essay questions./

19.

Neutral Order 1

The parliamentary debate was almost over./ After hearing the final speech,/ the politician/ met with/ the secretary/ for a/ long debriefing./

Neutral Order 2

The parliamentary debate was almost over./ After hearing the final speech,/ the secretary/ met with/ the politician/ for a/ long debriefing./

Biased Order 1

The parliamentary debate was almost over./ After giving the final speech,/ the politician/ met with/ the secretary/ for a/ long debriefing./

Biased Order 2

The parliamentary debate was almost over./ After giving the final speech,/ the secretary/ met with/ the politician/ for a/ long debriefing./

20.

Neutral Order 1

Tempers had flared in the nursery./ After shouting at the babysitter,/ the toddler/ went to/ the mother/ and hugged/ her tightly./

Neutral Order 2

Tempers had flared in the nursery./ After shouting at the babysitter,/ the mother/ went to/ the toddler/ and hugged/ her tightly./

Biased Order 1

Tempers had flared in the nursery./ After kicking the babysitter,/ the toddler/ went to/ the mother/ and hugged/ her tightly./

Biased Order 2

Tempers had flared in the nursery./ After kicking the babysitter,/ the mother/ went to/ the toddler/ and hugged/ her tightly./

21.

Neutral Order 1

The movie was in the final stages of production./ While viewing the final scene,/ the director/ criticised/ the actor/ for his/ poor technique./

Neutral Order 2

The movie was in the final stages of production./ While viewing the final scene,/ the actor/ criticised/ the director/ for his/ poor technique./

Biased Order 1

The movie was in the final stages of production./ While editing the final scene,/ the director/ criticised/ the actor/ for his/ poor technique./

Biased Order 2

The movie was in the final stages of production./ While editing the final scene,/ the actor/ criticised/ the director/ for his/ poor technique./

22.
Neutral Order 1
Road accidents can be very traumatic./ After seeing the
crash,/ the driver/ spoke to/ the priest/ about his/ terrible nightmares./
Neutral Order 2
Road accidents can be very traumatic./ After seeing the
crash,/ the priest/ spoke to/ the driver/ about his/ terrible nightmares./
Biased Order 1
Road accidents can be very traumatic./ After causing the
crash,/ the driver/ spoke to/ the priest/ about his/ terrible nightmares./
Biased Order 2
Road accidents can be very traumatic./ After causing the
crash,/ the priest/ spoke to/ the driver/ about his/ terrible nightmares./

23.
Neutral Order 1
The pub was quiet on Monday nights./ After describing the
cocktail,/ the barman/ recommended another to/ the customer/ he was/ chatting to./
Neutral Order 2
The pub was quiet on Monday nights./ After describing the
cocktail,/ the customer/ recommended another/ to the barman/ he was/ chatting to./
Biased Order 1
The pub was quiet on Monday nights./ After serving the
cocktail,/ the barman/ recommended another to/ the customer/ he was/ chatting to./
Biased Order 2
The pub was quiet on Monday nights./ After serving the
cocktail,/ the customer/ recommended another to/ the barman/ he was/ chatting to./

24.
Neutral Order 1
The music industry is worth millions./ Before releasing the
song,/ the singer/ argued with/ the producer/ over the/ new contract./
Neutral Order 2
The music industry is worth millions./ Before releasing the
song,/ the producer/ argued with/ the singer/ over the/ new contract./
Biased Order 1
The music industry is worth millions./ Before performing the
song,/ the singer/ argued with/ the producer/ over the/ new contract./
Biased Order 2
The music industry is worth millions./ Before performing the
song,/ the producer/ argued with/ the singer/ over the/ new contract./

25.
Neutral Order 1
The church was always full of lively chat./ After reading the interesting

sermon,/ the minister/ discussed it with/ the organist/ over a/ cup of tea./
Neutral Order 2
The church was always full of lively chat./ After reading the interesting sermon,/ the organist/ discussed it with/ the minister/ over a cup of tea./
Biased Order 1
The church was always full of lively chat./ After preaching the interesting sermon,/ the minister/ discussed it with/ the organist/ over a/ cup of tea./
Biased Order 2
The church was always full of lively chat./ After preaching the interesting sermon,/ the organist/ discussed it with/ the minister/ over a cup of tea./

26.
Neutral Order 1
All the studying had led up to this day./ After reading the question,/ the student/ just stared at/ the examiner/ and didn't/ say anything./
Neutral Order 2
All the studying had led up to this day./ After reading the question,/ the examiner/ just stared at/ the student/ and didn't/ say anything./
Biased Order 1
All the studying had led up to this day./ After answering the question,/ the student/ just stared at/ the examiner/ and didn't/ say anything./
Biased Order 2
All the studying had led up to this day./ After answering the question,/ the examiner/ just stared at/ the student/ and didn't/ say anything./

27.
Neutral Order 1
Everything had to be perfect for the dinner party./ After inspecting the dinner table,/ the butler/ called for/ the duchess/ in a/ loud voice./
Neutral Order 2
Everything had to be perfect for the dinner party./ After inspecting the dinner table,/ the duchess/ called for/ the butler/ in a/ loud voice./
Biased Order 1
Everything had to be perfect for the dinner party./ After polishing the dinner table,/ the butler/ called for/ the duchess/ in a/ loud voice./
Biased Order 2
Everything had to be perfect for the dinner party./ After polishing the dinner table,/ the duchess/ called for/ the butler/ in a/ loud voice./

28.
Neutral Order 1
The airport had been busy all summer./ While lifting the suitcase,/ the guard/ chatted to/ the passenger/ and was/ very friendly./
Neutral Order 2
The airport had been busy all summer./ While lifting the suitcase,/ the passenger/ chatted to/ the guard/ and was/ very friendly./
Biased Order 1
The airport had been busy all summer./ While searching the suitcase,/ the guard/ chatted to/ the passenger/ and was/ very friendly./
Biased Order 2

The airport had been busy all summer./ While searching the suitcase,/ the passenger/ chatted to/ the guard/ and was/ very friendly./

Appendix F:

Experimental Materials for Experiment 7

(Slash marks ('/') indicate eyetracking region boundaries)

1.
Good Context Plausible
The service was slow in the restaurant./
The waiter/ served/ the meal/ but the/ kitchen/ was in chaos./
Good Context Anomalous
The service was slow in the restaurant./
The menu/ served/ the meal/ but the/ kitchen/ was in chaos./
Neutral Context Plausible
It was almost nine o'clock./
The waiter/ served/ the meal/ but the/ kitchen/ was in chaos./
Neutral Context Anomalous
It was almost nine o'clock./
The menu/ served/ the meal/ but the/ kitchen/ was in chaos./

2.
Good Context Plausible
There was a new play on in the theatre./
The actor/ recited/ the speech/ that/ began/ the first act/.
Good Context Anomalous
There was a new play on in the theatre./
The spotlight/ recited/ the speech/ that/ began/ the first act./
Neutral Context Plausible
It was a warm July evening./
The actor/ recited/ the speech/ that/ began/ the first act./
Neutral Context Anomalous
It was a warm July evening./
The spotlight/ recited/ the speech/ that/ began/ the first act./

3.
Good Context Plausible
There were many horses in the stables./
The owner/ groomed/ the stallion/ ahead/ of the/ next race./
Good Context Anomalous
There were many horses in the stables./
The saddle/ groomed/ the stallion/ ahead/ of the/ next race./
Neutral Context Plausible
It was a gorgeous summers day./
The owner/ groomed/ the stallion/ ahead/ of the/ next race.
Neutral Context Anomalous
It was a gorgeous summers day./
The saddle/ groomed/ the stallion/ ahead/ of the/ next race./

4.

Good Context Plausible
It was very busy in the kitchen./
The chef/ cooked/ the casserole/ while/ the soup/ simmered gently./
Good Context Anomalous
It was very busy in the kitchen./
The knife/ cooked/ the casserole/ while/ the soup/ simmered gently./
Neutral Context Plausible
It had been a long morning./
The chef/ cooked/ the casserole/ while/ the soup/ simmered gently./
Neutral Context Anomalous
It had been a long morning./
The knife/ cooked/ the casserole/ while/ the soup/ simmered gently./

5.
Good Context Plausible
It was business as usual at the police station./
The detective/ questioned/ the suspect/ about/ the High/ Street robbery./
Good Context Anomalous
It was business as usual at the police station./
The handcuffs/ questioned/ the suspect/ about/ the High/ Street robbery./
Neutral Context Plausible
It was a cloudy afternoon./
The detective/ questioned/ the suspect/ about/ the High/ Street robbery./
Neutral Context Anomalous
It was a cloudy afternoon./
The handcuffs/ questioned/ the suspect/ about/ the High/ Street robbery./

6.
Good Context Plausible
Everything was going well on the long flight./
The passenger/ requested/ a blanket/ while/ the stewardess/ served tea./
Good Context Anomalous
Everything was going well on the long flight./
The plane/ requested/ a blanket/ while/ the stewardess/ served tea./
Neutral Context Plausible
Everything was going according to plan./
The passenger/ requested/ a blanket/ while/ the stewardess/ served tea./
Neutral Context Anomalous
Everything was going according to plan./
The plane/ requested/ a blanket/ while/ the stewardess/ served tea./

7.
Good Context Plausible
It was nice and quiet in the supermarket./
The shopper/ paid for/ the shopping/ while/ the bags/ were packed./
Good Context Anomalous
It was nice and quiet in the supermarket./
The checkout/ paid for/ the shopping/ while/ the bags/ were packed./
Neutral Context Plausible
It was quiet at that time in the morning./

The shopper/ paid for/ the shopping/ while/ the bags/ were packed./
Neutral Context Anomalous
It was quiet at that time in the morning./
The checkout/ paid for/ the shopping/ while/ the bags/ were packed./

8.
Good Context Plausible
Its easy to lose your temper in a classroom./
The teacher/ shouted at/ the pupil/ who was/ talking/ too loudly./
Good Context Anomalous
Its easy to lose your temper in a classroom./
The blackboard/ shouted at/ the pupil/ who was/ talking/ too loudly./
Neutral Context Plausible
It was just after lunchtime./
The teacher/ shouted at/ the pupil/ who was/ talking/ too loudly./
Neutral Context Anomalous
It was just after lunchtime./
The blackboard/ shouted at/ the pupil/ who was/ talking/ too loudly./

9.
Good Context Plausible
A lot of fans had come to the book-signing./
The author/ wrote/ some autographs/ once/ the formalities/ were over./
Good Context Anomalous
A lot of fans had come to the book-signing./
The novel/ wrote/ some autographs/ once/ the formalities/ were over./
Neutral Context Plausible
A lot of people had turned up./
The author/ wrote/ some autographs/ once/ the formalities/ were over./
Neutral Context Anomalous
A lot of people had turned up./
The novel/ wrote/ some autographs/ once/ the formalities/ were over./

10.
Good Context Plausible
There was a friendly atmosphere in the hairdressers'./
The barber/ dried/ the customer's hair/ while/ the discussion/ carried on./
Good Context Anomalous
There was a friendly atmosphere in the hairdressers'./
The scissors/ dried/ the customer's hair/ while/ the discussion/ carried on./
Neutral Context Plausible
The brown leaves suggested autumn was approaching./
The barber/ dried/ the customer's hair/ while/ the discussion/ carried on./
Neutral Context Anomalous
The brown leaves suggested autumn was approaching./
The scissors/ dried/ the customer's hair/ while/ the discussion/ carried on./

11.
Good Context Plausible
It was very quiet in the laboratory./

The scientist/ calculated/ some results/ and they/ looked/ quite interesting./
Good Context Anomalous
It was very quiet in the laboratory./
The test tube/ calculated/ some results/ and they/ looked/ quite interesting./
Neutral Context Plausible
It was almost time for a coffee break./
The scientist/ calculated/ some results/ and they/ looked/ quite interesting./
Neutral Context Anomalous
It was almost time for a coffee break./
The test tube/ calculated/ some results/ and they/ looked/ quite interesting./

12.
Good Context Plausible
Things were peaceful in the hospital./
The doctor/ interviewed/ the patient/ after/ the full/ medical examination./
Good Context Anomalous
Things were peaceful in the hospital./
The operation/ interviewed/ the patient/ after/ the full/ medical examination./
Neutral Context Plausible
It was the middle of the afternoon./
The doctor/ interviewed/ the patient/ after/ the full/ medical examination./
Neutral Context Anomalous
It was the middle of the afternoon./
The operation/ interviewed/ the patient/ after/ the full/ medical examination./

13.
Good Context Plausible
Everyone was working hard in the bank./
The manager/ helped/ the customer/ with/ some new/ account forms./
Good Context Anomalous
Everyone was working hard in the bank./
The chequebook/ helped/ the customer/ with/ some new/ account forms./
Neutral Context Plausible
Everyone was feeling tired./
The manager/ helped/ the customer/ with/ some new/ account forms./
Neutral Context Anomalous
Everyone was feeling tired./
The chequebook/ helped/ the customer/ with/ some new/ account forms./

14.
Good Context Plausible
It was a pleasant summer evening in the garden./
The gardener/ watered/ the daffodils/ with/ the new/ watering can./
Good Context Anomalous
It was a pleasant summer evening in the garden./
The barbeque/ watered/ the daffodils/ with/ the new/ watering can./
Neutral Context Plausible
The radio appeared to be broken./
The gardener/ watered/ the daffodils/ with/ the new/ watering can./
Neutral Context Anomalous

The radio appeared to be broken./
The barbeque/ watered/ the daffodils/ with/ the new/ watering can./

15.
Good Context Plausible
There was a lot of grumbling in the hotel./
The guest/ complained to/ the manager/ about/ the faulty/ shower head./
Good Context Anomalous
There was a lot of grumbling in the hotel./
The bedroom/ complained to/ the manager/ about/ the faulty/ shower head./
Neutral Context Plausible
It was just another boring day./
The guest/ complained to/ the manager/ about/ the faulty/ shower head./
Neutral Context Anomalous
It was just another boring day./
The bedroom/ complained to/ the manager/ about/ the faulty/ shower head./

16.
Good Context Plausible
There was tension on the building site./
The foreman/ sacked/ the bricklayer/ because/ of the/ recent thefts./
Good Context Anomalous
There was tension on the building site./
The scaffolding/ sacked/ the bricklayer/ because/ of the/ recent thefts./
Neutral Context Plausible
It was the last day of the week./
The foreman/ sacked/ the bricklayer/ because/ of the/ recent thefts./
Neutral Context Anomalous
It was the last day of the week./
The scaffolding/ sacked/ the bricklayer/ because/ of the/ recent thefts./

17.
Good Context Plausible
Everyone was in the television studio./
The producer/ argued with/ the director/ about/ the new/ studio lights./
Good Context Anomalous
Everyone was in the television studio./
The camera/ argued with/ the director/ about/ the new/ studio lights./
Neutral Context Plausible
It was the first day of the week./
The producer/ argued with/ the director/ about/ the new/ studio lights./
Neutral Context Anomalous
It was the first day of the week./
The camera/ argued with/ the director/ about/ the new/ studio lights./

18.
Good Context Plausible
There were important developments at the newspaper office./
The editor/ wrote/ a memo for/ the reporter/ concerning/ the new/ staff roles./
Good Context Anomalous

There were important developments at the newspaper office./
The article/ wrote/ a memo for/ the reporter/ concerning/ the new/ staff roles./
Neutral Context Plausible
There was music coming from somewhere./
The editor/ wrote/ a memo for/ the reporter/ concerning/ the new/ staff roles./
Neutral Context Anomalous
There was music coming from somewhere./
The article/ wrote/ a memo for/ the reporter/ concerning/ the new/ staff roles./

19.
Good Context Plausible
It was a beautiful day on the beach./
The swimmer/ swam near/ the shoreline/ where/ the children/ were playing./
Good Context Anomalous
It was a beautiful day on the beach./
The sandcastle/ swam near/ the shoreline/ where/ the children/ were playing./
Neutral Context Plausible
It was a windy day across the country./
The swimmer/ swam near/ the shoreline/ where/ the children/ were playing./
Neutral Context Anomalous
It was a windy day across the country./
The sandcastle/ swam near/ the shoreline/ where/ the children/ were playing./

20.
Good Context Plausible
There was a bad decision at the football match./
The player/ criticised/ the referee/ while/ other players/ stayed calm./
Good Context Anomalous
There was a bad decision at the football match./
The ball/ criticised/ the referee/ while/ other players/ stayed calm./
Neutral Context Plausible
It was a bank holiday weekend./
The player/ criticised/ the referee/ while/ other players/ stayed calm./
Neutral Context Anomalous
It was a bank holiday weekend./
The ball/ criticised/ the referee/ while/ other players/ stayed calm./

21.
Good Context Plausible
Work started early down on the farm./
The farmer/ whistled to/ the sheepdog/ while/ the sheep/ ran around./
Good Context Anomalous
Work started early down on the farm./
The tractor/ whistled to/ the sheepdog/ while/ the sheep/ ran around./
Neutral Context Plausible
It was the first week of the month./
The farmer/ whistled to/ the sheepdog/ while/ the sheep/ ran around./
Neutral Context Anomalous
It was the first week of the month./
The tractor/ whistled to/ the sheepdog/ while/ the sheep/ ran around./

22.
Good Context Plausible
It was silent in the examination hall./
The student/ read/ the question/ about/ the first/ world war./
Good Context Anomalous
It was silent in the examination hall./
The desk/ read/ the question/ about/ the first/ world war./
Neutral Context Plausible
A bus was passing by outside./
The student/ read/ the question/ about/ the first/ world war./
Neutral Context Anomalous
A bus was passing by outside./
The desk/ read/ the question/ about/ the first/ world war./

23.
Good Context Plausible
It was busy at the gym./
The athlete/ lifted/ heavy weights/ during/ the morning/ training session./
Good Context Anomalous
It was busy at the gym./
The treadmill/ lifted/ heavy weights/ during/ the morning/ training session./
Neutral Context Plausible
It was less busy in the morning./
The athlete/ lifted/ heavy weights/ during/ the morning/ training session./
Neutral Context Anomalous
It was less busy in the morning./
The treadmill/ lifted/ heavy weights/ during/ the morning/ training session./

24.
Good Context Plausible
The room was packed for the press conference./
The journalist/ asked/ difficult questions/ about/ the new/ housing policies./
Good Context Anomalous
The room was packed for the press conference./
The microphone/ asked/ difficult questions/ about/ the new/ housing policies./
Neutral Context Plausible
The room was completely full./
The journalist/ asked/ difficult questions/ about/ the new/ housing policies./
Neutral Context Anomalous
The room was completely full./
The microphone/ asked/ difficult questions/ about/ the new/ housing policies./

25.
Good Context Plausible
It was chaos on the battlefield./
The soldier/ rescued/ the civilian/ from/ the machine/ gun fire./
Good Context Anomalous
It was chaos on the battlefield./
The landmine/ rescued/ the civilian/ from/ the machine/ gun fire./

Neutral Context Plausible
Everybody was running around./
The soldier/ rescued/ the civilian/ from/ the machine/ gun fire./
Neutral Context Anomalous
Everybody was running around./
The landmine/ rescued/ the civilian/ from/ the machine/ gun fire./

26.
Good Context Plausible
There was a delay at the train station./
The travellers/ drank/ coffee/ while/ an announcement/ was made./
Good Context Anomalous
There was a delay at the train station./
The tickets/ drank/ coffee/ while/ an announcement/ was made./
Neutral Context Plausible
It was turning into a pleasant day./
The travellers/ drank/ coffee/ while/ an announcement/ was made./
Neutral Context Anomalous
It was turning into a pleasant day./
The tickets/ drank/ coffee/ while/ an announcement/ was made./

27.
Good Context Plausible
There was a great atmosphere at the rock concert./
The crowd/ cheered/ loudly/ when/ each/ new song started./
Good Context Anomalous
There was a great atmosphere at the rock concert./
The drums/ cheered/ loudly/ when/ each/ new song started./
Neutral Context Plausible
The heavy rain was still pouring down./
The crowd/ cheered/ loudly/ when/ each/ new song started.
Neutral Context Anomalous
The heavy rain was still pouring down./
The drums/ cheered/ loudly/ when/ each/ new song started./

28.
Good Context Plausible
All the seats were full in the cinema./
The audience/ enjoyed/ the film/ about/ the invasion/ from Mars./
Good Context Anomalous
All the seats were full in the cinema./
The screen/ enjoyed/ the film/ about/ the invasion/ from Mars./
Neutral Context Plausible
It was Friday evening./
The audience/ enjoyed/ the film/ about/ the invasion/ from Mars./
Neutral Context Anomalous
It was Friday evening./
The screen/ enjoyed/ the film/ about/ the invasion/ from Mars./

29.

Good Context Plausible
The hunting trip was very exciting./
The hunter/ fired/ the rifle/ making/ a very/ loud bang./
Good Context Anomalous
The hunting trip was very exciting./
The deer/ fired/ the rifle/ making/ a very/ loud bang./
Neutral Context Plausible
The sun was about to set./
The hunter/ fired/ the rifle/ making/ a very/ loud bang./
Neutral Context Anomalous
The sun was about to set./
The deer/ fired/ the rifle/ making/ a very/ loud bang./

30.
Good Context Plausible
There was a special event at the art gallery./
The artist/ opened/ the exhibition/ once/ the crowds/ had arrived./
Good Context Anomalous
There was a special event at the art gallery./
The painting/ opened/ the exhibition/ once/ the crowds/ had arrived./
Neutral Context Plausible
People were entering the building./
The artist/ opened/ the exhibition/ once/ the crowds/ had arrived./
Neutral Context Anomalous
People were entering the building./
The painting/ opened/ the exhibition/ once/ the crowds/ had arrived./

31.
Good Context Plausible
It was all very impressive at the graduation ceremony./
The student/ smiled at/ the professor/ while/ the degrees/ were awarded./
Good Context Anomalous
It was all very impressive at the graduation ceremony./
The certificate/ smiled at/ the professor/ while/ the degrees/ were awarded./
Neutral Context Plausible
Everyone stood up when the music started./
The student/ smiled at/ the professor/ while/ the degrees/ were awarded./
Neutral Context Anomalous
Everyone stood up when the music started./
The certificate/ smiled at/ the professor/ while/ the degrees/ were awarded./

32.
Good Context Plausible
It was the annual school trip to the zoo./
The lion/ roared at/ the schoolboy/ while/ all the/ children jumped./
Good Context Anomalous
It was the annual school trip to the zoo./
The cage/ roared at/ the schoolboy/ while/ all the/ children jumped./
Neutral Context Plausible
It was the annual school trip./

The lion/ roared at/ the schoolboy/ while/ all the/ children jumped./
Neutral Context Anomalous
It was the annual school trip./
The cage/ roared at/ the schoolboy/ while/ all the/ children jumped./

Appendix G:

Experimental Materials for Experiment 8

(Slash marks ('/') denote eyetracking region boundaries)

1.
Good Context
The army had already begun attacking the city./
The soldier/ was protected by/ the child/ during/ all the/ heavy/ shooting./
Neutral Context
The buildings in the city were very tall./
The soldier/ was protected by/ the child/ during/ all the/ heavy/ shooting./

2.
Good Context
The robbery had gone wrong./
The policeman/ was chased by/ the burglar/ down/ the dark/ empty/ street./
Neutral Context
It was the middle of the night./
The policeman/ was chased by/ the burglar/ down/ the dark/ empty street./

3.
Good Context
It was opening night in the new restaurant/
The waitress/ was served by/ the customer/ while/ the barman/ poured/ drinks./
Neutral Context
A lot of people had turned up/
The waitress/ was served by/ the customer/ while/ the barman/ poured/ drinks/

4.
Good Context
It was a busy morning in the GP's surgery/
The doctor/ was treated by/ the patient/ inside/ the new/ treatment/ room/
Neutral Context
It had been a productive morning/
The doctor/ was treated by/ the patient/ inside/ the new/ treatment/ room/

5.
Good Context
It was very cold in the haunted house./
The ghost/ was terrified by/ the woman/ inside/ the old/ drawing/ room./
Neutral Context
It was a particularly cold winter./
The ghost/ was terrified by/ the woman/ inside/ the old/ drawing/ room./

6.
Good Context
The spelling test results were announced./

The teacher/ was praised by/ the pupil/ because/ the score/ was/ impressive./
Neutral Context
The announcement had been made./
The teacher/ was praised by/ the pupil/ because/ the score/ was/ impressive./

7.
Good Context
Harassment in the workplace was a serious issue./
The boss/ was bullied by/ the trainee/ after/ the weekly/ staff/ meeting./
Neutral Context
The corridor was getting rather cluttered./
The boss/ was bullied by/ the trainee/ after/ the weekly/ staff/ meeting./

8.
Good Context
There had been a lot of tickets sold for the boxing match./
The boxer/ was punched by/ the referee/ during/ the exciting/ fourth/ round./
Neutral Context
A huge number of tickets had been sold./
The boxer/ was punched by/ the referee/ during/ the exciting/ fourth/ round./

9.
Good Context
The participants in the marathon were getting exhausted./
The spectator/ was applauded by/ the runner/ during/ the final/ uphill/ stretch./
Neutral Context
People were starting to get tired./
The spectator/ was applauded by/ the runner/ during/ the final/ uphill/ stretch./

10.
Good Context
There was a grand ceremony in the palace./
The king/ was rewarded by/ the farmer/ while/ the band/ played/ loudly./
Neutral Context
The big day had finally arrived./
The king/ was rewarded by/ the farmer/ while/ the band/ played/ loudly./

11.
Good Context
A crowd gathered outside the prison./
The guard/ was released by/ the prisoner/ after/ five years/ hard/ labour./
Neutral Context
A large crowd had gathered./
The guard/ was released by/ the prisoner/ after/ five years/ hard/ labour./

12.
Good Context
The clothes shop sold very expensive suits./
The tailor/ was measured by/ the customer/ inside/ the big/ changing/ room./
Neutral Context

The town centre was full of people./
The tailor/ was measured by/ the customer/ inside/ the big/ changing/ room./

13.
Good Context
The weather was perfect for the wedding./
The priest/ was blessed by/ the bride/ before/ the service/ began/ properly./
Neutral Context
It was Saturday afternoon./
The priest/ was blessed by/ the bride/ before/ the service/ began/ properly./

14.
Good Context
The construction site was a sexist environment./
The builder/ was whistled at by/ the blonde/ while/ the other/ builders/ cheered./
Neutral Context
Sometimes it's necessary to work late./
The builder/ was whistled at by/ the blonde/ while/ the other/ builders/ cheered./

15.
Good Context
It was very luxurious inside the limousine./
The chauffeur/ was driven by/ the celebrity/ around/ the vibrant/ city/ centre./
Neutral Context
The streets were all lit up./
The chauffeur/ was driven by/ the celebrity/ around/ the vibrant/ city/ centre./

16.
Good Context
The new course of biology lectures had started./
The professor/ was taught by/ the class/ about/ the evolution/ of/ mammals./
Neutral Context
The holidays had finally come to an end./
The professor/ was taught by/ the class/ about/ the evolution/ of/ mammals./

17.
Good Context
Helicopter flights can be very exciting./
The pilot/ was flown by/ the tourist/ around/ the impressive/ mountain/ range./
Neutral Context
The autumn leaves had begun to fall./
The pilot/ was flown by/ the tourist/ around/ the impressive/ mountain/ range./

18.
Good Context
The circus had arrived in town./
The clown/ was entertained by/ the children/ during/ the/ first/ performance./
Neutral Context
Everyone was feeling relaxed./
The clown/ was entertained by/ the children/ during/ the/ first/ performance./

19.
Good Context
There was along queue in the chemist's shop./
The pharmacist/ was advised by/ the customer/ about/ good hay/ fever/ medicine./
Neutral Context
There certainly was a long queue./
The pharmacist/ was advised by/ the customer/ about/ good hay/ fever/ medicine./

20.
Good Context
There was an article about the controversial novel in the newspaper./
The reporter/ was interviewed by/ the author/ about/ people's/ strong/ reactions./
Neutral Context
People had become quite interested./
The reporter/ was interviewed by/ the author/ about/ people's/ strong/ reactions./

21.
Good Context
The bus finally arrived at the bus stop./
The passenger/ was paid by/ the driver/ while/ the ticket/ was/ printing./
Neutral Context
It was half past eight in the morning./
The passenger/ was paid by/ the driver/ while/ the ticket/ was/ printing./

22.
Good Context
Some people's teeth are in a very bad state./
The dentist/ was warned by/ the child/ about/ maintaining/ good oral/ hygiene./
Neutral Context
It was the middle of a busy week./
The dentist/ was warned by/ the child/ about/ maintaining/ good oral/ hygiene./

23.
Good Context
The crime rate in the city was rising./
The mugger/ was robbed by/ the woman/ while/ several/ people just/ watched./
Neutral Context
It was a hot day in the city./
The mugger/ was robbed by/ the woman/ while/ several/ people just/ watched./

24.
Good Context
The fire in the building was blazing out of control./
The fireman/ was rescued by/ the girl/ while/ the roof/ began/ collapsing./
Neutral Context
The radio reported the news./
The fireman/ was rescued by/ the girl/ while/ the roof/ began/ collapsing./