Depth of Processing and Semantic Anomalies

Jason Thomas Bohan

Department of Psychology

University of Glasgow

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

December 2007

Abstract

The traditional view of language comprehension is that the meaning of a sentence is composed of the meaning of each word combined into a fully specified syntactic structure. These processes are assumed to be generally completed fully and automatically. However, there is increasing evidence that these processes may, in some circumstances, not be completed fully, and the resultant representation, underspecified. This is taken as evidence for shallow processing and is best typified, we argue, when readers fail to detect semantically anomalous words in a sentence. For example, when asked, "how many animals did Moses take on the Ark?" readers often incorrectly answer "two" failing to notice that it was Noah and not Moses who built the Ark. There has been surprisingly little work carried out on the on-line processing of these types of anomalies, and the differences in processing when anomalies are detected or missed. This thesis presents a series of studies, including four eye-tracking and one ERP study that investigates the nature of shallow processing as evidenced when participants report, or fail to report, hard-to-detect semantic anomalies. The main findings are that semantic anomaly detection is not immediate, but slightly delayed. Anomaly detection results in severe disruption in the eye movement data, and a late positivity in ERPs. There was some evidence that non-detected anomalies were processed unconsciously in both the eye movement record or in ERPs, however effects were weak and require replication. The rate of anomaly detection is also shown to be modulated by processing load and experimental task instructions. The discussion considers what these results reveal about the nature of shallow processing.

Acknowledgments

I would like to thank my supervisor, Professor Tony Sanford, for giving me this very special opportunity as well as the honour of working with him. He has been an excellent supervisor who has taught me so much and it has been a real pleasure working with him.

I would also like to thank Hartmut Leuthold for all his help and patience in setting up my ERP study, and also to Patrick Sturt for his help with eye tracking.

I want to thank Linda Moxey for her invaluable advice, for generously giving me so much of her time when I most needed it, and for being such a good friend.

I owe a very special thank you to Ruth Filik who has helped me in so many ways to complete this work. She has given me advice, assistance, encouragement, and even helped to keep me fit (even though I didn't always want her to!).

Finally I would like to thank Campbell Seaton who has always been there for me and who always believed I would get there, even when I didn't! Thank you for everything.

Table of Contents

Author's declaration

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

……..

Jason T Bohan

Publications

Within this thesis, Experiment 4 has been accepted for publication.

Chapter 4

Bohan, J., and Sanford, A.J. (2008) Semantic Anomalies at the borderline of consciousness: An eye-tracking investigation. *The Quarterly Journal of Experimental Psychology*, 61(2), 232-240.

Introduction

The topic of this thesis is shallow processing in language, which is demonstrated when readers fail to detect semantically anomalous words in text. Shallow processing, in this sense, refers to the idea that the contributions of syntactic and semantic processes to comprehension may, in some circumstances, not be carried out fully, and that the resultant mental representation of a text may be underspecified.

One of the best ways to demonstrate that shallow processing has occurred, we will argue, is when readers fail to detect semantically anomalous words. To illustrate what we mean by semantic anomaly, consider the question in [1]:

[1] Can a man marry his widow's sister?

Nearly 90% of people who were asked this question responded "yes" in this study. In other words, they answered that it was possible for a man who is dead to marry his living wife's sister! This is because the word *widow* refers to a woman who has outlived her husband and the male is, therefore, in no position to re-marry (Sanford and Bohan, forthcoming). Readers appeared to understand this question, and even answered it, and yet the anomalous nature of the question was mostly undetected. This is a clear demonstration of shallow semantic processing. Our study also suggests that failure to detect semantic anomalies may be an excellent indicator of when shallow processing has occurred.

Until recently there has been little empirical evidence of shallow semantic processing. However evidence has grown in the last few years and in this chapter, a broad range of evidence is reviewed that demonstrates shallow processing in language comprehension. This is followed by a more detailed review of research on semantic anomalies,

exemplified by the *widow's sister* example above. Before this however, it is important to consider how the idea of shallow processing fits into the standard views of language comprehension.

Traditional models and psycholinguistic assumptions

The orthodox view of language comprehension is that the meaning of a sentence is composed of the meanings of each individual word combined into a fully specified syntactic structure. It is assumed that the processes involved are "generally completed fully and relatively automatically" (Christianson, Williams, Zacks, & Ferreira, 2006, p.206). The notion that these processes may not, in some circumstances, be carried out fully, is controversial within psycholinguistics. So, while MacDonald, Pearlmutter, & Seidenberg (1994) acknowledged that, "the communicative goal of the listener can be achieved with only a partial analysis of the sentence," they viewed "these as degenerate cases." (p. 686). While shallow processing is controversial within psycholinguistics, Christianson et al. point out that this is not the case in other areas of cognitive psychology where the notion that mental representations may, in some situations, be underspecified, has been accepted more readily. For example, it has been shown that the subjective visual perception of a scene is not based on a true and exact representation of that scene (Henderson and Hollingworth, 1999; Irwin, 1996; Simons & Levin, 1997), and that human judgement and decision making may be influenced more by heuristics and biases than by a full consideration of available information (Kahneman, Slovic, & Tversky, 1982).

Earlier, Just & Carpenter (1980) argued that the interpretation of a sentence occurred incrementally, and that semantic information for each word is fully retrieved (if possible) during the incremental process. For example, they wrote, "readers interpret a word while they are fixating it, and they continue to fixate it until they have processed it

as far as they can" (p30). There is good evidence supporting incremental semantic processing, for example, through the ease with which incongruous words are identified in text. Traxler & Pickering (1996) illustrated this when they compared readers eye movements in [2a,b]:

> [2a] That's the pistol with which the man shot the gangster yesterday afternoon.
>
> [2b] That's the garage with which the man shot the gangster yesterday afternoon.

Traxler & Pickering reported that initial fixations on the word *shot* were longer in [2b] than [2a], which suggests that the word's meaning was accessed and integrated as soon as it was fixated. Further evidence demonstrating incremental semantic processing was provided by Altmann & Kamide (1999). They used a visual world paradigm, in which participants listen to statements while viewing a pictorial scene containing sentence relevant and non-relevant objects. They observed that there were more saccadic eye movements towards objects that were restricted by a preceding verb, even before the referent has been uttered (e.g. looking towards a picture of a cake while listening to a sentence about somebody eating something). This suggests that semantic analysis occurs incrementally because enough semantic information is being processed at the verb to predict soon-to-be encountered referents.

In general terms, the principle of compositionality (Fodor & Pylyshyn 1988) also supports the assumption of incremental and immediate semantic processing. Compositionality assumes that the meaning of a word is fully retrieved and combined, under the rules of syntax, to produce a final interpretation. Furthermore, the meaning of a word is stable across different sentences (i.e. the semantic properties are 'context-independent'). However, a strict interpretation of compositionality has been criticised by connectionist theorists. For example, McClelland, St.John & Taraband (1989)

argued against the idea that a word can only ever contribute the same meaning to all sentences in which it appears as "representing an impoverished view of the comprehension process" (p.322). By way of illustration they considered the semantic contribution of the word *ball* in [3a, 3 b, and 3c]

[3a] The hostess threw the ball for charity.

[3b] The slugger hit the ball over the fence.

[3c] The baby rolled the ball to her daddy.

In these three sentences the word *ball* in [3a] is different from [3b and 3c], however even in [3b and 3c] we are likely to imagine different types of balls, (one a hard sports ball, and the other a softer ball suitable for a young child). Compositionality, McClelland et al. argued, fails to capture these different shades of meaning, unless each shade was to have a separate lexical entry. More directly relevant to the present thesis is McClelland et al.'s observation that the *Moses* illusion not only demonstrates that semantic retrieval is not exhaustive, but also that the "meaning" of a to-be-retrieved word is sometimes difficult to define. To illustrate, consider the question, "How many of each animal did Moses take on the Ark?" Many readers fail to notice that the question is anomalous because it was Noah who built the Ark and not Moses (Erickson & Matteson, 1981). In direct conflict with the notion of compositionality, failure to detect this anomaly indicates that word meaning has not been retrieved fully. It is also difficult to define what meaning is in this example because the names Moses and Noah merely represent pointers to encyclopaedic knowledge.

While full and incremental processes are assumed in many process models, it may be more appropriate to say that, with the exception of Just & Carpenter's early work, that psycholinguistic research has just not been concerned with proving these assumptions.

Rather, a lot of work has been more concerned with discovering how immediately meaning is accessed and combined in processing. It is also very clear from the evidence that a great deal of processing is immediate (for example, on prediction Altmann & Steedman, 1988; Altmann & Kamide 1999; Kamide, Altmann & Haywood, 2003; Hagoort & van Berkum 2007; on pronoun resolution, Garrod, Freudenthal, & Boyle, 1994; Gordon, & Hendrick, 1998; Sanford, & Garrod, 1989; Sanford, Filik, Emmott, & Morrow, (in press); Sanford, Garrod, Lucas, & Henderson, 1983; Van Berkum, Zwitserlood, Bastiaansen, Brown, & Hagoort, 2004; and on the detection of syntactic and pragmatic anomalies, Braze, Shankweiler, Ni, & Palumbo, 2002; Ni, Fodor, Crain, & Shankweiler, 1998). However, it is not clear that current accounts really do make the assumption that semantic information is fully retrieved immediately on encountering every word. It may be that while researchers concerned with shallow processing have presented a "full processing" assumption to provoke debate, this has obscured the real questions which are, what is meant by *shallow processing*, and what modulates it in sentence comprehension? In the next section we begin by considering the scope of evidence for shallow processing in sentence comprehension.

Shallow processing: Scope of the evidence

The idea that semantic processing may be shallow is to be found in a wide range of literature. The principal summary position statements on shallow processing appeared in Sanford & Sturt (2002), while the use of the closely-related expression "Good Enough Representation" appeared first in Ferreira & Henderson (1999), and was expanded upon in Christianson, Hollingworth, Halliwell, & Ferreira (2001), and Ferreira, Bailey & Ferraro (2002).

Sanford & Sturt (2002) argued that many processes may be shallow or incomplete, and the final representation, underspecified. They also argued that a full and detailed

analysis is often neither necessary nor desirable in many situations. Furthermore, whether processing is shallow or deep is a dynamic property of the system, and may be modulated by linguistic devices such as focus. In support of their arguments they presented a diverse range of evidence from anomaly detection and text change detection studies. These results are detailed later.

Ferreira et al. (2002) challenged the assumption that sentence meaning is derived from an incremental analysis of the linguistic input. They argued that the language system adopts, in appropriate situations, a heuristic approach that gives rise to a 'good-enough' representation for the purposes of communication. They argue that while the language system essentially parses a sentence correctly, the interpretation relies on constant reinforcement of the syntactic structure. Reinforcement may come from existing schemas or a supportive context. Without this reinforcement the syntactic structure may be 'lost', and this, potentially, gives rise to interference from some types of information, for example from pragmatic knowledge, with the final interpretation of the sentence. Two lines of evidence that Ferreira et al. provide to support their arguments come from the misinterpretation of garden path, and passives sentences, and this is outlined a little later.

In discussing these two approaches, Sanford & Graesser (2006) highlight the important distinctions between the terms, shallow processing, underspecification, and good enough representations. The term *shallow processing* is used by Sanford & Sturt (2002) to refer to language processes that could have been carried out more thoroughly or extensively in different contexts. A possible outcome of shallow processing is that an *underspecified* representation is held. By way of an example, Sanford & Graesser consider the statement [4]:

[4] Every kid is up a tree.

There are several possible interpretations of (4), for example, just one tree exists that all kids are up; that there are many trees with one kid up each; that some trees have one or more kids up them; and that some trees have no kids up them. While a full semantic analysis would consider all of these options, an underspecified representation would not represent all of them. However, an underspecified representation may actually be '*good enough*' for the task in hand, so that a detailed and precise semantic analysis is unnecessary.

Evidence for shallow processing comes from a diverse range of research utilising a variety of methodologies. The methods adopted include incidental anomaly detection, text change detection, reading time, eye tracking studies, and memory tasks. The results have been interpreted as evidence for shallow processing in respect of semantic, syntactic, and interpretative processes. In the following sections, we review this evidence.

Failing to detect anomalies

Shallow processing may be inferred in situations where readers fail to notice mistakes, inconsistencies or anomalies in text. *Semantic anomaly* is a term used to refer to cases when an individual word is incorrectly used, normally within a highly constraining context. For some anomalies, readers may typically fail to detect them. For example, in the Moses illusion introduced earlier, many people fail to notice that the question, "How many of each animal did Moses take on the Ark?" is anomalous because it was Noah who built the Ark and not Moses (Erickson & Mattson 1981; Barton & Sanford 1993). Some studies of anomalies have used an incidental anomaly detection paradigm, where readers are presented with a text under naturalistic conditions and after reading are

asked to report anything anomalous in the text. Failure to notice semantic anomalies is strong evidence for shallow processing because if lexical recovery and integration into the sentence representation was immediate and complete, then these anomalies should be easily detected. Because such anomalies are at the centre of this thesis, literature in this area will be reviewed in detail in a later section of this chapter.

Failing to notice text changes

Sanford (2002) adopted a different methodology, text change detection, to illustrate that the level of representation across a text is graded, with the meaning of some words being more fully represented than others. This technique was adapted from the visual change blindness paradigm, where participants are presented with consecutive displays of a figure where some element is changed between displays. This technique has been used successfully to explore the role of attention and the detail of representation in visual processing (Simons & Levin 1997). In a similar fashion, text change detection requires participants to read two presentations of the same text and to detect whether an individual word has been changed on the second presentation. The logic of this approach is that a change to a word that has been more extensively processed will be more detectable than a change to a more shallowly processed word. Results showed that small semantic changes, e.g. changing *finished* to *completed*, were less detectable than larger semantic changes, such as if *finished* was changed to *started*. However, small distance changes were more detectable if they were placed in a prominent position, such as a main clause, which suggests that certain sections of text (i.e. those in a focussed position) will be more fully, or more deeply processed, than other words in the sentence. These results challenge the assumption that all words are processed equally, and that all changes are equally detectable. Examples of change detection are also discussed more fully in later parts of this review.

Examples of incomplete semantic commitment: ambiguous nouns, metonymy, aspectual coercion, and anaphoric reference

Evidence from studies involving ambiguous noun resolution, metonymy, and more recently aspectual coercion, demonstrate that at least on some occasions processing can proceed without immediate interpretation of some expressions. Delayed interpretation is the equivalent of a temporarily underspecified representation of the message.

Ambiguous nouns

Frazier & Rayner (1990) contrasted situations when readers made an immediate commitment to a word's interpretation, to situations when commitment may be delayed. They used eye tracking to demonstrate that when faced with an ambiguous noun, such as *bank*, which could refer to either a savings bank or river bank, readers attempted to assign word meaning as early as possible, this being in line with the dominant meaning. If the subsequent context supported the subordinate meaning, however, processing was disrupted as evidenced by an increase in reading time. While homonymous words, such as *bank*, with distinctly different meanings, seemed to require immediate commitment, polysemous words, that is words with more closely related meanings, for example, *newspaper* which may be used to refer to either the object or the institution, required no such commitment (as evidenced by eye measures when the subsequent context supported one interpretation over another). These results demonstrate that in some situations immediate commitment is necessary, whereas in other situations, minimal (or incomplete) commitment is acceptable.

Metonymy

Similar results have been found with metonymic expressions, such as *Vietnam*. Words such as these may be used literally (to refer to the country), or metonymically (to refer to Vietnam war). Frisson & Pickering (1999) used well-established metonymic expressions such as *Vietnam* and *convent* (which may be used to refer to the building or the institution), and found no difference in the eye-record data between literal and metonymic meanings, which suggests that with these types of word, readers did not need to make a full commitment to the word's meaning.

Unbounded events

Underspecification has also been demonstrated for bounded and unbounded events by Pickering, McElree, Frisson, Chen & Traxler (2006). Pickering et al., argued that readers sometimes make minimal commitments when making decisions on the temporal properties of events. They used both self-paced reading and eye tracking paradigms to present sentences containing either bounded or unbounded verbs, for example [5a and 5b].

> [5a] The insect glided effortlessly until it reached the garden (unbounded event)
>
> [5b] The insect hopped effortlessly until it reached the garden (bounded event)

While *hopped* is a discrete event, *gliding* can carry on indefinitely. Readers found that participants thought both sentences made sense, and that there was no evidence for any processing difficulty in [5b] when they realised that the event (*hopping*) was carrying on

(*until*…)[1]. They argued that this provided further evidence for underspecification because readers have failed to fully represent the bounded (or telic) meaning of the verb *hopped.*

The lack of an effect between conditions, Pickering et al. argue, could not be explained by experimental insensitivity. To support this claim they cite experiments where a different type of semantic coercion, complement coercion, was employed and where the eye movement record data supported full and early interpretation. To illustrate this we can compare, [5c] and [5d].

[5c] The author began the book …

[5d] The author read the book …

The verb *began* requires an event to complement it, however, the noun phrase *the book* refers to an entity rather than an event. This coerces the reader into interpreting the phrase as an event, such as writing or reading the book. When sentences such as these are compared to control sentences such as [5d] the eye movement record shows immediate disruption, consistent with full and early interpretation (Pickering, Traxler & McElree 2005; Traxler, McElree, Williams & Pickering 2005). These differences in the results between aspectual and complement coercion, they argue, suggest that the lack of results found in the case of aspectual coercion is not due to a lack of experimental sensitivity, but due to the fact that readers are in fact underspecifying the telicity of events.

[1] These results contrast with earlier studies supporting early commitment to aspect (Piñango, M. M., Zurif, E., & Jackendoff, R. (1999); Todorova, M., Straub, K., Badecker, W., & Frank, R. (2000)). However Pickering et al. (2005) argue that the methodology these studies employed (either a lexical decision or stop-making-sense concurrent task to reading) actually induced readers to commit to a full interpretation of aspect, and did not reflect what would occur normal circumstances.

Anaphoric reference

In a similar vein, underspecification has also been shown to occur when readers attempt to establish anaphoric reference. Sanford & Sturt (2002) argued that in some situations the antecedents of some anaphors are not fully defined and this ambiguity does not necessarily pose a problem for readers. To illustrate this they gave the example of [6]

[6] Mary bought a brand new Hitachi radio. It was in Selfridge's window.

They argue that there is an ambiguity in what *it* actually refers to. It could be that Mary bought that specific radio in the window, or it could just refer to the type of radio she bought. Either case is possible, but from the text what actually happened is not fully specified.

Koh, Sanford, Clifton & Dawydiak (in press) demonstrated that plural anaphoric reference may, in some circumstances, also be underspecified. They investigated the "conjunction cost" which is the processing difficulty observed when a singular pronoun is used to refer to one member of a pair introduced in a conjoined noun phrase, as in [7]

[7] Last night John and Mary went to an Italian restaurant.

[7a] They really enjoyed the food.

[7b] He really enjoyed the food.

When two characters are conjoined in such a way [7], there is a preference for a subsequent plural pronoun [7a] rather than a singular pronoun [7b]. The use of a singular pronoun results in longer reading times in self-paced reading studies, and in eye-tracking studies, with disruption occurring either at the pronoun or just following it (Albrecht & Clifton, 1998; Moxey, Sanford, Sturt & Morrow, 2004). This is the

conjunction cost. Koh et al., argue that one explanation for this cost is that the use of a singular pronoun indicates a shift in thematic subject. So, a story originally about two individuals in a common role, is changed to one about two individuals playing different roles. There are some actions, however, that may be carried out by just one of the individuals on behalf of *both* of them, so that the common role is not divided. Such actions, they hypothesised, would not result in a conjunction cost. For example, in [7b] above, the use of a singular pronoun distinguishes John from Mary (and leads to the possible inference that she did not enjoy the meal as much as he did), whereas in [7c] asking for a table is an action performed on behalf of both parties.

[7c] He asked for a table.

Such actions are described as 'number-indifferent' actions by Koh et al. because the action preserves a common role, and for the purposes of the story, it does not matter which character performs the action. In a self-paced reading time study (Experiment 1) where participants read scenarios that involved number-sensitive and number-indifferent actions, Koh et al. reported the expected conjunction costs for actions such as enjoying food that signalled out one individual [7b], but no such cost with actions that were assumed to be carried out on behalf of both parties [7c]. Furthermore, using a text change detection paradigm (Experiment 4), where the plural pronoun *They* was changed to a singular pronoun (*he* or *she*) in the second presentation, they reported lower rates of detection in number-indifferent [7c] compared to number-sensitive [as in 7a changing to 7b] actions (rates of detection were 38.7% vs 44% respectively). Koh et al. argued that in scenarios containing 'number-indifferent' actions there is no need to discriminate whether one or both characters performed the action, therefore there not will be a conjunction cost. If there is no specification of who performed the action in

any detail the representation of this action is underspecified and therefore in a change blindness paradigm fewer changes will be detected.

Reference to parts-and-wholes

A somewhat different situation provided some further evidence of underspecified pronominal reference. Poesio, Sturt Artstein, & Filik (2006) investigated what they termed mereological cases of anaphoric references, which they argued are frequently used in an underspecified manner. Furthermore, as in the previous examples, underspecification does not appear to result in any comprehension difficulties. A mereological pronoun is a pronoun which is used to refer to an object that is made up of more than one entity, and the pronoun may refer to the whole object or the individual parts. For example, *it* in [8], could refer to the engine, the boxcar, or the whole train.

[8] The engineer hooked up the engine to the boxcar and sent it to London.

This 'merged' use of pronouns is likened to the ploysemous nouns reported by Frazier & Rayner (1990) because they can be interpreted in respect of both antecedents (in the same way V*ietnam* may refer to the country or war without commitment to either sense), and this is also similar to the number-indifferent scenarios that licence the use of either singular or plural pronouns. Poseio et al. report an analysis of naturally occurring dialogue from the TRAINS corpus collected at the University of Rochester (Gross, Allen, & Traum 1993). They asked naïve participants to identify mereological expressions naturally occurring within this corpus. They reported that participants had little difficulty in identifying this type of ambiguous pronoun use, and that it was relatively common. They also observed that interpretation of pronouns such as *it* in [8] was not consistently interpreted in one way (that is, some reported it as referring to the engine, others to the boxcar, and others to both). To investigate whether or not these

types of expressions incur any processing costs Filik, Sanford & Sturt (2005) reported the experimental results from eye-tracking and change detection experiments on the online processing of mereological expressions. They contrasted the use of singular (*it*) and plural (*them*) pronouns in mereological statements to neutral statements where the individual items (engine and boxcar) are referred to in a conjoined phrase, but not necessarily joined together, as in [8a].

[8a] The railwayman saw the engine and the boxcar and sent it to London

Since the railwayman in [8a] only saw the engine and the boxcar they are still individual objects and not necessarily unified into a train. They found that the use of the singular pronoun *it* caused difficulty when it was forced to refer to a single conjunct in a coordinated noun phrase (the conjunction cost) in [8a], but no difficulty was observed when the plural pronoun *them* was used. In contrast, little difficulty was observed in merelogical statements such as [8] with both singular and plural pronouns, and both caused less disruption than *it* in [8a]. In a second experiment, this time using text change detection, they used items similar to [8 and 8a], and reported that participants were less likely to notice that the plural pronoun *them* changed to the singular *it* in merelogical cases. They argued that the failure to notice the change in pronouns reflected that the antecedent was underspecified in mereological cases and hence fewer changes were detected. Both experiments supported the interpretation that pronoun use in these cases is underspecified because both the singular or plural pronoun could refer equally to either antecedent (the engine or the boxcar) or both.

These studies all demonstrate the wider significance of shallow processing and show that examples of underspecification can be found in many different situations. In all these examples, the arguments put forward are that occasional underspecification is the

norm, rather than due to the result of a "failure" to process properly due to lack of attention or limited capacity.

Garden path sentences and semantic inertia

A somewhat different line of argument for shallow processing has been made by Christianson and colleagues. The syntactic representation of a sentence is presumed to be computed incrementally (e.g. Altmann & Steedman, 1988; Frazier, 1979; Frazier & Rayner, 1982; Kamide, Altmann & Haywood, 2003; Marslen-Wilson, 1973; Pickering & Traxler, 1998; Sedivy, Tannenhaus, Chambers, & Carlson, 1999; Sturt & Lombardo, 2005) . Evidence for this claim has been provided from reading studies using 'garden path' sentences. For example, when readers reach the word *fell* in, "The horse raced past the barn fell" (Bever 1970), they realise that they have miss-parsed the sentence and that the verb *fell* is in fact the main verb of the sentence. *Raced* which had originally been taken as the main verb, has to be re-parsed as the past participle of a reduced relative clause (and so the sentence in full should read, "The horse *that was raced* past the barn fell"). Garden path sentences disrupt the flow of reading and normally trigger a reanalysis of the sentence (Frazier & Rayner 1982). Because this happens so quickly the assumption has been made that re-analysis is carried out fully and leads to a final correct interpretation of the sentence. However, Christianson, Hollingworth, Halliwell & Ferreira (2001) have demonstrated that the initial misinterpretation of a garden-path sentence may persist, even after a seemingly correct re-interpretation had been achieved. To demonstrate this they presented their participants with garden path sentences such as [9]

[9] While Anna dressed the baby played in the crib.

In sentences like these, readers normally assume that “the baby” is the direct object of “dressed”, however when they reach the second verb, “played”, they realise they have misunderstood the sentence and are forced to reanalyse it. The assumption is that the reanalysis is carried out accurately. Readers were allowed to read these sentences at their own pace and then were asked to answer one of two questions:

[9a] Did the baby play in the crib?

[9b] Did Anna dress the baby?

After reading sentences such as [9], participants were asked questions to gauge their final representation of their meaning. For example, [9a] was used to assess whether the original misanalysis of the “the baby” (as the direct object of “dressed”) has been successfully restructured as the subject of “played”; and question [9b] was used to assess whether readers had altered their understanding of the sentence as a whole, so that they understood that Anna was actually dressing herself and not the baby. They compared how accurately participants were in answering these questions when they had read the sentences in either garden-path or non-garden-path versions. Their results showed that participants had no problem answering [9a] correctly, and understood that the baby was playing in the crib, following non-garden path sentences. However, participants who had read the garden-path version were more likely to incorrectly respond “yes” to [9b]. The authors conclude that participants’ original misanalysis, which was that Anna was dressing the baby, persisted even though they had subsequently successfully restructured “the baby” as the subject of “played”. The process of reanalysis, therefore, must be incomplete and can be taken as yet further of evidence of a language system that tolerates shallow processing in appropriate situations.

Pragmatic normalization

Some of the earliest evidence for local semantic processes being over-ridden by global processing comes from pragmatic normalisation. An early illustration of this was provided by Fillenbaum (1974) who reported that pragmatically unusual sentences, such as *The dog was chased by the cat, John got dressed and had a bath,* or *Get a move on or you will catch the bus*, were consistently "normalized" in line with real-world expectations when participants were requested to paraphrase them. More recently, Ferreira (2003) demonstrated that readers sometimes use world-knowledge to interpret passive sentences. Passive sentences are syntactically challenging for readers because the thematic roles are in an atypical order. Ferreira asked participants to identify the agent or patient of an event in a sentence, or to make plausibility judgments of the likelihood of the event, in sentences such as [10a] and [10b]:

[10a] The dog bit the man. (active)

[10b] The dog was bitten by the man. (passive)

Readers made far more mistakes when answering questions about a sentence with a passive construction than an active one. They were more likely to understand "who-did-what-to-whom" in light of pragmatic information rather than on an accurate analysis.

While Ferreira's materials used non-canonical sentences such as passive constructions, there are also examples of readers not fully analysing the syntactic relations of sentence elements with seemingly easier sentences. This has been demonstrated with what has become known as the *depth-charge* sentences reported by Wason & Reich (1979). They asked participants to paraphrase statements such as [11a] and contrasted the interpretation of these sentences to similarly structured sentences such as [11b].

[11a] No head injury is too trivial to be ignored

[11b] No missile is too small to be banned.

The meaning of [11b] is clearly that all missiles should be banned, but by the same token, it would follow in [11a] that this sentence actually means that all head injuries should be *ignored*. This presumably is not what would be intended from a common-sense perspective. When participants were asked to paraphrase sentences such as [11a] they often did so in line with the common-sense interpretation that head injuries should be taken seriously. This suggests that instead of carefully parsing these sentences, and utilising local semantics, readers are interpreting its meaning based on pragmatic knowledge. Thus the influence of pragmatics and situation-specific knowledge may override a full local semantic interpretation of a message in some situations.

Interim conclusions

Taken together these studies suggest that the meanings of words may only be partially recovered, that the integration of word meaning and referential structure between sentence elements may be underspecified, that the parsing of a sentence may be incomplete, and that world-knowledge may be used to pragmatically interpret a sentence rather than a full syntactic and semantic analysis of the text. These findings also suggest that shallow processing may produce an underspecified or possibly a *good enough* representation rather than an exhaustive analysis (Ferreira, Bailey & Ferraro, 2002; Sanford & Sturt, 2002). However, the most consistent and striking demonstration of shallow processing is when readers fail to notice semantic anomalies in text. Semantic anomalies were initially treated as curiosities in the literature, at best demonstrating a failure of the memory system to fully retrieve information. However,

semantic anomalies are an important example of shallow processing and potentially offer important insights into the dynamic processes of language comprehension.

Semantic anomalies

The topic of this thesis centres on the on-line processing of semantic anomalies, and this section provides a brief review of research in this area. A semantic anomaly is the use of an expression with an inappropriate meaning. The term "semantic illusion", coined by Erickson & Mattson (1981), occurs when readers or listeners manage to construct a seemingly coherent representation of a message, even though the discourse contains a semantically inappropriate word which, if taken literally, renders the whole message meaningless or flawed in a serious way. The ease with which readers can detect anomalies can vary considerably. Some anomalies are easy to detect because they share neither semantic similarity with the "correct" word, nor have any relationship to the context. So, for example, in the sentence, "He spread his warm bread with socks" (Kutas & Hillyard 1980), readers quickly detect *socks* as anomalous. The word, *socks,* shares no semantic features with a potentially correct target word, such as *butter*, nor does it have an obvious relationship to a context about eating. *Socks* is therefore an easy-to-detect thematic violation within such a sentence. However, other semantic violations are much harder to detect. These violations occur when words are again semantically inappropriate, but these anomalies may share some semantic features with the "correct" word, and they may also have a strong relationship to the overall context. What makes these harder-to-detect semantic anomalies so interesting is that readers often fail to detect the anomalous word yet at the same time manage to construct a coherent representation of the message, this being achieved as if the correct word had actually been used. Two of the best known examples of hard-to-detect anomalies are:

Erickson & Mattson's (1981) "Moses illusion" [12], and Barton & Sanford's (1993) "Survivors problem" [13].

[12] How many animals of each kind did Moses take on the ark?

[13] Where should the survivors be buried?

The problem with these questions is that, in the Moses illusion, it was of course Noah who built the ark, and with the survivor's problem, you do not bury people who are alive! In both examples, readers manage to construct a meaningful coherent representation of the questions, and even provide appropriate answers to them, yet if the anomalous word had been correctly detected and processed, the questions would be meaningless. Non-detection of semantic anomalies such as these suggests that readers do not carry out an exhaustive lexical semantic analysis of all words in a sentence and hence is strong evidence for shallow processing. These results, therefore, present a challenge to models of language processing. There have been attempts to explain semantic anomalies in terms of a failure of the memory process, be that inadequate encoding, retrieval or matching of features. Others have focussed on the dynamic aspects of text coherence and investigated the role of scenarios, task demands, and focus in guiding the extent of depth of processing. A wide range of methodologies have been employed, including paper and pen verification/question-answering tasks, reading time, eye-tracking and ERP paradigms. In this section we review the main results reported in the literature, relating them to explanations of failures to detect, before turning to factors that may modulate detection rates.

Explanations of the effects

We will consider four explanations for the non-detection of semantic anomalies. These are that:

(a) The information is not encoded initially.

(b) The anomaly is detected but then 'corrected'.

(c) The semantic features of the anomalous word are not fully compared with the 'correct' word.

(d) There is a strong semantic relationship between the anomalous word and the context.

Are anomalies missed due to a lack of encoding?

One of the simplest explanations for the non-detection of a semantic anomaly is that the anomalous word has not been encoded al all. This position is easily dismissed. Erickson & Mattson (1981) investigated this hypothesis by simply requesting their participants to read their materials out-loud. As is common with all of this research, an anomaly was counted as missed only if the participants indicated in a post-test knowledge check that they knew what the correct term should have been. They used four experimental items including the Moses illusion. While these four items varied in their rates of non-detection, with the Moses illusion itself giving the highest rates of non-detection at 81%, all items were frequently missed by participants. This was despite the fact that participants had clearly spoken the anomalous word out-loud. However, this was a relatively unsophisticated manner of guaranteeing encoding and a better alternative has been to inspect reading times.

Van Oostendorp & de Mul (1990) inspected reading times for detected and non-detected anomalies and found no evidence to support the lack of encoding hypothesis. Non-detected anomalies took longer to accept as truthful in a verification task compared to detected anomalies that were rejected as untruthful. This they argued, was evidence against the lack of encoding hypothesis because when anomalies went undetected

participants were taking longer to read the sentence compared to when they read the sentence and detected the anomalies, whereas a lack of encoding would suggest shorter reading times.

Similar results were reported by, Reder & Kusbit (1991) who used a self-paced reading paradigm and reported that the time spent reading words that were reported as distorted (anomalous) was faster than those than those that were missed. This again suggests that the failure to detect an anomaly is not due to insufficient time for reading and hence encoding the anomalous word. The results from all three studies suggest that readers spent time to encode the information because they either spoke the word aloud, or that reading times were longer when anomalies were missed compared to detected anomalies. No data using eye-tracking has been obtained to date, and will be the subject of our own work.

Do readers 'correct' detected anomalies?

Another possibility is that readers actually do detect semantic anomalies, but quickly correct the anomaly to what they think the communicator really intended (this is sometimes referred to as the Conversational postulate based on the work of Grice 1975). If this were the case, however, then when participants are forewarned that semantic anomalies will be in the text, and are explicitly requested to report them (for example in Erickson & Mattson 1981), all anomalies should be detected. This does not occur. Also, when participants have had the anomaly pointed out to them during subsequent experimental debriefing stages, they are reported as expressing genuine surprise at their failure to notice the anomalies (Barton & Sanford 1993). Furthermore, if readers were correcting anomalies, this would also suggest that detection is in some way under the control of the reader. Reder & Kusbit (1991) clearly demonstrated that this was not the case by manipulating the task instructions given to participants. The experimental

instructions either stressed accuracy in reporting detected anomalies (*literal condition),* or they were instructed to answer the question and ignore any anomalies (*gist condition*). They reported that not only did participants find it difficult to detect anomalies, searching for anomalies in the literal condition increased the number of errors they made. Their results demonstrated two points. First, detection was not under the participant's control, and second, participants were not 'politely' ignoring and correcting detected anomalies. These results are mirrored in other studies where participants were fore-warned or explicitly asked to report anomalies (for example, Van Oostendorp & Kok 1990; van Jaarsveld, Dijkstra, & Hermans 1997; Büttner 2007). There is no evidence, therefore, for pragmatically-driven acceptance of semantic anomalies.

Non-detection due to shared semantic features

Erickson & Mattson (1981) argued that the locus of the Moses illusion lay in the high semantic similarity between *Noah* and *Moses*. They demonstrated this by varying the semantic similarity between the anomalous target words. So, *Moses* shares a high degree of semantic similarity with *Noah* (both were Old Testament characters, they received messages from God, and both were leaders), whereas a non-biblical name, such as *Nixon*, has none of these shared features. Using this, and three other illusions, they manipulated the anomalous word based on shared semantic similarity. Their results clearly showed that when the target anomalous word was replaced with one that shared few semantic features, anomaly detection was virtually 100%.

This issue was further explored by Van Oostendorp & De Mul (1990). They used terms which had been empirically defined as semantically related or not and asked participants to verify statements that might contain a semantic anomaly. They reported that highly related words produced larger illusions (i.e. Moses and Noah) than low

related names (Moses and Adam). The average susceptibility to the illusions was 29% in the high-related and 16% in the low-related condition, which clearly shows an effect of semantic similarity. They argued that the apparent cohesion of the representation of a sentence is partly based on the semantic relatedness between words. If the relationship is high, then words will only be processed shallowly, if the relationship is low, then they will be processed more extensively.

Van Oostendorp & Kok (1990) further demonstrated the importance that relationship strength has on the extent of semantic illusion. They strengthened the relationship between pairs of words by using a paired-associate learning task before the anomaly detection stage (e.g. learning to associate either Moses or Adam with Ark or Animals) to increased the rates of the illusion (from 17% in the non-paired-associate task to 32% when in the paired task for low related names, and 30% to 44% in the high related condition). These results suggested not only that greater overlap of attributes between two names leads to increased likelihood of confusion, but also that the stronger the relation is between names and concepts associated with the correct target, the higher the rate of non-detection.

Van Oostendorp & de Mul (1990) also collected reading time data in their study (their materials were presented as whole sentences on a computer screen and participants responded via a button box to indicate 'true', 'false', 'don't know'). They reported that when participants detected the anomalies, reaction time was longer when the anomalous target was highly related to the correct word, which shows that making a correct anomaly detect decision is harder when the words are highly semantically related. However, when anomalies were missed (in their experiment that was when participants incorrectly judged statements as being true when in fact they contained an anomalous word), there was no difference between high and low semantic similarity conditions, in

respect of judgement time. They also observed that reaction time in this situation was longer compared to successful detection. However they discount this as being of no interest. This is a shame because it is crucial to ask why readers fail to detect an anomaly, and the two observations that semantic similarity has little effect on judgement times, and judgement times are longer when readers miss anomalies, are both relevant to this question. The longer judgment times might suggest that these conditions are being processed quite deeply, but for whatever reason deeper processing does not result in detection. Secondly, whatever processing is occurring, it is not influenced by semantic similarity in the same way, which questions the nature of this processing. These are themes which will be explored in more detail later.

Van Oostendorp and colleagues discuss the importance of semantic similarity and relationship strength in anomaly detection, and they argue that these factors influence language processing within working memory only. Lynne Reder and her colleagues (Reder & Cleermans 1990; Reder & Kusbit 1991; Kamas, Reder & Ayers 1996) similarly argued that lexical semantic features are important, however they believe that the semantic illusion occurs because of a failure to match semantic features of the anomalous word in working memory with the semantic features of the "correct" word in long term memory. They call this 'partial mismatching'. Reder & Cleermans (1990) demonstrated that this was not simply a failure to retrieve the relevant information from long term memory by priming participants' knowledge relevant to their experimental items. In an initial study phase participants simply read factual statements containing the non-anomalous terms for experimental items. They were then presented with experimental items that were judged for truthfulness, and some of these contained anomalous words which they were asked to report. Their results showed that priming relevant information, and hence improving memory accessibility, did not improve overall detection rates. Priming did, however, affect error rates in that the speed of

correct responding was increased, and there was a reduction in the rate of incorrect (or, 'don't know') responses. They concluded that failure to detect semantic anomalies occurred because the semantic features of the anomalous word and of the correct word retrieved from long term memory were only partially compared. When there were enough shared semantic features an anomaly may go unnoticed.

This was further investigated by Kamas, Reder & Ayers (1996) who used a similar priming method as used by Reder & Cleermans (1990) and Reder & Kusbit (1991) to improve memory accessibility. However, this time the primed semantic features could either be features that were shared, dissimilar or irrelevant to the experimental question. So, for example, a question that primed shared features for Noah and Moses was "What religions study the story of Moses?", and a question that primed the differences between them was, "What sea did Moses part?" These questions preceded the presentation of the anomalous question. They found that there was no difference in detection rates when semantically similar features had been primed (59% successful detection) and an irrelevant question (58% successful detection), however, priming semantic features that distinguished the two terms did significantly increased detection (70% detection). They concluded that anomalies are missed in many situations because a partial matching process detects enough shared features in *Noah* and *Moses* to pass inspection. When there are fewer or no shared features, such as between *Noah* and *Nixon*, this would result in a more detailed lexical analysis of the anomalous name. Therefore, in their priming studies, improving accessibility to shared semantic features fails to uncover the anomaly in a processing system that relies on partial matching and is tolerant of occasional mismatches. However, priming that activates features emphasising the differences between names increases over all detection.

These studies suggest, therefore, that semantic illusions are, at least in part, due to the shared semantic features, or a strong semantic relationship, between the anomalous and the correct word. Illusions occur, according to Reder and colleagues, due to partial matching of shared semantic features.

Global fit theory

A different possibility is explored by Barton & Sanford (1993), who argue that semantic anomalies may go unnoticed because the anomalous word has a strong fit to the current context (scenario). They suggest that how well the word fits this context affects the extent of processing the word will receive. Their theoretical account draws on the Scenario Mapping and Focus theory of language comprehension (Sanford & Garrod 1981; 1998). In this theory it is argued that a primary process in language processing is to establish a coherent representation of a sentence using long term memories for situations and events to interpret new linguistic information. Therefore, to comprehend a piece of text the reader maps the message onto situation-specific knowledge, or scenarios, which are akin to schemas or scripts, as quickly as possible. This primary process, mapping incoming linguistic information onto scenarios, occurs *before* interpretation. The scenario remains in implicit focus, while it is relevant, and helps to interpret the actions and consequences of characters and entities within the explicit focus of the story. Interpretation, therefore, is open to influence from the activated background scenario. This process is also assisted by producers of language who try to manipulate the relevant background information, and may use appropriate linguistic devices to guide attention, such as focus and emphasis. Evidence supporting this latter conclusion will be discussed in a later section.

Sanford & Garrod (1998) argued that when a word is first encountered it is checked in terms of its relevance to the current context. They (and Barton & Sanford 1993) argue

that this is a *fast passive process*, and is simply a statistical-type test of the word's fit to the context. If the relevance of the word is high then the word may receive extra processing (especially if it is in focus) or it may not. However, if the word does not pass this simple test of association it is likely to be detected as being "out of context", and is therefore likely to be attended to and processed more deeply. This means that the context of the story will affect the detection of an anomaly, and also that the amount of processing that the word receives is modulated by the context. As a consequence of this, the global context can over-ride the local semantics.

Evidence to support these claims was provided by Barton & Sanford (1993). They used one very reliable semantic anomaly in their investigations, which was presented to readers under naturalistic reading conditions. The 'survivors problem' is presented below [14]

> [14] There was a tourist flight travelling from Vienna to Barcelona. On the last leg of the journey, it developed engine trouble. Over the Pyrenees, the pilot started to lose control. The plane eventually crashed right on the border. Wreckage was equally strewn in France and Spain. The authorities were trying to decide where to bury the survivors.
>
> Where should the survivors be buried?

In this example, participants often fail to notice that the word *survivors* is inappropriate because it means that people are alive and therefore should not be buried. Participants were presented with the scenario and asked to provide an answer to the question. Directly afterward they underwent a detailed debriefing. Detection of the inappropriate use of the word *survivors* was established through the answer to the question (i.e. if they mentioned the anomaly as part of their answer) or during the debriefing session if they indicated that they had noticed the anomaly.

To illustrate the influence of scenario, Barton & Sanford manipulated the background scenario which could be an airplane crash or a bicycle crash (Experiment 3). While the term *survivors* is highly relevant in a disaster situation such as an airplane crash, it has less relevance to a more mundane situation of two cyclists crashing into each other. When the final question, "where should the survivors be buried?" was asked, detection rates increased from 33% for the airplane crash scenario to 80% for the bicycle crash scenario. On the basis of this, they argued that the first step in mapping new linguistic material on to a background scenario is based on a simple test of fit to the situation. Words that have a good fit (and are therefore highly relevant) will fit the global context well, such as *survivors* in an airplane crash. As such, these words are likely to be processed less deeply than words that do not fit the overall context, unless the word is focussed or emphasised in some way. However, words that do not fit the context well, such as *survivors* in a bicycle crash, will receive more attention and be processed more deeply and so are more likely to be detected. This makes sense because words which clearly do not belong to a situation will require more processing.

Also, whether or not detection takes place depends on how easy it is to retrieve anomaly-relevant core meanings for the word. So, for example, scenario-relevant words for people who have survived a disaster situation may include, *injured*, *wounded*, *maimed*, as well as *survivors*. These words may be similar and may be used in similar contexts, however they have different core meanings. This was established by Barton & Sanford who pre-tested these materials by asking participants to provide definitions for these 4 words. They were also presented with one of these words within the survivors question and asked to describe whether the writer meant that the individuals were alive at the time of writing, which was used to assess whether being alive was a presupposed meaning in the words. Responses showed that it was only the word *survivors* that was defined by participants as having a core meaning of, *being alive*. Whereas with *injured*,

wounded, and *maimed*, being alive was only presupposed (indicated by responses to how the words would be used) but was not a core meaning provided in participants' definitions.

In Barton & Sanford's (1993) Experiment 1, the rates of anomaly detection for *injured*, *wounded*, *maimed*, and *survivors*, were compared in an incidental anomaly detection paradigm as outlined above. Detection rates were lowest for the three injured terms (there were no differences between these terms), and was highest for the word, *survivors* (see figure 1.1). This demonstrates, they argued, that words that fit the global context may only be processed shallowly. However, if the word has as its core meaning information relevant to anomaly detection, then it is more likely to be detected.

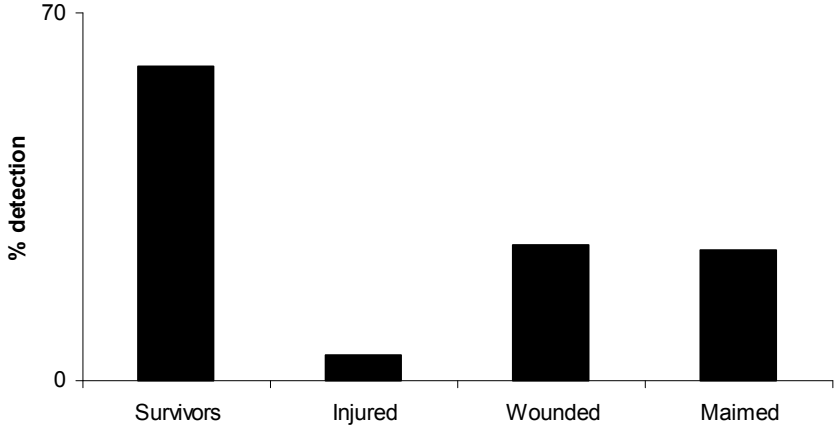


Figure 1.1: Rates of detection (%) comparing different noun phrases in the survivors problem from Barton & Sanford (1993), Experiment 1 (reprinted with permission)

Barton & Sanford also demonstrated that the scenario may even override careful analysis of local noun phrases, further challenging assumptions of compositionality. They presented the air-crash scenario with the qualifying noun phrase, *surviving dead*. This internally incoherent noun phrase was detected much less frequently than the basic term *survivors* (detection rates dropped from 66% to 23%). They argued that the word

dead was such a good fit to the overall context that this suppressed further analysis, leading to decreased detection rates. These results have been replicated by Daneman, Lenneretz, & Hannon, (2006), and Hannon & Daneman, 2001.

In sum, these results suggest that readers use scenario-based expectations to guide the comprehension process. If a word or phrase fits the global context it may only receive cursory analysis, unless the word is either focussed or emphasised. However, if a word does not fit the context analysis may be more extensive, and this may lead to detection. Whether or not detection occurs is also partly determined by how easy it is to retrieve anomaly relevant information from the word's core meaning. If information that is highly relevant to the anomaly is easily accessible, then detection is also more likely.

Factors influencing anomaly detection and depth of processing

Are there any other factors which influence the likelihood of anomaly detection? While shared semantic features and the overall context are important, it seems reasonable to expect that other factors may also modulate anomaly detection. In this section we will consider four possible factors that may do this; linguistic focus or emphasis; cognitive load; experimental task demands; and, individual differences.

Focus

One obvious candidate for modulating depth of processing is linguistic focus. This is because focus serves to highlight some information in a sentence as particularly relevant, over other information. Erickson & Mattson (1981) originally investigated the impact that focus devices have on detection rates. They converted their original materials, where anomalies were placed within questions, into statements, as in [15].

This, they argued, would reduce any focussing effects caused by a question-answering paradigm. Participants were required to verify statements for truthfulness.

[15] Moses took two animals of each kind on the Ark.

They concluded that since anomalies were not always detected in such statements like [15], focus is not the cause of anomaly non-detection. However, they do not make any direct comparisons between detection rates for statements and questions. Based on their reported descriptive statistics, the average rate of non-detection for statements was 26.5%, and for the two question-focussed versions they were 52% (Experiment 1) and 48% (Experiment 3). This difference appears to support a focus effect in rates of anomaly detection.

Baker & Wagner (1987) criticised Erickson & Mattson's focus manipulation on the grounds that it did not sufficiently signal what information was focal and what was pre-supposed. Instead they employed logical subordination as a way to clearly distinguish focal information from 'extra' information. They also tested their materials in both an auditory and text based version (the text version was not time restricted and target sentences were placed within larger bodies of text). Participants were explicitly requested to detect and report false information. Their materials concerned common facts and contained false information in either the main clause, as in [16a]

[16a] The Emerald City, the home of the Wizard of Oz, was named after the precious *red* stone.

or in the subordinate clause, as in [16b]

[16b] The Emerald City, named after the precious *red* stone, was the home of the Wizard of Oz.

Their participants detected more errors in the main clause (89%) than they did if the information was placed in the subordinate clause (81%). These results suggest that focussing serves as a guide to the important aspect of a message. As a consequence of focus, this information is likely to be more extensively processed, hence increasing the likelihood of anomaly detection.

Further work that has demonstrated the effect of focus on anomaly detection has been carried out by Bredart & Modolo (1988) and Bredart & Doquier (1989). Bredart & Modolo employed clefting as a way of manipulating focus. They asked participants to make true / false judgements to statements that they had been forewarned might contain anomalous information. In, what they termed, the *narrow focus condition*, the anomalous name was placed within the cleft phrase of the sentence, whereas in the b*road focus condition,* other information was placed in the cleft phrase, as in [17a and 17b]:

> [17a] Narrow focus condition: It was Moses who took two animals of each kind on the Ark (narrow focus is on *Moses*)
>
> [17b] Broad focus condition: It was two animals of each kind that Moses took on the Ark. (narrow focus is on *two animals*)

Their results also clearly showed that when the anomalous phrase was placed in the cleft position participants were much more likely to detect it (90.8% vs 65.9%).

In a follow-up study Bredart & Doquier (1989) manipulated focus using typographical devices to focus attention on target words. This permitted a focus manipulation that did not alter the surface structure of the items. In their study they placed either the anomalous term or other information (focussed) in uppercase and underlined, for example [18a]

[18a] MOSES decided to take two animals of each kind on the Ark.

Or, other information (un-focussed) in uppercase and underlined, for example [18b]

[18b] Moses decided to take TWO animals of each kind on the Ark.

Again, the results were consistent with the effect of focus increasing detection, so that when Moses was in focus, the average detection rate was 86.5%, but was significantly less in the unfocussed condition, 68.3%.

Converging evidence for focus effects from other methodologies

Converging evidence for focus effects in modulating depth of processing comes from studies using text change detection. In this method, participants read a target passage twice, with a short delay between presentations, and are required to report any words that change across presentations. The rationale is that detecting a semantic change will indicate to what level of detail the word has been encoded into the discourse representation. Words that have been processed more deeply should be encoded in more detail, and so changes will be more detectable. Likewise, words that have been processed more shallowly are likely to be represented at a coarser grain of analysis. As such, words that are in focus should be processed more deeply than words not in focus, and so changes should be more detectable.

Using text change detection Sturt, Sanford, Stewart & Dawydiak (2004) demonstrated that placing a word in a focussed position (in this case they used cleft constructions) participants were more likely to report that a word had changed in the second presentation. They presented short passages, such as [19] followed by a target sentence

that placed one of two noun phrases in a pseudocleft construction (*Jamie* in 19a, or *cider* in 19b).

[19] Everyone had a good time at the pub. A group of friends had met up there for a stag night.

[19a] It was Jamie who really liked the cider, apparently. [Focus on Jamie]

[19b] What Jamie really liked was the cider, apparently. [Focus on cider]

Two changes were made to the target word *cider*. The change was either to another word with a similar semantic meaning (*beer*) or to a more distant semantic meaning (*drink*). They reported main effects for both focus and semantic distance. When *cider* was focussed [19b], participants were more likely to notice the word change. Also, if the change had involved a large semantic change (to the superordinate category, *drink*) it was more likely to be detected. They also reported a crucial interaction, which was that when a word is focussed, small semantic changes are more detectable (see figure 1.2).

In a second experiment they changed their focus manipulation so that a prior sentence focussed the reader's attention on some aspect of the subsequent sentence. A word in this second sentence would then be changed. An example is shown below. In [20a] the sentence focuses the reader on to the man, whereas in [20b] the focus is broader and on the general events. The target sentence [20c] was always identical and the word *hat* was changed to a semantically similar or dissimilar word (*cap* vs *dog*).

[20a] Everybody was wondering which man got into trouble (focussed)

[20b] Everybody was wondering what was going on that night. (unfocussed)

[20c] Target Sentence - In fact, the man with the (hat / cap / dog) was arrested

They reported similar results, with main effects for both focus and distance, and an interaction between focus and distance, such that close semantic changes were more detectable when in a focussed position.

To explain these interactions, Sturt et al. proposed the granularity theory. The granularity theory (developing the ideas of Hobbs 1985) proposes that the meaning of a word may be represented at differing levels of granularity. In some situations a word may have been more extensively processed and as a consequence it is represented more fully (i.e. a finer level of granularity). Focus is one factor which determines the level of granularity, so that a word that is in focus will be more extensively processed, and hence represented in more detail. For example, when *the man* is in focus, details about what type of headgear he happens to be wearing will be a critical detail and be processed more extensively. Whereas broad focus, in [14c], concerns the general events, and so *hat* is not a critical detail and is held at a coarser grain of representation. Such an account can explain the interaction reported in both of their experiments.

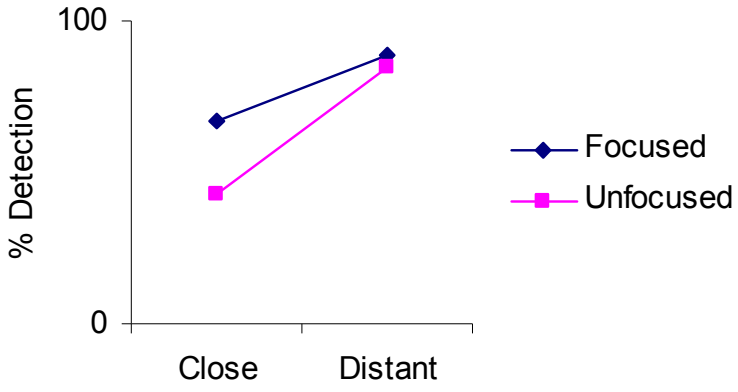


Figure 1.2: Illustration of the interaction found in focused and unfocused conditions with close and distant semantic changes found in Sturt et al. (2004)

Further demonstration of the effect of focus in a change detection paradigm was provided by Sanford, Sanford, Molle & Emmott (2006). They reported results from both text and auditory change detection studies (in the auditory version participants listened to two consecutive voices reading the same passage where one may change in the second reading). They used two 'attention capture devices' which served to guide attention selectively to the relevant points of the message. In the text-based version, they employed italicisation as an attention capturing device. This was chosen because italics are often used in texts to signify important or surprising information, and so may work in a similar way to the focus devices reported earlier. They reported main effects for italics (focus) and semantic distance, and in line with the granularity theory, an interaction between italics and distance (italicised words in the close semantic distance condition were detected more frequently when italicised than in the distant condition).

In their second experiment, they presented an auditory version of materials adapted from Sturt et al. (2004) where focus was manipulated via the prior context. In [21a] the initial sentence asks the implicit question *which money*? This question is answered in the second sentence, *the money from the wallet*. The focus here is on the wallet, and because the money in it is being discriminated from any other money, it is contrastive information, which leads to contrastive focus.

> [21a] Narrow focus:
> They wanted to know which money had been stolen.
> The money from the *wallet* had gone missing.
> Thefts in the area were becoming all too common.

In [21b] the implicit question is on, what happened? Because this refers to the general event depicted in the whole sentence this is referred to as, broad focus.

[21b] Broad focus:
They wanted to find out what had happened.
The money from the wallet had gone missing.
Thefts in the area were becoming all too common.

When information is in narrow focus speakers naturally use a specific type of pitch accent, which begins low and rises to a high pitch on the stressed syllable (this is described as L+H* using Pierrehumbert's (1980) system of notation). In [21a] stress would be placed on the word wallet. No comparable stress pattern would be observed for sentences in broad focus. Participants listened to short passages as in [21a,b] spoken consecutively by male and then female speakers, where the critical word (*wallet* in 21a) changed to either a semantically similar (*purse*) or dissimilar word (*bank*). When in the narrow focus condition this word was emphasised by the speaker, but not in the broad focus condition. They replicated the results of Sturt et al. (2004), so that when *wallet* was in focus through vocal stress, participants detected more changes. Also, if the change was to a semantically dissimilar word, it was more likely to be detected. Finally, they reported a similar interaction, between focus and semantic distance, so that when a word was in focus it had a bigger effect on detection rates for close semantic changes than distant semantic changes. The results for both experiments supported earlier predictions of the granularity theory. It appears, therefore, that words may be placed in focus, or signalled as important, by using a wide range of stylistic devices. Words that are singled out in this way receive more attention and more processing, and hence will be represented in more detail. In a task such as change detection this finer grain of representation results in changes being more detectable.

Load

A further interesting finding using change detection is the effect of processing load. Language processing is assumed to place a demand on working memory which has a

limited capacity system (e.g. Just & Carpenter's (1992) Capacity Theory of comprehension). If sentences are harder to process and require more processing effort, it could be expected that a reader's ability to detect changes will subsequently decrease as there are fewer resources available for detection. Some evidence relating to this hypothesis was provided by Glenberg, Wilkinson & Epstein (1982) who reported that the detection of contradictory information was greater in shorter texts (one paragraph) compared to a longer texts (three paragraphs), where overall length may be seen as analogous to processing difficulty.

More direct evidence comes from Sanford, Sanford, Filik & Molle (2005), who demonstrated that processing load could affect rates of text change detection. They manipulated load by using object-extracted relative clauses (high load) compared to subject-extracted relative clauses (low load). It is known that object extracted relative clauses are harder to understand than subject-extracted relative clauses, and that the difficulty in processing occurs on the embedded verb (see Gibson 1998). To investigate whether increasing cognitive load would result in a coarser grain of representation, as indexed by the ability to detect word changes, Sanford et al. changed the embedded verb *talked* to a semantically close (*spoke*) or distant (*listened*) alternative on the second presentation. See [22] for an example item.

> [22] There is an increasing demand for therapists and counsellors in many areas of modern life.
>
> The child who the psychologist talked to had hurt the woman. **High Load**
> The child who talked to the psychologist had hurt the woman. **Low Load**
> It is important that all victims receive a high standard of emotional support.

It was found that changes to the embedded verb in object-extracted relative clauses (high load) led to fewer detections compared to low load sentences. To investigate the impact of processing load on change detection, in two further experiments they

employed referential load as a technique to manipulate processing load, and measured subsequent rates of change detection (again with close and distant semantic changes). The increased processing cost associated with referential load has been demonstrated by Warren & Gibson (2002). They reported that there is a processing cost when using full noun phrases, compared to first or second person pronouns, which is reflected in longer reading times on embedded verbs. This difference in processing cost, they argued, is due to the "givenness" of prior referents in the discourse, and different expressions vary in respect of how accessible these referents are (Ariel 1990). Therefore, first and second pronouns are assumed to be highly accessible within a context, but full noun phrases may either introduce a new character or require a prior antecedent, both of which incur a processing cost to establish. Sanford et al. used materials as in [23] where the verb *met* was changed to either *seen* (close semantic change) or *missed* (distant change), where this was preceded by a full noun phrase, *the student*, (high load), or a pronoun, *I*, (low load).

> [23] The college frequently held social functions for visiting academics. The professor who {*the student / I*} had recently **met** (seen / missed) at the party was famous, but no one could figure out why.

Sanford et al. also measured rates of change detection for surrounding regions (prior adverb and subsequent noun phrase) and, as can be seen in figure 1.3, they found that load did indeed result in fewer changes being detected, and that the effect of load was localised at the embedded verb. They tentatively concluded that processing load may lead to a general dampening effect on the ability to detect because of a less accessible memory trace for the word. The effect of cognitive load has not been investigated in the anomaly detection literature but may be another factor in the modulation of semantic anomaly detection rates. Experiment 1 of this thesis examines this contention.

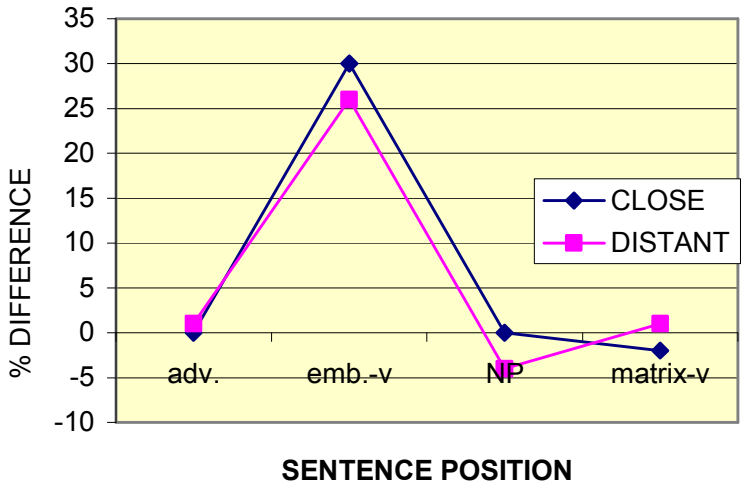


Figure 1.3: The effect of load (% difference between high and low) on rates of change detection per region of a sentence (reprinted with permission from Sanford et al. 2005)

Task demands

Some studies have specifically looked at the effect that different experimental tasks can have on overall anomaly detection results (Kamas, Reder & Ayers 1996; van Jaarsveld, Dijkstra & Hermans 1997; Büttner 2007). This work fits the idea that different tasks place differing demands on language use, so that language users strategically adopt different processing styles suitable for a current situation.

Van Jaarsveld, Dijkstra & Hermans (1997) compared anomaly detection rates when task instructions emphasised accuracy (with no time limit) to detection rates when instructions requested speed and accuracy. They found that the group who were instructed to respond accurately missed 18.3% of the illusions, whereas the group who were instructed to respond faster missed 32.9% of the illusions. They concluded that depth of processing was modulated by task requirements, and were therefore partly under participants' control.

A different conclusion was reached by Kamas, Reder & Ayers (1996) who argued that depth of processing, as reflected in anomaly detection, was not under participants' strategic control. They compared detection rates for semantic anomalies which were presented within a question, when participants were only required to detect anomalies (and to ignore the questions – this was the single task group) with participants who monitored for anomalies while also providing answers to the question (dual task group). They found that the single task group correctly identified more anomalies than the dual task group, but they also reported that the error rate increased as well. To investigate the effect of error rates they performed a signal detection analysis and based on these results they argued that the increased detection rates were not due to increased sensitivity, but merely a shift in response criterion. This indicated that depth of processing, and so anomaly detection, was not under strategic control of their participants.

Büttner (2007) investigated task demands by comparing anomaly detection rates in questions to those in statements. In Experiment 1 she compared detection rates for semantic anomalies that were embedded in either a question that participants answered, or in a statement that they judged for truthfulness. So, for example, in [24a] participants were required to provide the answer, *coat*, whereas in [24b], they judged the truthfulness of the statement. In both versions they were forewarned that anomalies may be in the text and that they should report them (in this example, Joseph has been replaced with the incorrect name, Jacob).

[24a] What many coloured garment was Jacob given by his father?

[24b] Jacob was given a coat of many colours by his father.

More semantic anomalies went unreported when presented in questions (47.5%) compared to statements (31.3%). One potential explanation for these differences, Büttner argued, was the different memory demands of questions (where the answer needs to be retrieved from long term memory) and statements (where all the items are present and only requires recognition check to be completed). To control for this, she contrasted statements with multiple choice questions. So, for example, in [25a,b], *Pacific Ocean* has replaced *Atlantic Ocean*.

> [25a] What famous ship tragically sunk in the Pacific Ocean after hitting an iceberg? (Titanic or Bismarck?)
>
> [25b] The famous ship Titanic tragically sank in the Pacific Ocean after hitting an iceberg. (True or False?)

The same effects were also observed with multiple choice questions (Experiment 2) leading to fewer detections with statements than with questions (43.9% vs. 29.9%). Therefore, she concluded that the different task demands required for answering questions or monitoring for truthfulness, modulates anomaly detection. Büttner argued that the lower rate of detection in the question-answering group may have been due to the increased difficulty of the dual task, or to a pragmatic obligation to answer a question that interfered with the monitoring task. However, further investigation into the role of task demands and instructions are required to clarify whether or not these do modulate anomaly detection. This is a theme which will be explored further later on in this thesis.

Individual differences

While these manipulations have shown that detecting semantic anomalies is influenced by semantic features, syntactic constructions of the sentence, and task demands, Hannon & Daneman (2001a) focussed on individual differences and ability to detect anomalies.

They manipulated semantic similarity between the correct target and an impostor word[2], and how strongly related the target words were to the current context. They argued that the ability to resist closely related impostor words would be related to how efficiently individuals could access and reason with knowledge stored in long term memory. They used a knowledge access test developed by Hannon & Daneman (2001b) and found that participants with a lower score on the knowledge access test found it harder to detect highly semantically related words. They also thought that the ability to integrate new information with a developing text representation would be related to working memory. They demonstrated that high working memory capacity was related to increased ability to detect anomalies when there was a strong supporting context.

Hannon & Daneman (2004) demonstrated that comprehension skill, as measured by the Nelson Denny, was related to anomaly detection. High scorers on the Nelson Denny seemed particularly adept at detecting locally anomalous noun phrases, such as *tranquilising stimulants*. Readers with lower reading skill scores seemed more influenced by the global scenario, rather than computing the semantics of individual noun phrases. These results were replicated by Daneman, Lennertz, & Hannon (2006) who eye-tracked their participants and compared the eye movement data for anomalies that were detected and missed, and compared both to a non-anomalous control version. There was no evidence from early measures of reading time for immediate detection. However, significant effects were found in late measures, with longer look back fixation time (second pass) to the control condition, and also a greater number of re-fixations on the anomalous phrase when the anomalies were detected compared to the control.

[2] Hannon and Daneman (2001a) use the term impostor to refer specifically to the influence of knowledge-based processes on anomaly detection. They used materials such as, *What passenger liner was tragically sunk by an iceberg in the Pacific ocean?*, where Pacific is the impostor term because it should read Atlantic Ocean. They argued that the failure to detect impostor terms was due to incomplete knowledge retrieval (that is, only accessing some of the stored semantic knowledge in long term memory), or partial matching of the semantic features of Pacific and Atlantic.

These results suggested that anomaly detection did not occur immediately, but was slightly delayed. Also, the non-result between missed anomalies and non-anomalous controls suggested that missed anomalies were not processed differently from non-anomalous words.

Summary

In this chapter we described some of the psycholinguistic assumptions of traditional models of language comprehension, such as incrementality and compositionality, along with some of the work challenging the strict acceptance of these assumptions. Empirical evidence, employing a diverse range of experimental techniques, has demonstrated that language processing may, on many occasions, be far from complete or unambiguous. Instead, processing may often be shallow, or incomplete, and the resultant message underspecified. Semantic anomalies were taken as a key example.

Susceptibility to semantic illusions cannot be explained due to a lack of encoding nor to the possible 'correction' of personally detected anomalies. It has been shown that illusions are more likely to occur if there are shared semantic features between the anomalous and correct word. Also, the relationship between the anomalous word and the global context has been shown to be important, with words that have little contextual association being more detectable than others. Global fit theory argues that words with a strong fit to the context may only receive cursory analysis (enough to establish a relationship), and so an anomalous word with a strong global fit may not be detected as anomalous. Whereas a word with a low fit to the context will be more salient and so will be given more attention, processed more deeply, and hence is more likely to be detected if anomalous. This argument suggests that depth of processing may be variable. Factors that may modulate depth of processing include focus, or

attention-capturing devices, and the processing demands of the discourse. Both of these have been shown to influence anomaly or text change detection.

Questions addressed in this thesis

The majority of work on anomalies has employed off-line experimental techniques, such as paper and pen detection studies. Many of these have also relied on either a limited set of experimental items, or items that required purely encyclopaedic knowledge. Off-line studies provide little temporal information on anomaly detection, and they cannot provide information on what type of disruption, if indeed any, is caused by detection. More importantly, many studies do not distinguish between cases of successful anomaly detection and detection failure, nor do they compare these to suitable control versions of the experimental sentences. The importance of comparing these different conditions is to specify how processing may differ in these conditions. For example, is there evidence for more attentive reading when anomalies are detected compared to cases where they are missed? How do cases of undetected anomalies differ from non-anomalous control conditions?

The aim of this thesis is to attempt an answer to the following questions:

- What do on-line techniques, such as eye-tracking and event-related potentials (ERPs), reveal about the processing of semantic anomalies (e.g. time course of detection; the exact nature of these effects)?

- What processing differences, if any, exist between cases when readers detect and fail to detect semantic anomalies, evidenced via these techniques? This question is relevant to the issue of whether conscious judgements underestimate the impact of anomalies. For example, undetected anomalous words may be

registered by the language system in some way, but for some reason this does not reach conscious awareness. Also, analyses that incorporate both instances of detected and undetected anomalies, may underestimate the extent of disruption caused by detecting an anomaly, in comparison to missing an anomaly or controls. This leads to specific questions that can be asked in respect of detected and undetected anomalies, such as:

  - What differences can be observed between anomaly detection and appropriate control sentences?

  - What differences can be observed between undetected anomalies and appropriate control sentences?

- Do factors such as sentence processing load and task instruction modulate anomaly detection?

One good way of investigating these issues is by utilising an ERP paradigm. The advantage of adopting such an approach is that ERP studies have shown that easy-to-detect semantic anomalies, such as *socks* in, "He spread his warm bread with socks," evokes an exaggerated negative-going waveform that reaches a peak 400 msec post-stimulus. This peak is called the N400 component and is thought to reflect the ease of semantic integration (Kutas & Hillyard 1980). A more extensive review of this literature will be provided in Chapter 7. However, this methodology and established findings may be utilised to ask the following questions:

- In what ways do the detection of hard-to-detect semantic anomalies differ from easy-to-detect anomalies as evidenced with ERPs?

- Does the detection of hard-to-detect semantic anomalies produce an N400 effect similar to that observed with easier-to-detect anomalies?

- In cases where readers fail to detect semantic anomalies, is there evidence for an N400 effect (that is, evidence for unconscious detection)?

Chapter 2: Developing and pre-testing experimental items

The experimental literature on semantic anomalies has broadly relied on two types of materials, either easy-to-detect or hard-to-detect semantic anomalies. The "easy-to-detect" anomaly studies, which are most often cited in the ERP literature, have used materials with obvious thematic role or animacy violations. For example, "he spread the warm bread with *socks*" (Kutas & Hillard 1980), and "at breakfast the eggs would *eat* …" (Kuperberg, Kreher, Sitnikova, Caplan, & Holcomb 2007). Research based on such materials suggests that readers have no difficulty in detecting these types of violations, and they do so rapidly. However, harder-to-detect semantic anomalies have been employed by other researchers precisely because they are so often missed by readers. Typically, these studies have used off-line techniques to investigate the nature of the illusions, for example, providing answers to semantically anomalous questions (Barton & Sanford 1993), or proofreading for anomalous terms (Baker & Wagner 1987). Two examples of these materials are the *Moses* illusion and the *survivors* problem. These illusions are 'classic' in that they produce very strong and reliable effects with participants frequently failing to detect the anomaly. However, there are only a limited number of strong semantic illusions suitable for our research. The use of only one or two anomalous items clearly limits the experimental design and statistical analyses open to the experimenter, and would be too restrictive for the present purpose. Because my research aim was to use conventional eye-tracking and EEG techniques to investigate how hard-to-detect semantic anomalies are processed on-line, more semantic anomalies had to be produced. In producing the materials, there were a number of constraints that had to be met, concerning the criteria of definition, structure of materials, overall expected detection rates, and conformity to established properties.

Defining aspects

The characteristics of 'classic' hard-to-detect anomalies are that they rely on words (or names) which are highly related to the context of the story, but are used in an anomalous way. So, for example, in the survivors problem, the word *survivors,* which has a strong contextual fit to the scenario of a disaster, is used anomalously because it refers to people who are being buried. Examples such as these are quite rare in the literature, and often materials have been created by substituting so-called impostor terms, e.g. "What kind of tree did Lincoln chop down?" when it should have been George Washington (Reder & Kusbit 1991). These types of anomalies rely upon encyclopaedic knowledge and retrieval of specific facts rather than more general semantic properties. The *survivor's* problem (Barton & Sanford, 1993) is more subtle in that it relies on general semantic knowledge, and how well readers integrate this knowledge into their developing comprehension of a story.

Structure of materials

Existing materials have in the main been unsuitable for the carefully controlled comparisons required in eye-tracking and ERP studies. Constructing materials requires that a target word, or a critical region, can be clearly defined and easily compared across conditions. With some of the existing examples, there is a difficulty in defining which word is 'critical'. So, for example, in the *Moses* illusion, "How many of each animal did Moses take on the Ark", the question is which word would be taken as the critical word, *Moses* or *Ark*? Furthermore, would it be fair to exchange the anomalous name *Moses* to *Noah* for a suitable control version? The materials had to be constructed in such a way that the critical word was clearly defined, and that it always remained the same in both anomalous and non-anomalous versions.

Detect rates

The aim of the experiments reported in this thesis was to compare measurements when people detect anomalies with cases where they miss them, and hence an overall detection rate of 50% was optimal. In order to satisfy this criterion, it was necessary to pre-test a large pool of anomalies. Additionally, it was desirable that there were at least some detections and some misses of all stimuli. Finally, it would be crucial to check in the experiments proper that everyone was aware that each anomaly was indeed anomalous, even if they missed them during the main experiment.

Conformity to established properties

In addition to these ground-rules, it is clearly desirable that the final set of materials conforms to established properties of borderline anomalies. Specifically, detection rates should be modulated by manipulations of high versus low sentential processing load and linguistic focus. Experiment 1 reported later in this chapter will demonstrate that these materials conform in an expected way to load manipulations. Research conducted more recently, and carried out as part of two supervised undergraduate projects, will also be discussed because these projects demonstrate that the materials conform to expected patterns of modulation through focus manipulations. Since these two questions justify the use of these materials in the main experiments, they will be reported together in the present chapter.

Developing and pre-testing materials

The semantic anomalies used in subsequent chapters were adapted from an initial corpus of 46 items that were extensively pre-tested to ensure that readers would reliably fail to detect the anomalies some of the time. The *survivor's* problem, reported by

Barton & Sanford (1993), was used as a model for the initial set of experimental items. In their work, Barton & Sanford presented participants with a detailed description of an event (i.e. a plane or bicycle crash). This was followed by a question which participants were required to answer and that contained a semantic anomaly ("where should the *survivors* be buried?"). The anomalous phrase was counted as detected if participants made reference to the anomalous term in their answers, or if they identified the anomaly during a subsequent debriefing session. A similar approach was adopted in the development of experimental items reported here.

Materials: The experimental items were written as short stories that described a stereotypical scenario. At the end of the story there was a question and the semantic anomaly was always placed in this final question. The anomalies relied on stereotypical scenarios which might involve characters placed in an incorrect role, or events incorrectly described. For example, one story described the event of an aeroplane being forced to land by terrorists, and asked the participants whether the authorities should negotiate with the *hostages* (rather than hijackers). See table 2.1 for examples. Each participant read 8 experimental items which were presented along with 52 filler items in a fixed random order. The filler items were written in a similar style (and were of a similar length) to the experimental items. Each experimental item was presented to 15 participants.

Assessing the strength of the semantic anomalies. There were two versions of the pre-test procedure, one where materials were presented auditorially and the other a written version. Table 2.1 provides three examples with associated detection rates. Appendix 1 contains a full list of the 46 anomalies, as they existed at this stage, along with non-detection rates for auditory and text-based versions.

Table 2.1: Three examples of experimental items as they were originally presented in both auditory or text versions. The anomalous word has been placed in bold. Detection rates in both audio and visual versions are presented as a percentage.

1) Pan Am flight 004 from Chicago was forced at gunpoint to land at New York's John F. Kennedy Airport. The emergency services responded quickly and all were in attendance around the international terminal building. Time was running out for the airport police. They knew that people would be killed.
   Q. Under these circumstances, should the authorities meet the demands of the **hostages** or stand up to international terrorists?

   Non-anomalous word is hijackers
   Detection rates in Audio version= 20% Visual = 27%

2) The future of the NHS has been a major electoral issue. There is increasing concern from nursing unions, that their members are under-paid. UNISON has threatened strike action if a new government does not improve the present situation. However, critics argue that strike action could dangerously affect the people in their care.
   Q. After considering these arguments, would you support a national strike until there is a reasonable pay settlement for all **patients** in NHS hospitals?

   Non-anomalous word is nurses
   Detection rates in Audio version= 33% Visual = 53%

3) In a report published last year, it was claimed that the level of general knowledge of British students is extremely poor. This is present even amongst university students. In the recent Scottish Universities Mastermind quiz, one Paisley University student finished last due to a poor performance in the general knowledge round.
   Q. In your opinion is it fair to tar all students with the same brush, just because one student couldn't answer the question, 'Where's Amsterdam?, because her knowledge of **history** was so poor?

   Non-anomalous word is geography
   Detection rates in Audio version= 60% Visual = 67%

There were 3 parts to the pre-test procedure, and these were the same for both auditory and text versions; all items were tested in both versions. At the outset, participants were led to believe that the study concerned student attitudes towards contemporary events and that their task would be to express their opinions about these events. They were

also informed that an anomaly had recently been discovered in one story already pre-tested. They were shown this example, which was a story concerning a cloned *cow* called Dolly (which should have been sheep and was a well-reported story at the time). As a secondary task they were asked to monitor for further anomalies and to record these anomalies in their booklets, along with their answers to questions, should they notice any.

In ***part 1*** of the auditory version participants listened to a series of stories that had been recorded on to audio tape. The materials were recorded by a female speaker who had a clear voice, and presented through good quality speakers in a quiet room. At the end of each story there was a question, after which the experimenter paused the audiotape whilst participants wrote their answers down in an answer booklet. In the text based versions participants were given a booklet with all the stories and questions printed and answers were written straight into the booklet. This section gave us a measure of incidental anomaly detection.

In ***part 2*** of both versions participants were given a written copy of experimental items and asked to point out any anomalies in the text. They also indicated if they had noticed the anomaly in part 1. This section gave us a measure of detection when instructions were explicitly to detect anomalies.

In ***part 3*** the experimental items were re-presented with the anomaly presented in bold and with an accompanying explanation. Participants indicated if they did or did not understand the anomaly and were given space to write feedback. Finally, they were debriefed as to the purpose of the study.

Detection rates were based on responses in part 2 of the procedure, and were only counted if participants indicated that they understood the anomaly in part 3. Items that

were always detected were removed from the final list and the feedback provided in part 3 of the questionnaire was also used to either re-write or remove items that were unsuitable for further testing. The final list consisted of 46 items that were adapted and used in subsequent studies.

Overall detection rates averaged approximately 30% in the auditory version, and 42% in the text version. The rate for the written materials was considered suitable for the planned future empirical work because it would permit the acquisition of data under both detect and non-detect conditions. The distribution of detection rates is illustrated in figure 2.1 for both auditory and text versions. It can be seen that the text version lead to higher rates of detection overall, and that there is a consistency of items across the two versions (i.e. items with a high detection in the auditory version are also detected frequently in the text version).

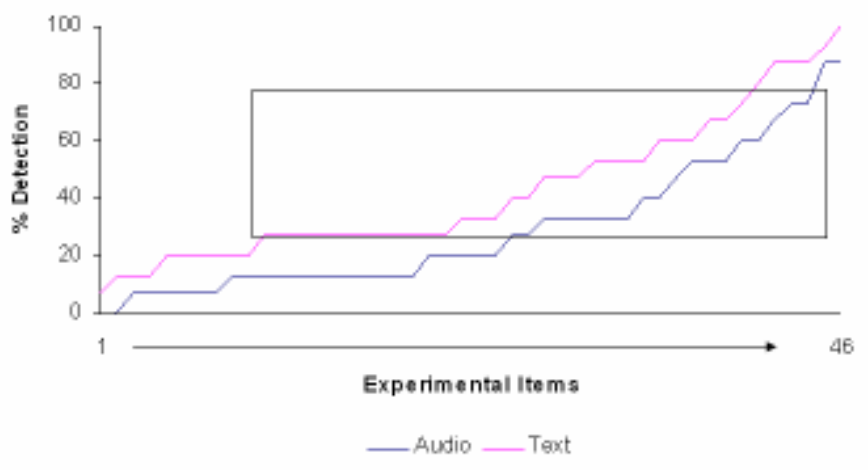


Figure 2.1: This figure is a simple schematic diagram to illustrate the distribution of scores across items. It can be seen that detection rates were variable across items, and with many items scoring within the middle range of detection rates (30% - 80%). This also illustrates that the detection rates for the auditory version was consistently higher compared to the text-based version. Those scoring in the central region (between approximately 30% and 80% were preferentially chosen for future experimental manipulations (indicates by the central black square).

With this pool of materials, care was taken to ensure that the target anomalous word was clearly defined in all items, and was anomalous as soon as the word was encountered (i.e. it was not reliant on subsequent information rendering the word anomalous).

Use of the materials in Experiments

Some small details of the materials were changed over subsequent experiments, for instance the overall length of the preceding contextual information and the positioning of the anomalous word itself, but the pre-test provided a sensible starting point for the final development of materials. In the eye-tracking experiments, only materials that lay in the range of 30-80% detection were used, as shown by the box in figure 2.1.

A crucial question, beyond evidence for detection failure, was whether or not these items would behave in a similar way to other semantic anomalies reported in the literature, that is, do they have similar properties to other *Moses*-type anomalies? To demonstrate this, in the following section, Experiment 1 reports that sentential load can modulate anomaly detection. Later, two further experiments are described that demonstrate the influence of focus on rates of anomaly detection.

Experiment 1: Increasing processing load reduces detection rates of semantic anomalies

Under what circumstances are semantic anomalies detected? Numerous studies have demonstrated that semantic anomalies are readily missed by readers (for example, Erickson & Mattson 1981; Barton & Sanford 1993), and that detection can occur in some situations and not others. Detecting a semantic anomaly in part relies on how deeply a word is processed, i.e. the extent of lexical retrieval and semantic integration of

a word. As such, anomaly detection may be influenced by the semantic similarity between correct and semantically anomalous words (Van Oostendorp & De Mul 1990), and by the relationship between the anomalous word and the global fit it has to the context (Barton & Sanford 1993). Other factors, however, may also modulate anomaly detection, and these could include the limitations imposed by computational resources utilised in reading, and also the overall complexity of a sentence containing a semantic anomaly.

Language comprehension is assumed to place demands on working memory, and working memory resources are assumed to be limited (Just & Carpenter 1992, Kintsch 1988). A limited capacity working memory system may affect language processing in one of two ways: when memory capacity is exceeded items may either be forgotten or lost, or overall processing speed may be slower (Carpenter, Miyake & Just 1995). Individual differences in working memory capacity can be measured using Daneman & Carpenter's readings span task (1980). This test measures the processing and storage capacity of working memory processes whilst reading and was used by Hannon & Daneman (2001) to demonstrate that working memory capacity was correlated with higher rates of anomaly detection. They concluded that a larger working memory capacity permitted fuller semantic analysis and semantic integration, resulting in increased anomaly detection.

Manipulating sentence complexity influences the efficiency of language processing. For example, Gibson (1998, 2000) has argued that sentence comprehension places two demands on limited computational resources. These demands are the integration of new words into the current sentence, and the storage of currently relevant sentence elements. Gibson's distance-based dependency locality theory (DLT) suggested that increasing the distance between dependent syntactic heads increases the complexity of the

sentence, and hence difficulty for the reader. Increasing the number and complexity of computations between syntactic heads is assumed to affect processing efficiency. So, for example, Warren & Gibson (2002) demonstrated that an intervening full noun phrase, compared to a first or second person pronoun, increased participants' complexity judgements and resulted in longer reading times.

Sanford, Sanford, Filik & Molle (2005) demonstrated that increased sentence complexity could modulate depth of semantic processing in a text change detection paradigm. In Experiment 4 they defined complexity in the same way as Warren & Gibson, and presented participants with passages containing a target sentence where a critical word was changed between two presentations. If the same sentence, prior to the critical word, contained a full noun phrase (as opposed to a first or second person pronoun), participants were less likely to detect the word changing. This effect was localised at the embedded verb.

Increasing sentence complexity, and hence increasing the demands placed on working memory, should impact on the likelihood of detection of semantic anomalies. In the experiment presented here complexity has been manipulated by inserting supplementary information before the target word in the high load condition, and by moving this supplementary information to after the target word in the low load condition. If there are increased processing costs as more items in a sentence need to be remembered (and more demands are made on working memory), then a semantic anomaly placed in a more complex sentence should be detected less often compared to one placed in an easier sentence.

Method

Design and Materials

This experiment used a within-subjects design. Each experimental item was designed so that it could be presented in either a high or low load condition which was either anomalous or non-anomalous.

16 experimental items were adapted from the original corpus of semantic illusions (items were mostly taken from those that had led to detection rates in the middle range as illustrated in figure 2.1). Eight semantic anomalies were viewed by any one participant, four of them in the high load and four in the low load condition. These were presented in a fixed random order along with 52 filler items which had been written in a similar style to the experimental items. All items were four sentences in length and purported to report a recent news event. At the end of the story there was a short question concerning an issue raised by the story. The semantic anomalies were always placed in the final question. Processing load was manipulated by including parenthetical information either before the target word (high load) or after the target word (low load). In the example below, there is a description of a plane being hijacked. Participants were asked to answer a question about the story; however careful reading of the question reveals that the question is anomalous because it implies that negotiations are with the *hostages* rather than *hijackers*. In the story extra information about the 'officials' is given, that they must "ensure the safety of their passengers". This extra parenthetical information is placed between "officials" and the anomalous word "hostages" in the high load condition, but appears after the anomalous word in the low load version. For example:

Pan Am flight 004 from Chicago was forced at gunpoint to land at New York's John F. Kennedy Airport. The emergency services responded quickly and all were in

attendance around the international terminal building. Time was running out for the airport police. They knew that people would be killed.

Low load
Question: Under these circumstances, what difficulties would the officials at John F. Kennedy Airport, who must negotiate the demands of the **hostages**, be facing when they must ensure passenger safety and possible further threats to airport security?

High load
Question: Under these circumstances, what difficulties would the officials at John F. Kennedy Airport, who must ensure the safety of their passengers, be facing when they must negotiate the demands of the **hostages** and possible further threats to airport security?

Participants

60 participants were recruited from the undergraduate population of the University of Glasgow's Psychology department. They were paid £4 for their participation.

Procedure

The experimental procedure was very similar to the pre-test procedure reported earlier. Participants were informed that the study concerned student attitudes to current events. They were informed that they would be shown a series of stories taken from daily newspapers and they would be asked a question about the story. They were also informed, however, that an anomaly had been discovered in a previous version of the study ("Dolly the cow"). They were shown the item in full and the anomaly explained. As a secondary task they were asked to circle any anomalies in the booklet that they noticed. Emphasis was placed on question-answering rather than anomaly detection.

Participants read 60 items and answered the corresponding questions. At the end of this, they came to part 2 of the questionnaire. The instructions asked them to re-read 16 of the previous news stories and informed them that there were anomalies in some of these stories. Eight of the re-presented experimental items contained a semantic

anomaly and the other eight were the non-anomalous control versions. Their task was to circle any semantic anomalies that they detected. They also indicated by ticking a box whether they had noticed the anomaly in part 1 or 2 of the questionnaire.

The third part of the questionnaire presented the eight semantic anomalies again. The anomalous word was highlighted in bold and an explanation was provided for why the word was anomalous. Participants indicated by ticking a box whether they had noticed the anomaly in part 1, part 2, or if they had only just noticed it in part 3 of the questionnaire. They could also respond if they did not understand the anomaly and space was provided for feedback. Participants were debriefed at the end.

Results

Detection was assessed on the basis of participants' responses to part 3 of the questionnaire. In this section, the anomalies were re-presented with an explanation of the anomaly. Participants responded whether they had detected the anomaly at all, and if so at which point (part 1, 2 or 3 of the questionnaire). They scored 1 point for each anomaly detected and raw scores were converted into percentages for easier comparison.

In part 1, when participants were asked to focus on answering the questions, only 10% of the anomalies were detected. In part 2 when experimental items were re-presented and participants asked to identify anomalous statements, an additional 25.5% of anomalies were detected. Overall detection rates equalled 35.5%.

The anomalies that were detected were then separated into high and low load conditions and overall percentages re-calculated. Detection rates expressed as percentages for each part of the questionnaire are summarised in table 2.2 below.

	Part 1 Mean % (SE)	Part 2 Mean % (SE)	Total Mean % (SE)
Low Load	10 (1.9)	30 (2.3)	40 (3.6)
High Load	10 (2.3)	21 (3.4)	31 (3.5)

Table 2.2: Mean (standard error) percentage of detections in each section for both high and low processing loads

High and low load conditions were compared with a series of paired t-tests for each part of the questionnaire. In part 1 there was no difference between high and low load $p<0.1$. However, there were significant differences between high and low loads in part 2, by subjects $t(59)= 2.5$ $p<0.02$, and by items $t(15)= 2.4$ $p<0.03$. There was also a significant effect in part 3, by subjects $t(59)= 2.0$ $p<0.05$, and by items $t(15)= 2.6$ $p<0.02$. These results support the hypothesis that increasing processing difficulty decreases rates of anomaly detection.

Discussion

It was hypothesised that extra parenthetical information provided before a semantically anomalous word would place a greater demand on a limited capacity working memory system, and this would result in decreased anomaly detection. This hypothesis was upheld and results demonstrated that anomaly detection is in part modulated by sentence complexity. These results support previous findings in the semantic anomaly literature that word processing is not uniform but is in fact variable. The nature of this variability is that in some situations words will be processed more deeply and their anomalous nature detected by readers, whereas in other situations reading may be shallower resulting in anomaly non-detection.

This experiment has shown that variability in word processing is affected by the difficulty of the sentence, in this case by adding parenthetical information. There is a cost in processing efficiency when working memory capacity nears its limit, as argued

by both Just & Carpenter (1992) and Gibson (1998). This cost may be due to either memory failure or unsuccessful integration. Stretching limited resources appears to influence readers who have adopted a shallower reading strategy. A shallow reading strategy would not involve exhaustive checks that a word is semantically appropriate within a context. Such an argument was proposed by Barton & Sanford (1993), who argued that a shallow reading strategy may involve checking that a word fitted a global context, but not one that checked the core meaning of every word. Therefore, what may have happened in the present study is that anomalous words passed a simple check of global fit, and this was sufficient for comprehension in the high load condition. A more thorough, deeper semantic analysis that could have resulted in anomaly detection was not instigated, presumably because the resources were diverted in maintaining storage. In the low load condition, fewer demands were placed on working memory, and so the resources were available to both perform a global check, as well as a more thorough semantic analysis. However, while more were still detected in the low load condition, many still went unreported. This demonstrates that even though these sentences were relatively easy to read, shallow processing may still occur, as reflected in non-detection of anomalies.

Experiment 1 demonstrates that these materials are adaptable for experimental purposes, in that sentential load can modulate anomaly detection. While the influence of sentential load on anomaly detection has not been demonstrated before, it has been shown to influence depth of processing in respect of text change detection (Sanford, Sanford, Filik, & Molle 2005). Experiment 1 provides converging evidence that sentential load can modulate depth of processing in a semantic anomaly detection task.

In the next section, two studies are reported using these materials to demonstrate that linguistic focus can modulate depth of processing, and hence anomaly detection. While

these studies are not part of the PhD itself, they are described in some detail because they support our general claim that these materials behave in a similar way to existing materials reported in the literature. This is important to do because, we believe, it strengthens any general conclusions that we may want to draw. These studies were carried out by fourth-year University of Glasgow Psychology undergraduate projects that were supervised in part by the author.

Demonstrating the effect of focus devices on rates of anomaly detection

Linguistic focus has been shown to modulate semantic anomaly detection, whereby anomalies that are placed within the focus of a sentence are detected more frequently, than the same anomalies placed in an unfocussed position. For example, Bredart & Modolo (1988) used *it*-cleft structures to manipulate linguistic focus with a set of *Moses*-type anomalies, and showed that this increased detection of anomalies. Using similar materials, Bredart & Docquier (1989) showed that emphasis through upper case and underlined letters has the same effect. To investigate the role of focus on anomaly detection Bohan, Sanford, Clark, & Glen (in prep.) adapted items from the original corpus of semantic anomalies (additional items were also included which had been developed in the same way), and have reported two off-line studies demonstrating that focus does modulate anomaly detection. Furthermore, their results support existing literature that has employed similar focus devices. Their research employed two devices, the manipulation of typographical features and *it*-cleft structures, to demonstrate the role that focus and emphasis has on anomaly detection.

Bohan et al.[3] used typographical devices to focus readers' attention using the newly developed semantic anomalies. In one experiment, the anomalous words were presented in bold lettering. The anomalies were achieved by manipulating a prior context word. This meant that both anomalous and non-anomalous versions of the stories could be presented, and the critical word itself always remained the same (this manipulation was developed for eye-tracking purposes and will be described in more detail in Chapter 3). An example is:

> There was a daring and violent bank raid in Glasgow this month in front of twenty terrified cashiers. The Clydesdale bank in Govan was **ransacked / defended** by a squad of armed **police**, who carried loaded shotguns.

Here it would be anomalous for the police to ransack the bank (but not for them to defend it).

There were 4 possible conditions of any experimental item;

- the *critical word* may be in bold (police),
- the *context word* may be in bold (ransacked),
- an *irrelevant word* in the story in bold (loaded),
- no word was placed in bold.

These variants were used for both anomalous and non-anomalous versions. At the end of each story there was a question based on the passage, and the participant's task was to answer the question. In a similar procedure to the pre-test, participants were informed that anomalies were in the text and that, as a secondary task, they should circle detected anomalies. Following this, participants were re-presented with the experimental items and requested to identify semantic anomalies. Finally, in part 3 the anomalies were

[3] The data for this experiment was collected by Kirsten Glen, a 4th year psychology student, whose work was supervised by Bohan and Prof. Sanford.

explained, participants indicated if they understood the anomaly and indicated at what point detection occurred.

Table: 2.3 below summarises anomaly detection rates in both parts 1 and 2 of the questionnaire (figures for part 2 include both those anomalies detected in part 1 plus the additional detections in part 2). As with the earlier pre-test, detection is lower in part 1 of the questionnaire where question answering was the main task, compared to part 2 when anomaly detection was made the main task.

Detection was enhanced when the anomalous word was in bold (36% and 69% for parts 1 and 2), or when the context word was in bold (36% and 68%), and there was no reliable difference between these conditions. Detection was lower when no words were placed in bold (27% for part 1 and 60% for part 2), and there was a virtually identical rate of detection in the other-bold condition (27% and 61%). Bohan et al. reported reliable effects of focus on detection rates for both focus conditions (anomalous, and context words in bold) when compared to either unfocussed manipulations (other word was in bold, or no word in bold).

	Part 1	Part 2
Anomalous word in bold	36	69
Context word in bold	36	68
Other word in bold	27	61
No bold	27	60

Table: 2.3: Mean anomaly detection (%) for typographical focussing devices in parts 1 and 2

These effects showed that anomalous information that had been presented in bold lettering led to a higher rate of anomaly detection, compared to either when an anomaly-irrelevant word is in bold (other word), or when no word in the text is in bold. The same pattern of effects was observed when a prior context manipulation (which

rendered the target word as anomalous or not) was also placed in bold. This context word/phrase was closely linked to the critical anomalous word itself, and placing this context word/phrase in bold appears to be just as effective for increasing anomaly detection as placing the anomalous word itself in bold. The use of typographical features, such as bold lettering in this case, has the effect of a focussing device because it communicates to the reader that 'this information is important'. It was argued that because both the target and context information were placed in focus in this way, participants were more likely to read this information more carefully, resulting in higher rates of anomaly detection. This occurred both under conditions of incidental anomaly detection (in part 1), and also when explicitly searching for anomalies (in part 2).

In Bohan et al's second experiment[4] they employed *it*-cleft constructions to place contextual information in a focussed position. Bredart & Modolo (1988) have shown that cleft constructions can modulate anomaly detection. They asked participants to evaluate statements for truthfulness, for example:

(1) Moses put two of each kind of animal on the ark.

(2) It was Moses who put two of each kind of animal on the ark.

In sentence (2) the focal information is on *who* built the ark, and therefore the name Moses is in a focal position. Participants were more likely to detect that the name is anomalous in this context, compared to sentence (1). Sturt, Sanford, Stewart & Dawydiak (2004) demonstrated similar results with cleft constructions, but employed a different methodological technique, text change detection. They presented short passages, such as (3) followed by a target sentence that placed one of two noun phrases in a cleft construction (*Jamie* in 3A, or *cider* in 3B).

[4] The data for this experiment was collected by Fiona Clark, a 4th year psychology student, whose work was supervised by Bohan and Prof. Sanford.

(3) Everyone had a good time at the pub. A group of friends had met up there for a stag night.

(3A) It was Jamie who really liked the cider, apparently. [Focus on Jamie]

(3B) What Jamie really liked was the cider, apparently. [Focus on cider]

When the critical noun phrase, *cider*, was placed in a focal position (3B) participants were more likely to detect when this word was changed to either *beer* or *drink*.

These two studies, therefore, suggest that cleft constructions can affect both anomaly and change detection. Bohan et al explored this issue further. They re-wrote the semantic anomalies used in their previous study, but focussed contextual information through the use of clefting, and contrasted detection rates to unfocussed, non-clefted, constructions. In the example below, the word *vacated* is anomalous in the context of an already empty table:

Introductory sentence	The restaurant manager regretfully told Mary that they were very busy and she would have to wait.
Focussed version	It was the empty table by the window that was likely to be *vacated* in about twenty minutes.
Unfocussed version	The empty table by the window was likely to be *vacated* in about twenty minutes.

In line with our prediction, when contextual information was placed in focus using the *it*-cleft construction, detection of subsequent semantic anomalies was higher at 77%, compared to 65% in the control version. This difference was statistically reliable.

The results from these two focus manipulations demonstrate that these materials behave in a way that is consistent with other *Moses*-type materials. In line with previous work (e.g. Bredart & Modolo 1988; Bredart & Docquier, 1989), when anomalous words were focussed either through typographical devices, or through *it*-cleft constructions,

semantically anomalous words were detected more frequently. Our argument is that higher rates of anomaly detection occur because words which are currently in focus are processed more deeply.

In summary. So far this chapter has described the development and pre-testing of a basic set of semantic anomalies. These materials form the basis for research subsequently reported in this thesis. The materials have been extensively pre-tested to ensure that they are detected some of the time, but not all of the time. Three experiments have been reported to illustrate that they are suitable for experimental manipulation, and that they have similar properties to other *Moses*-type illusions. This has been demonstrated in the replication of results found with linguistic focus. Also, a new result which supports evidence from a converging methodology (text change detection) demonstrated that semantic anomalies may be modulated by sentential load.

General framework for future research

One aim of this thesis is to investigate the on-line processing of semantic anomalies. Another is to collect data on participants' accuracy in anomaly detection. This has been achieved by requesting participants to directly report anomalies as they were detected. With the materials developed so far it should be possible to acquire eye-movement and EEG data partitioned into instances when anomalies are accurately detected and when they are not. Both of these conditions can then be compared to a baseline, non-anomalous condition. In each experimental chapter, the data will be partitioned in this way and analysed initially with an omnibus one-way ANOVA (with three levels; anomalous detect, anomalous non-detect, non-anomalous). The ANOVA comparison should illustrate any differences between these three conditions. However, it is in the specific planned comparisons that we hope to clarify whether there are any processing differences between the three conditions. These comparisons are:

1. Detected anomalous compared to non-anomalous control. This comparison should illustrate whether the detection of an anomaly disrupts normal reading. If detection does disrupt reading, then this comparison should illustrate the repair processes concomitant with anomaly detection. Related to this, the comparisons should also provide important information on the time course of anomaly detection, that is, does detection occur as soon as the anomalous word is encountered, or is it delayed?

2. Detected anomalous compared to undetected anomalous. This comparison will use eye-movement and EEG data taken from just the anomalous condition partitioned into cases of detected and non-detected anomalies. Comparing this data will show whether there are any processing differences between detected and missed anomalies. For example, it may be that there is no difference in processing between these two conditions, with non-detection equally as disruptive as detection. Alternatively, there may be significant disruption when anomalies are detected, but little disruption in the non-detected cases.

3. Undetected anomalous compared to non-anomalous control. A crucial question is whether or not there is any evidence for unconscious detection of semantic anomalies? If undetected anomalies are unconsciously processed then we may reasonably expect to see some disruption in the eye-movement data, even when anomalies go unreported. If, however, there is no processing cost, then we should expect there to be no difference between these conditions.

These comparisons will be made, and questions asked, in each of the subsequent experimental chapters.

Conclusions

- In summary, this chapter has reported the development of a basic set of experimental semantically anomalous items.

- An extensive pre-testing procedure has established that these items are suitable for future empirical investigation because they produced rates of detection of 42% in the text version, which is ideal for analyses between anomalous detect and non-detect. Also, these items have a clearly definable anomalous word and have a target word that can be controlled across conditions which will permit more reliable comparisons across experimental conditions.

- The load and focus studies show that the detection rates under baseline conditions is indeed near 50% giving us confidence in using these materials.

- It has been demonstrated that processing load can modulate rates of anomaly detection. When processing load was increased by inserting parenthetical information prior to the anomalous word, detection rates decreased in comparison to less complex sentences. Such a result has not been reported in the anomaly literature to date, however these results do mirror load manipulations reported with text change detection studies, which provides converging evidence for the influence of load on depth of processing.

- The materials have been shown to conform to what is known about *Moses*-type materials. Two focus manipulations, one with typographical emphasis and one using cleft constructions, have mirrored existing results reported in the literature. This provides additional weight to the argument that these materials are comparable to those already reported in the literature.

- Subsequent experiments will all compare anomalous detect, anomalous non-detect, and non-anomalous conditions. These comparisons will employ a general omnibus ANOVA for all three conditions, and then subsequent planned comparisons between conditions.

Chapter 3 Eye tracking semantic anomalies: A preliminary exploration

Previous studies have nearly always used off-line measurements to investigate borderline anomalies, and while they illustrated the effects of load and focus on detection rates, they have not thoroughly examined the time course of detection and the pattern of disruption that anomalies might cause. To investigate the time-course of anomaly detection, eye tracking is an ideal methodology. Any immediate disruption should be observed with early measures such as the duration of the first fixation, first pass reading times, and first pass regressions out. More sustained difficulties might be reflected in later measures such as total number of fixations on a target word and the number of regressions back to the word. Eye tracking will help to answer questions such as; do borderline anomalies disrupt the reading process, and if so, how? How does anomaly detection compare to either non-detection or an anomaly not being present? And, when readers fail to detect an anomaly, are there any differences compared to a control condition?

The initial study reported in this chapter was a preliminary exploration because there were two potential difficulties in the experimental design. One difficulty concerns the task instructions: if participants are to be asked to report anomalies, they will have to be informed that anomalies will be present. Our concern was whether our materials would be robust enough to go undetected on a sufficient number of occasions to afford a proper statistical comparison of data partitioned into detect and non-detect trials. The second problem is that in order to ensure that detection is accurate participants must describe the anomalies to the experimenter. Simply pressing a key does not guarantee accurate recognition of the anomalies. Since the DPI eye-tracker requires participants to use a bite bar to minimise head movements, participants must come off and then go

back on to the bite bar on a large proportion of trials, and calibration procedures must be repeated at regular intervals. Logistically, it was not known whether participants could be accurately tracked with this procedure, because of the regular disruption involved. Experiment 2 was the first attempt (to our knowledge) to combine eye-tracking with trial-by trial verbal reporting.

Experiment 2

Method

Design and Materials

This was a within subjects design, where participants read one version of a given experimental item, containing a target word that might be either anomalous or non-anomalous. The anomalous / non-anomalous status of critical words was determined by a prior context manipulation in the same sentence. An example material is:

> A jumbo jet was forced at gunpoint to land. Negotiation by
> the authorities with the **hostages** was brief. The siege lasted for two days.

In this example, the anomalous word is *hostages* because the authorities should be negotiating with the hijackers. In the control, non-anomalous, version, the word *negotiation* was changed to *communication*, which meant that the story now made sense because the target word, hostages, would not be anomalous.

52 passages were presented to participants. There were 26 experimental items, modified from the initial pre-tested pool where detection rates had been found to vary between 30% and 80%. Passages were three sentences in length, and varied between 2 and 5 lines of text on the screen. There were 26 fillers, half of which contained very

obvious anomalies. The materials were randomised and placed into two files, with experimental items appearing in the same order in each file. Each file contained only one version (anomalous or non-anomalous) of the twenty-six experimental items, with half of the items being anomalous and half non-anomalous. If an item appeared in one condition in file 1, then it appeared in the other condition in file 2. Thus each file contained 13 experimental items in each condition.

Procedure

A Generation 5.5 Fourward Technologies Dual Purkinje Image eye-tracker (with an angular resolution 10 minutes of arc) was used. Text was displayed on a computer monitor approximately 80 cm from the participant giving ~4 characters/degree of visual angle. Gaze location was monitored every millisecond. A bite-bar and head rest minimised head movements. The tracking procedure was explained to participants at the start, and they were instructed to read for normal comprehension. A calibration procedure was completed at the start, and calibration was checked at the start of each trial. A fixation spot ensured that when the text appeared participants were looking at the start of the text. Half of the participants saw file 1, and half saw file 2. Thus each participant saw each material only once, but over all subjects, each material was seen in each condition an equal number of times.

Due to the exploratory nature of this first eye tracking study, two slightly different versions of the procedure were used.

Version 1

In version 1, participants were explicitly instructed to search for anomalies in the text. They were initially shown a couple of example anomalies, e.g. “Mary ate some

ROCKS", and were warned that there might be similar ones in the text. They were instructed to respond in two ways: First, after every story there was an on-screen prompt which read, "Did this story make sense?" They responded, *Yes* or *No* by pressing the appropriate hand-held button (one in each hand). Secondly, if they had detected an anomaly they were also requested to describe the anomaly to the experimenter. To do so they indicated that they wanted to come off the bite bar by "knocking" on the table. This permitted the experimenter to turn off the tracking beam. The experimenter recorded all responses. Participants were not informed if they had been correct. The participant then resumed their position on the bite bar and the calibration procedure was repeated.

Version 2

In a second version of the same experiment, comprehension questions were included and replaced the previous, "Does this story make sense?" prompt. New instructions were also written that emphasised accuracy in answering these questions. A subtle subterfuge was also introduced with participants being informed that the study was in fact a pilot study, and that some stories accidentally contained anomalous words. In order to "help" the experimenter they were asked to point out any anomalies that they detected. They were asked to do this in the same way as in version 1 ("knocking" on the table). An obvious anomaly was placed at the start (not an experimental item), and participants were reminded of this secondary task if they did not voluntarily point it out.

Participants

22 first year psychology students of the University of Glasgow participated in the study, 10 in version 1 and 12 in version 2. There were 9 men and 13 women, and they ranged

in ages from 17 to 24, all had normal vision and were native English speakers. They were paid £6, or given course credit, for their participation.

Eye-tracking analysis

For the purposes of analysis, each experimental item was separated into six regions. So for example,

> A jumbo jet was forced at gunpoint to land. $_1$ /{*Negotiation /communication*}$_2$/ by the authorities with the$_3$/ **hostages$_4$** / was brief.$_5$/The siege lasted for two days. $_6$/

Region 1 was the **Introduction region** which included the entire first sentence and its purpose was to set the context of the story; region 2 was the **Context region** and it was by manipulation of this region that determined the anomalous or non-anomalous status of the later critical region; region 3 was the **Pre-critical region** which separated out the context and critical regions; region 4 was the **Critical region** which contained the target word; region 5 was the **Post-critical region** which contained the few words to the end of the sentence; region 6 was the **End region** which contained the whole final sentence of the story.

Fixations of less than 80ms were combined with adjacent fixations within one character position, and remaining fixations of less than 80ms were excluded from the analysis. Fixations of over 1200ms were also excluded. Four measures are reported that are normally taken to indicate early processing. *First fixation* duration records the length of the initial fixation within a region. *First pass reading time* records the time spent within a region before leaving to the right or left. *First pass regressions* measures the proportion of trials on which readers looked back to previous sections of the text

before progressing forward. *Regression path* time measures the time spent within a region before progressing to the next region, but also includes the time spent in regressions back to previous regions. Also, three measures normally associated with late processing were included in the analysis. *Number of fixations* reports the total number of fixations within a region. *Regressions-in* measures the proportion of regressions back to this region from subsequent regions. *Total time* measures the total amount of time spent within a region.

Results

Detection rates

The overall detection rate was 46% (131 out of a possible 286). This was an ideal rate of detection permitting detect and non-detect comparisons with the eye-tracking data. Of the 22 subjects two detected fewer than 20% of the 13 anomalous items they read, and only one detected more than 80% (the range of detection being 8 - 85%). The 26 items also varied on how frequently the anomalies were detected. Some items were detected less frequently, e.g. "The government rejecting the pay offer", and "symphonies being sung by divas" were only detected 5% of the time, whereas others, for example, "reading problems such as anorexia" (45%), and "the broken wing of a hot air balloon" (41%) were detected more frequently. There were five items that were detected less than 10% of the time.

There was a difference in the overall detection rates between the two procedures, with detection for version 1 at 52% and for version 2 at 40% (the main difference in the two procedures being the emphasis on searching for anomalies). The detection rates for individual items across the two procedures were compared using an independent *t*-test, and this difference was not reliable ($t(50)= 1.5$ $p>0.1$). The same data was also

analysed using a pearson's correlation, and there was a strong positive correlation (r=0.57) between detection rates for items. This suggests that items were detected at similar rates across procedures. Therefore, it was felt that the subsequent analyses of the eye tracking data could be carried out on the whole data set rather than treating them separately.

Eye-tracking analysis

Several comparisons were made with the eye-tracking data. The anomalous data had been separated into detect and non-detect based on the verbal reports provided by participants. One way ANOVAs were carried out first of all, comparing detect, non-detect and non-anomalous conditions. This was followed by a series of post-hoc t-test making the more important direct comparisons between detected anomalies and non-anomalous data; detected and missed anomalous data; and, missed anomalous and non-anomalous data. These pairwise comparisons are reported in section 3.2.4.

Omnibus analyses of detected, missed anomalous and non-anomalous data

The omnibus analyses compared the data from the anomalous condition separated into instances when anomalies were detected and missed, and the non-anomalous data. The anomaly detect and non-detect data were averaged by subjects and items for each measure, and compared to the non-anomalous control condition which had also been averaged across subjects and items. A series of one-way ANOVAs with three levels (detect, non-detect, and non-anomalous) were then carried out. There was considerable disruption observed with a range of measures in the critical and post-critical regions, as well as the context and pre-critical regions. The first and last sentences were treated as whole regions (introduction and end regions) and because these regions were so large

and only served the purpose of setting the context and finishing the story the data from these regions will not be reported.

In the **critical region**, there were significant effects observed with total time, number of fixations and regressions in. On average the *total time* that readers spent in this region depends on the presence and detection of an anomaly. This was significant by subjects $F1(2,42)= 13.1$ $p<0.001$, and by items $F1(2,50)= 17.1$ $p<0.001$ (see figure 3.1). The *number of fixations* was significant by subjects $F1(2,42)= 12.2$ $p<0.001$, and by items $F2(2,50)= 7.3$ $p<0.002$. Similarly, the *regressions in* to the critical region was significant by subjects $F1(2,42)= 11.9$ $p<0.001$, and by items $F2(2,50)= 7.3$ $p<0.002$.

In the **post-critical region** effects were observed with total time, number of fixations, first pass regressions, and regression path. The *total time* spent in the post-critical region was influenced by the presence and detection of anomalies, which was significant by subjects, $F1(2,42)= 13.2$ $p<0.001$, and items $F2(2,50)= 13.2$ $p<0.001$. Similar effects were also observed with *first pass regressions*, which was significant by subjects, $F1(2,42)= 7.9$ $p<0.001$, and items $F2(2,50)= 5.2$ $p<0.009$ (see figure 3.2). So too was r*egression path* which was significant by subjects, $F1(2,42)= 12.8$ $p<0.001$, and items $F2(2,50)= 13.3$ $p<0.001$.

Significant effects were also observed in earlier regions of the text, both the context and pre-critical regions, with number of fixations, regressions in, and total time (items only). In the **context region** there was an effect of anomaly with the *number of fixations* measure which was significant by subjects, $F1(2,42)= 4.0$ $p<0.03$, and items $F2(2,46)= 4.3$ $p<0.02$. The *regressions in* measure was significant by subjects, $F1(2,42)= 7.6$ $p<0.002$, and items $F2(2,50)= 3.7$ $p<0.03$. The *total time* measure was non-significant by subjects, $F1(2,42)= 1.5$ $p>0.1$, but was significant by items $F2(2,46)= 3.7$ $p<0.03$.

A similar pattern of effects was observed in the **pre-critical region**. The presence and detection of anomalies effected the *number of fixations* in this region, which was significant by subjects, $F1(2,42)= 16.9$ $p<0.003$, and items $F2(2,50)= 4.4$ $p<0.02$. The *regressions in* measure was also significant by subjects, $F1(2,42)= 8.2$ $p<0.001$, and items $F2(2,50)= 7.5$ $p<0.001$. The *total time* measure was also significant by subjects, $F1(2,42)= 4.4$ $p<0.02$, and items $F2(2,50)= 3.2$ $p<0.05$.

Overall, these analyses show that there is a clear impact on the eye movement depending on whether an anomalous word is present and detected, present and undetected, or absent. Significant effects with characteristically late measures, total time, number of fixations and regressions in, are observed in the context, pre-critical and critical regions. Effects in the post-critical region are observed with a mixture of early and late measures, first pass regressions, regression path, number of fixations and total time. While the ANOVA results demonstrate that there are strong effects within the data, further analyses are needed to determine which individual comparisons are significantly different. In the following sections the crucial comparisons between anomaly detect, non-detect, and non-anomalous conditions, are reported.

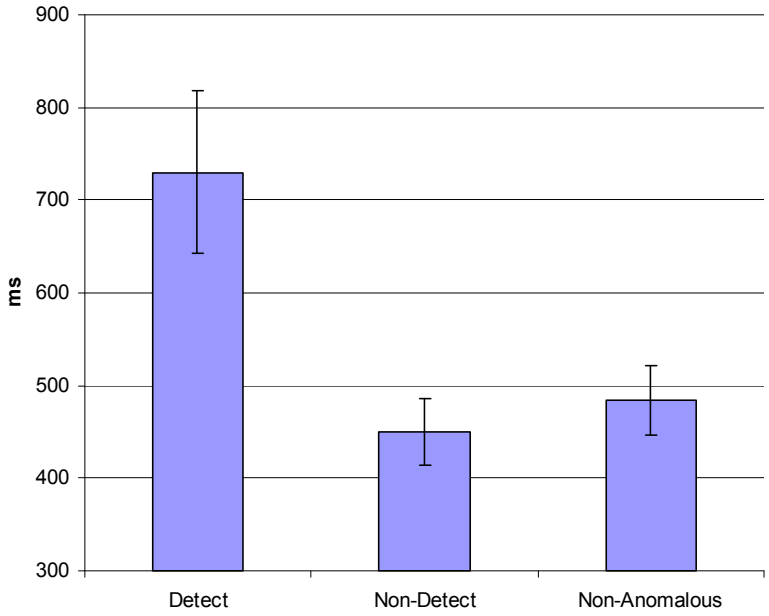


Figure: 3.1. Total time in the critical region in each condition

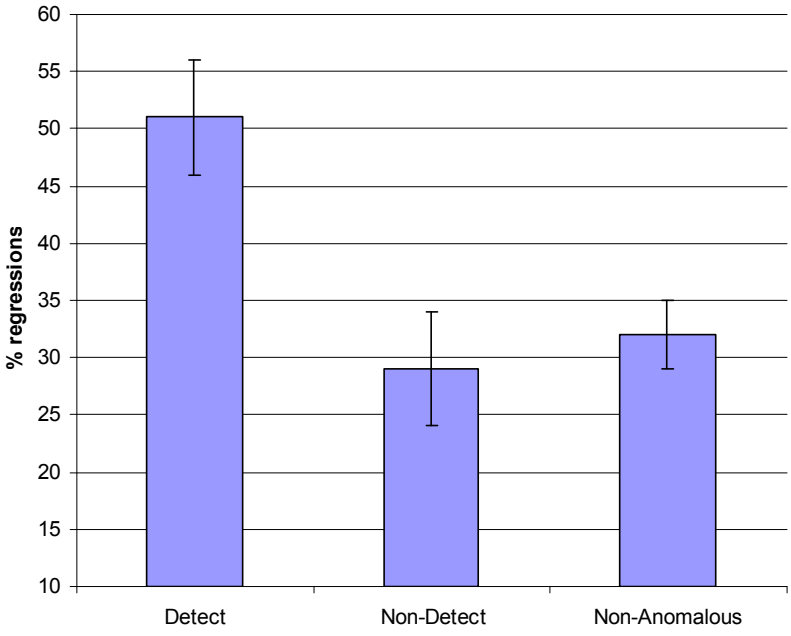


Figure 3.2: First pass regressions in the post-critical region in each condition

Detected anomalies vs. non-anomalous controls

Instances where participants correctly reported an anomaly as being present were separated from those where the anomaly was missed. The anomaly detect data was then averaged by subjects and items for each measure, and compared to the non-anomalous control condition, which had also been averaged across subjects and items in the same way. Descriptive statistics are summarised in table 3.1.

There were significant effects observed in the context, pre-critical, critical, and post-critical regions. These effects were with late measures, total time, number of fixations and regressions in, and these were again found in the context, pre-critical, critical region. A mixture of early and late measures (first pass regression, regression path, total time and number of fixations), were observed in the post-critical region.

Table 3.1: Summary of anomalous detect, anomalous non-detect, and non-anomalous conditions (mean, standard error) in the context, pre-critical, critical and post- critical regions. Significant differences in t-test comparisons between conditions are also indicated.

	First fixation (ms)	First Pass (ms)	First pass regression (%)	Regression path (ms)	Total time (ms)	Number of fixations	Regression-in (%)
Context region							
Detect	272 (12)	428 (29)	14 (3)	585 (45)	793 (68)	3.1 (0.3)	44 (5)
(Sig diffs)						**t1, t2**	**t1, t2**
Non-Detect	290 (17)	454 (32)	17 (3)	651 (57)	666 (64)	2.5 (0.2)	23 (4)
(Sig diffs.)							**t1**
Non-Anomalous	271 (14)	425 (23)	12 (2)	601 (49)	698 (62)	2.6 (0.2)	30 (4)
Detect-Non-Anom (sig diffs)						**t1, t2**	
Pre-Critical region							
Detect	263 (14)	659 (56)	25 (3)	1438 (161)	1653 (160)	7.7 (0.7)	60 (6)
(Sig diffs.)					**t1**	**t1, t2**	**t1, t2**
Non-Detect	260 (9)	680 (49)	22 (4)	1216 (82)	1262 (72)	5.5 (0.4)	38 (4)
(Sig diffs.)						**t1, t2**	
Non-Anomalous	255 (7)	706 (31)	30 (4)	1443 (81)	1396 (60)	6.5 (0.5)	41 (4)
Detect-Non-Anom (sig diffs.)					**t1, t2**	**t1, t2**	**t1, t2**
Critical region							
Detect	265 (18)	301 (25)	19 (6)	483 (46)	730 (88)	2.7 (0.3)	43 (5)
(Sig diffs.)					**t1, t2**	**t1, t2**	**t1, t2**
Non-Detect	260 (13)	332 (19)	16 (5)	466 (50)	450 (35)	1.8 (0.1)	20 (4)
Non-Anomalous	271 (14)	330 (19)	16 (3)	551 (75)	484 (37)	1.8 (0.1)	20 (2)
Detect-Non-Anom (Sig diffs.)					**t1, t2**	**t1, t2**	**t1, t2**
Post-critical region							
Detect	266 (10)	774 (67)	51 (5)	2390 (223)	1778 (126)	7.3 (0.5)	28 (5)
(Sig diffs.)			**t1, t2**	**t1, t2**	**t1, t2**	**t1, t2**	
Non-Detect	243 (9)	781 (46)	29 (5)	1503 (117)	1311 (74)	5.5 (0.3)	21 (3)
(Sig diffs.)	**t2**	**t2**					
Non-Anomalous	260 (8)	841 (33)	32 (3)	1597 (124)	1317 (53)	5.7 (0.3)	25 (2)
Detect-Non-Anom(Sig diffs.)			**t1, t2**	**t1, t2**	**t1, t2**	**t1, t2**	

Significant subjects and items t-test analyses are indicated where t1 and t2 are positioned between rows (Significant differences) Detect and Non-detect; Non-detect and Non-anomalous; the final row in each region illustrates significant t-test comparisons made between Detected anomalies and Non-anomalous conditions.

In the **critical region** significant effects were again found with total time, number of fixations and regressions-in. The *total time* on the anomalous phrase was longer compared to the control condition (730 vs. 484), which was significant by subjects $t1(21)= 3.6$ $p<0.002$, and by items $t2(25)= 4.4$ $p<0.001$ (see figure 3.1 for an illustration of all three conditions for ease of comparison with total time in the critical region). The *number of fixations* was also higher when an anomaly had been detected (2.7 vs. 1.8), which was also significant by subjects $t1(21)= 3.4$ $p<0.002$, and by items $t2(25)= 2.6$ $p<0.02$. Finally, there were more *regressions in* to this region when an anomaly was detected compared to the control non-anomalous condition (43% vs. 20%), which was also significant by subjects $t1(21)= 5.3$ $p<0.001$, and by items $t2(25)= 2.7$ $p<0.01$.

In the **post-critical region** there were significant effects with first pass regressions, regression path, total time and number of fixations. There were more regressions back within this region, as measured by *first pass regressions*, when an anomaly was detected compared to the non-anomalous condition (51% vs. 32%), which was significant by subjects $t1(21)= 2.9$ $p<0.008$, and items $t2(25)= 3$ $p<0.006$ (see figure 3.2 for an illustration for all three conditions with this measure in this region). Readers also appeared to be slowing down in this region when an anomaly was detected, as measured by *regression path* (2390 vs. 1597), which was also significant by subjects $t1(21)= 3.1$ $p<0.005$, and by items $t2(25)= 4.1$ $p<0.001$. The consequences were that there was more *total time* spent within this region when an anomaly was detected (1778 vs. 1317), which was also significant by subjects $t1(21)= 4$ $p<0.001$, and by items $t2(25)= 4.3$ $p<0.001$. Finally, there was also a greater *number of fixations* made within this region when an anomaly was detected (7.3 vs. 5.7), which was reliable by subjects $t1(21)= 2.9$ $p<0.008$, and by items $t2(25)= 3$ $p<0.005$.

Other regions are also affected by conscious detection of anomalies. There were a greater *number of fixations* made to the **context region** when anomalies were detected (3.1 vs. 2.6), which was significant by subjects $t1(21)= 2.2$ $p<0.04$, and by items $t2(24)= 2.6$ $p<0.01$. Also, there more *regressions in* to the **pre-critical region** when an anomaly was detected (60% vs. 41%), which was significant by subjects $t1(21)= 2.9$ $p<0.009$, and by items $t2(25)= 2.6$ $p<0.02$.

Overall, there were no significant effects from early measures until after the critical word, in the post-critical region. Anomaly detection then appears to trigger regressions back to previous regions, that is, the context, pre-critical and critical regions. The consequence being that there are more fixations made in, and more time spent in, these regions.

Detected vs. non-detected anomalies

The anomaly data was separated into instances when anomalies were correctly reported by participants and when they went unreported. All instances of detected anomalies were averaged by participants and region for each measure, and compared to instances when the anomalies went unreported, which were also averaged by participants and region for each measure. The same procedure was carried out on the item analysis. Descriptive statistics are summarised in table 3.1.

The first observation with this comparison is that the effects obtained in the omnibus analyses, and anomaly detect versus non-anomalous comparisons, are again observed in the context, pre-critical, critical and post-critical region. Significant effects with late measures, total time, number of fixations, regressions in, are observed in the context, pre-critical, and critical regions, and a mixture of early and late measures in the post-

critical regions, first pass regressions, regression path, total time and number of fixations.

In the **critical region** significant effects were obtained with total time, numbers of fixations and regressions in. The *total time* spent within the critical region was much longer when the anomaly was detected than when it was missed (on average 730ms vs. 450ms), and this difference was significant by both subjects $t1(21)= 3.9$ $p<0.001$, and item analyses $t2(25)= 4.6$ $p<0.001$ (see figure 3.1). In the critical region a greater *number of fixations* was recorded when the anomaly was detected compared to non-detection (2.7 vs. 1.8) which was also significant by both subjects $t1(21)= 3.8$ $p<0.001$, and items $t2(25)= 3$ $p<0.006$). A similar pattern was obtained with the *regressions in* measure, with a higher percentage of regressions back to the anomalous word when it was detected than when it went undetected (43% vs. 20% regressions). These differences are statistically reliable by both subjects $t1(21)= 3.5$ $p<0.002$, and items $t2(25)= 3.7$ $p<0.001$.

In the **post-critical region**, as with the global analysis, there were significant effects with first pass regressions, regression path, total time and number of fixations. *First pass regression* showed a consistent trend with a higher average percentage of regressions when the anomalies were detected compared to when they were missed (51% vs. 29% regressions; see figure 3.2). This was significant by subjects $t1(21)= 3.8$ $p<0.001$, and by items $t2(25)= 2.4$ $p<0.023$. Readers had larger *regressions path* times in this region when they detected the anomaly (2390ms vs. 1503ms), which was also significant by subjects and items respectively, $t1(21)= 4.7$ $p<0.001$, $t2(25)= 4.3$ $p<0.001$. The *total time* that readers spent in this region was also longer when an anomaly was detected (1778ms vs. 1311ms), also significant by subjects, $t1(21)= 3.6$ $p<0.002$,and items, $t2(25)= 3.8$ $p<0.001$. Finally, there was a greater *number of*

fixations in this region when an anomaly had been detected (7.3 vs. 5.5), and this was significant by subjects, $t1(21)= 3.8$ $p<0.001$, and items, $t2(25)= 3.2$ $p<0.004$.

Significant effects were also obtained in regions prior to the critical region. The **context region** contained the phrase that determined the anomalous status of the critical word. Two measures showed significant effects, number of fixations and regressions in. There was a greater *number of fixations* in the context region when anomalies were detected (3.1 vs. 2.5), which was also significant by subjects $t1(21)= 2.5$ $p<0.02$, and items $t2(23)= 2.3$ $p<0.03$. There was also a greater number of *regressions in* to this region when an anomaly was detected compared to when it was missed (44% vs. 23% regressions), which was significant by subjects, $t1(21)= 4.05$ $p<0.001$, and items $t2(25)= 2.4$ $p<0.025$.

The **pre-critical region** included all the words following the context manipulation and prior to the critical word. In this region significant effects were obtained with total time, number of fixations and regressions in. The *total time* spent in the pre-critical region was greater when the anomaly was detected compared to when readers missed it (1653ms vs. 1262ms). These differences were significant by subjects $t1(21)= 2.4$ $p<0.025$, and approached significance by items, $t2(25)= 2$ $p<0.057$. The *number of fixations* was on average significantly higher when the anomaly was detected than when it was missed (7.7 vs. 5.5), by both subjects, $t1(21)= 3.2$ $p<0.004$, and item analyses, $t2(25)= 2.5$ $p<0.02$. There was also a significantly greater number of *regressions in* to this region when anomalies were detected (60% vs. 38% regressions), by subjects, $t1(21)= 3.7$ $p<0.001$, and by items $t2(25)= 3.4$ $p<0.002$.

This comparison serves to illustrate further the differences between anomaly detection and non-detection. Disruption to reading the detected critical word is shown in late

measures such as, total time, number of fixations and regressions in. Disruption to the post-critical region is shown by early (first pass regressions and regression path) and late measures (total time and number of fixations). As with the detected anomalies versus non-anomalous comparison, there was evidence of disruption in earlier regions of the text (context and pre-critical regions). These earlier regions showed significant effects from total time, number of fixations and regressions in. This pattern of results suggests that anomaly detection triggers extensive re-reading of the critical sentence, evidenced by regressions back in to those regions, and with overall longer time spent in, and more fixations made in, those regions. The lack of significant effects from early measures in the critical region suggests that detection is not immediate, but slightly delayed, probably occurring in the post-critical region.

Non-detected anomalies vs. non-anomalous controls

The instances when participants failed to report a semantic anomaly as being present were separated from the anomaly detect data. For each participant and item the data was averaged per region for each measure. This was then compared to the non-anomalous control condition, which had been averaged in the same way. Descriptive statistics are summarised in table 3.1.

The most striking observation of this comparison is the lack of significant effects in the critical region. The only reliable effect is obtained with the number of fixations in the pre-critical region. Other effects are observed in the context region with regression in (subjects only), and in the post-critical region with first fixation and first pass (items only).

In the **pre-critical region** there is a smaller *number of fixations* when the anomalies were missed compared to the non-anomalous control condition (5.5 vs. 6.5), which was significant by subjects $t1(21)= 2.9$ $p<0.009$, and items $t2(25)= 2.04$ $p<0.05$.

In the **context region** there was fewer *regression in* to this region when anomalies were missed compared to the control condition (23% vs. 30%). This was significant by subjects only $t1(21)= 2.5$ $p<0.02$, but was not significant by items $t2(25)= 1.3$ $p>0.2$.

In the **post-critical region** significant results were obtained with early measures for the items analyses only. The time spent on the *first fixation* in this region was on average shorter when readers missed the anomalous term compared to the control (243ms vs. 260ms). This difference approached significance for the subject analysis $t1(21)= 1.8$ $p<0.08$, but was significant by items $t2(25)= 2.3$ $p<0.03$. A similar effect was observed with the *first pass* measure, with this region being read faster when readers missed anomalies compared to controls (781ms vs. 841ms), but this was significant for the item analysis only, $t1(21)= 1.4$ $p>0.5$ (non-sig.), $t2(25)= 2.7$ $p<0.01$.

The lack of significant effects in these analyses might suggest that missing a semantic anomaly is similar to it not being there at all. On the other hand, the few differences that have been detected in the context and post-critical regions might suggest that anomalies were missed because the stories were read more superficially. It is not possible based on the present data to decide which of these is the correct explanation.

Conclusions & Ways Forward

At the start of this experiment there were two potential difficulties identified with this study. One concern was that the materials would not be robust enough in an explicit anomaly detection paradigm. Our other concern was whether participants could be

effectively tracked with the necessary repeated interruptions to allow the verbal detection reports. Neither of these concerns were warranted. Ceiling effects were not observed in the detection rates. In fact, a near optimum level of detection at 46% was reached. As for the eye-tracking procedure itself, there were no difficulties with repeating the calibration procedure in-between trials. Also, participants did not find the task difficult to learn or too tiring.

Two questions that we raised at the outset concerned the detection of hard-to-detect anomalies; how did this impact on the eye movement data? And, what was the time course of anomaly detection? The initial omnibus ANOVAs clearly illustrated that anomaly detection resulted in significant disruption in all regions of the text. There were significant effects observed with late measures (e.g. number of fixations, total time, regressions in) in the context, pre-critical, critical, post-critical regions. Measurements were always higher in the anomaly detect condition which indicated that anomaly detection resulted in extensive re-sampling of the text. Significant effects with early measures were only observed in the post-critical region. This suggests that participants had already moved beyond the anomalous word itself before detection occurs. In other words, this suggests that the timing of detection is slightly delayed, and does not occur immediately on encountering a borderline-detect anomalous word.

The specific comparisons between conditions revealed subtle distinctions. The pattern of effects reported in the omnibus analyses were mirrored in the comparisons between anomalous detect and non-anomalous conditions, and also between detected anomalous and non-detected anomalous conditions, for both the critical and post-critical regions. However, in the earlier context and pre-critical regions, more measures were significant in the anomaly detect / non-detect comparison. This illustrates that a key difference between anomaly detection and failure to detect is the extensive re-reading of the text

when detection occurs. Furthermore, this strongly suggests that the processing associated with anomaly detection is very different from processing associated with undetected anomalies.

The final comparison addresses the important question of whether there was any evidence for anomaly detection without explicit awareness. That is, we compared instances when anomalies were missed with the non-anomalous control condition. It was hypothesised that if the critical word was detected as anomalous it should be reflected in the eye movement data. If this had happened we would have expected to observe a similar pattern of effects as reported in the cases of conscious detection. However, there was no evidence that detection occurred without conscious detection. There was no difference between reading the critical word in either an anomalous or non-anomalous condition. There were also no differences observed with early measures in the context or pre-critical regions to suggest that readers were initially reading the target sentence differently (e.g. perhaps scanning or reading less attentively). There was some evidence to suggest that participants were less likely to return to these regions when an anomaly was missed, as evidenced by the number of fixations and regressions in measures. Also, in the post-critical region, there was some evidence to suggest that this region was initially read more quickly when anomalies were missed (as shown by two early measures length of first fixation and first pass reading times). These effects may be interpreted as reflecting that when anomalies were missed participants were either scanning the text or just superficially reading it. However, we would argue against this interpretation because there does not appear to be any differences in reading initially up to, and including, the critical word. Also, in the post-critical region, while there is some evidence for initial shorter fixations, there are no effects observed with late measures. If this region was read more quickly when anomalies were missed then we would have expected to observe effects with late measures as well. Finally, these

results are quite patchy and they would need to be replicated to support any further interpretation. At present, therefore, we cannot conclude that missing anomalies was or was not due to less-attentive reading. But, we can conclude that there is no evidence for unconscious anomaly detection in cases where they have not been consciously reported.

Overall, these results are clear and appear to follow a sensible pattern. But there are some weaknesses with the study that needed to be addressed in future work. Firstly, two variations of the procedure were used and this needed to be standardised. Secondly, it was decided that all participants should be given comprehension questions with the emphasis placed on reading for comprehension, rather than detecting anomalies. Thirdly, in the present study it was simply assumed that participants had the necessary knowledge to detect all anomalies. Future studies will include a post-study multiple-choice test. Finally, the stories used in this experiment varied in the total number of words and how they appeared on the screen (i.e. how many lines of text they were displayed over). It was decided that this too should be standardised across materials. In the next experiment, an attempt was made to replicate these results in a design which includes all these improvements.

Chapter 4: Eye-tracking Semantic Anomalies 2

Experiment 2 demonstrated that semantic anomalies could be eye-tracked effectively while also asking participants to report the anomalies that they detected. The results showed a clear effect when anomalies were detected but no detectable disruption when anomalies were not reported. While these were certainly robust effects, there are several methodological weaknesses in the preliminary study that need to be addressed, including standardising the procedure, standardising the length of materials and regions, adding comprehension questions, and checking participant's knowledge of the anomalies. Experiment 3 represents an attempt to replicate the results of Experiment 2 after making these important changes.

Experiment 3

Method

Design and materials

The design essentially replicated that of Experiment 2. A given material was designed so that it could appear in an anomalous or non-anomalous condition. Through participants' responses, the data recorded for the anomalous condition could be classified as detect or non-detect. Within-participant analyses (one way, 3 level ANOVAs) and pairwise comparisons were carried out.

Materials

There were 22 experimental items, each of which was produced as a non-anomalous version and an anomalous version. These materials were identical to the previous eye-tracking detection study except that they were controlled for overall length of the

passages (most passages were standardised at 32 words long), and length within each of the six regions was also controlled. The regions are described in the example below. Seven experimental items had a slightly larger context region than the typical example given below (3 contained 2 words, 4 contained 3 words). This region was, however, always the same size in both anomalous and non-anomalous conditions, for example one item describes an accident involving a *hot air balloon* (3 words) with a damaged **wing**, which is anomalous. In the non-anomalous condition this was changed to, *busy charter plane* etc. An example item with a description of how each region was constructed, and the number of words within each region, is given below:

> A North American jumbo jet was forced at gunpoint to land in Canada. The authorities 1/ {*negotiated / communicated*} 2/ with the scared and desperate 3/ **hostages** 4/ and calmed them down.5/ The siege lasted for two days. 6/
>
> 1. *Introduction region* - this included the first sentence and part of the second - total length = 15 words
>
> 2. *Context region* - this region was manipulated between versions to determine whether the critical term was anomalous or not - total length varied, sixteen = 1 word; two passages = 2 words; four = 3 words.
>
> 3. *Pre-critical region* - this region linked the context and critical region - total length = 5 words.
>
> 4. *Critical region* - this was the critical word which was either anomalous or not dependent on the context - total length = 1 word.
>
> 5. *Post-critical region* - this region is the area directly after the critical word up to the end of the sentence = total length = 4 words.
>
> 6. *End region* - the end region is the final sentence - total length = 6 words.

30 fillers were added to the materials, half of which contained obvious anomalies. These were again 3 sentences in length and all contained 32 words.

For a given participant, half of the test materials were presented in the anomaly condition, and half in the non-anomaly condition. By producing two files, half of the

participants were able to see a given material in one condition, and half saw it in the other condition.

During presentation the text appeared over 4 lines, and anomalies were never presented at the start or end of any line.

To ensure that participants fully understood these anomalies they were asked to complete a multiple choice questionnaire during the debriefing stage (see appendices). For example, in relation to the hijacking scenario above, participants were asked, "When a plane has been hijacked, who would the authorities negotiate with?" They were then given the option of circling either, hostages, hijackers, or psychologists. Only data from correct responses were included in subsequent analyses.

Participants

28 psychology undergraduate students at the University of Glasgow acted as participants. Some completed the experiment for course credit, and some were paid £6 for their participation. None had taken part in any previous anomaly study.

Procedure

A Generation 5.5 Fourward Technologies Dual Purkinje Image eye-tracker (with an angular resolution 10 minutes of arc) was used. Text was displayed on a computer monitor approximately 80 cm from the participant giving ~4 characters/degree of visual angle. Gaze location was monitored every millisecond. A bite-bar and head rest minimised head movements. The tracking procedure was explained to participants at the start, and they were instructed to read for normal comprehension. A calibration procedure was completed at the start, and calibration was checked at the start of each

trial. A fixation spot ensured that when the text appeared participants were looking at the start of the text.

Participants were asked to read the short stories in a manner supporting normal comprehension. They were informed that each passage would be followed by a comprehension question and that responses should be either "yes" or "no". Their main aim, it was explained, was to answer these questions correctly. After this had been explained, they were also informed that there might occasionally be anomalies in the text. Anomalies were defined as whole words that were out of context for some reason. To illustrate this they were given some examples, including the Moses illusion. They were asked to inform the experimenter whenever they noticed any such anomalies. A short practice block of four items (2 with anomalies, 2 without) was presented before the calibration and experimental stage, and participants were given the opportunity to ask questions. Emphasis was placed on reading normally and answering the questions correctly.

Participants were situated in a comfortable position on the bite bar, and held a response button in each hand. After participants had read a story they pressed either response button to progress to the comprehension question. They then responded by pressing either the left button for "yes" or the right for "no". If an error had been detected they 'knocked' on the table, the experimenter turned the tracking beam off, and the participant came off the bite bar and explained what they thought the error was. Participants always answered the question first, and afterwards explained detected errors. The experimenter wrote down all errors that participants detected. The participant then went back on to the bite bar and the calibration procedure was repeated. After the tracking stage had been completed participants were debriefed about the nature of the study. They were then asked to complete a multiple-choice knowledge

check questionnaire, to ensure that they understood the anomalies. They were also encouraged to make any comments about any aspects of the anomalies on the sheet and to the experimenter verbally.

Regions of analysis

The experimental items were split into six regions for the purposes of data analysis:

> A North American jumbo jet was forced at gunpoint to land in Canada. The authorities 1/ {*negotiated / communicated*} 2/ with the scared and desperate 3/ **hostages** 4/ and calmed them down.5/ The siege lasted for two days. 6/

The introduction (1) and end (6) regions are not reported in the formal analysis. This is because they are large regions and served only to introduce or tie-up the stories. The four remaining regions are reported, and these are the context (region 2), Pre-critical (region 3), the critical region (region 4), and post-critical region (region 5).

Results

The tracking procedure was regularly interrupted to allow participants to report detected anomalies. This procedure caused very little difficulty. In most cases re-calibration was carried out easily, and in some it did not have to be carried out at all with the tracking beam easily establishing a lock.

Question answering

The comprehension questions were answered correctly 92% of the time.

The final multiple-choice quiz was answered correctly 100% of the time, and no items were removed for any participant due to lack of knowledge.

Detection rates

The overall detection rate was 49.7 %. This detection rate is comparable to the previous study, which was 46%, and is suitable for the comparison of data generated by detected and undetected anomalies

Eye-tracking analysis

Fixations of less than 80ms were combined with adjacent fixations within one character position, and remaining fixations of less than 80ms were excluded from the analysis. Fixations of over 1200ms were also excluded. Data that contained two or more regions with no data in first pass, which may indicate tracker loss, was removed from the analysis. This affected less than 1.4% of the data. Four measures are reported that are normally taken to indicate early processing. *First fixation*, *first pass*, *first pass regressions*, and *regression path*. Also, three measures normally associated with later processing were included in the analysis, *number of fixations*, *regressions-in*, and *total time*.

The verbal reports were used to classify the data from the anomalous condition into anomaly detect and non-detect. One-way ANOVAs were carried out to compare performance on missed anomalies, detected anomalies and the control non-anomalous condition. Although an omnibus analyses was carried out initially, specific planned comparisons were made of data from detected anomalies versus control, detected anomalies versus undetected, and undetected versus control, following the logic of Experiment 2. On the basis of Experiment 2, it was anticipated that we would see a clear pattern of disruption in the anomaly detect data compared to both the control non-anomalous and anomaly miss data. We also expected to see most effects in the critical and post-critical regions, with regression-based effects in the context and pre-critical

regions too. As before, it was expected that the majority of reliable effects would be observed with late measures, and with some early measures in the post-critical region. Further, no difference was expected in the anomaly missed and non-anomalous comparison.

Omnibus analyses of anomaly detect, non-detect and non-anomalous

One-way ANOVAs by participants and by materials were carried out comparing anomalies that were missed, anomalies that were detected, and the control non-anomalous condition. There was considerable disruption observed with a range of measures in the pre-critical, critical, and post-critical regions. The main results are reported below and the follow-up t-test comparisons are reported in the subsequent sections.

Critical region: The were significant effects for the presence and detection of semantic anomalies with the *total time* measure in this region, which was significant by subjects $F1(2,48)= 5.9$ $p<0.005$, and items $F2(2,34)= 14.9$ $p<0.001$. Significant effects were also observed with *number of fixations*, by subjects $F1(2,50)= 5.5$ $p<0.007$, and by items $F2(2,38)= 6.2$ $p<0.005$ (see figure 4.1).

Post-critical region: In this region significant effects were observed with *first pass regressions,* by subjects $F1(2,54)= 6.4$ $p<0.003$, and by items $F2(2,38)= 9.5$ $p<0.001$ (see figure 4.2). There was also a significant effect for *regression path*, by subjects $F1(2,54)= 5.4$ $p<0.007$, and by items $F2(2,38)= 4.5$ $p<0.02$. *Total time* was also significant, by subjects $F1(2,52)= 5.5$ $p<0.007$, and by items $F2(2,34)= 3.4$ $p<0.05$.. And, the *number of fixations* was also significant, by subjects $F1(2,54)= 4.5$ $p<0.02$, and by items $F2(2,38)= 4.2$ $p<0.02$.

Pre-critical region: A significant effect was observed for *regressions-in* (detect = 63%, non-detect = 40%, non-anomalous = 46%), by subjects $F1(2,54)= 12.5$ $p<0.001$, and by items $F2(2,38)= 14.2$ $p<0.001$.

In sum, there is a clear impact on the eye movement data depending on whether an anomalous word is present and detected, present and non-detected, or absent. Significant effects are observed with early measures only in the post-critical region (as in Experiment 2), and in late measures in all regions except the context region (unlike Experiment 2). While the ANOVA results once again demonstrated that there are strong effects within the data, further analyses are needed to determine which individual comparisons are significantly different. In order to do this a series of paired t-tests were carried out. These compared anomaly detect to non-anomalous control data, anomaly detect to anomaly non-detect data, and anomaly non-detect to non-anomalous control data. These comparisons are reported in the following sections.

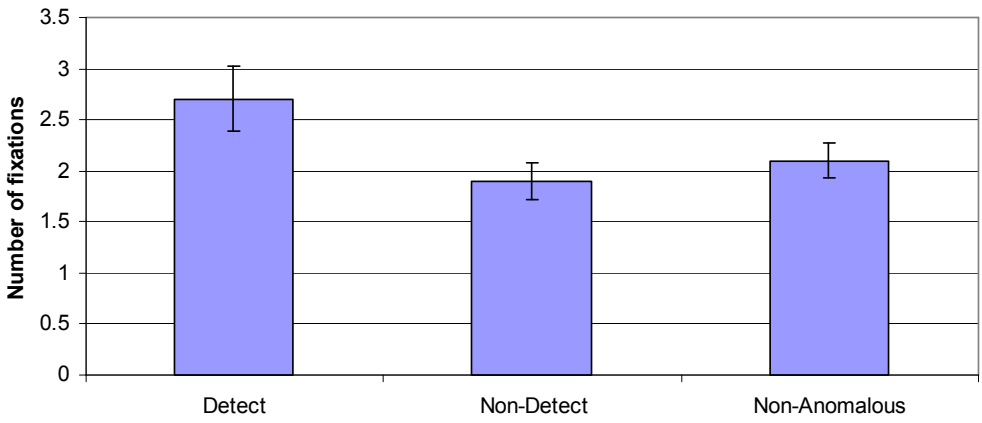


Figure 4.1: Number of fixations in the critical region separated into anomaly detect, non-detect and non-anomalous.

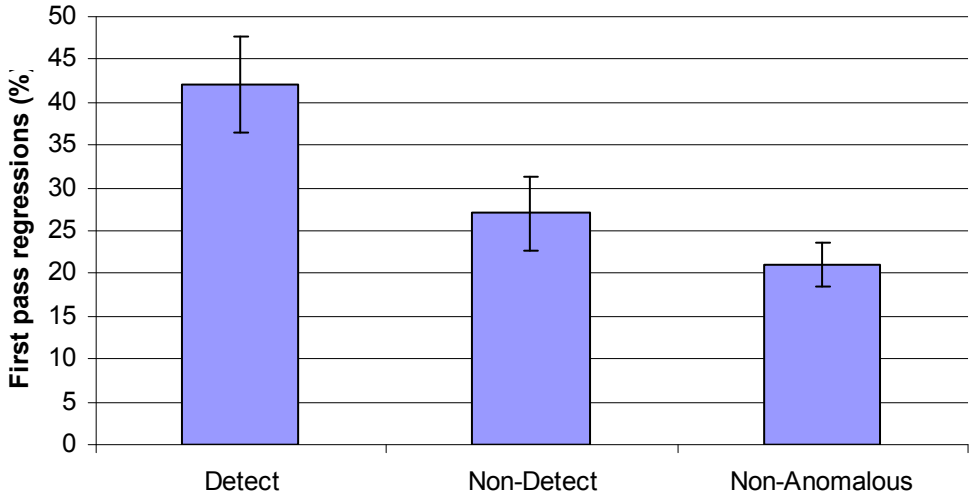


Figure 4.2: First pass regressions from the post-critical region separated into anomaly detect, non-detect and non-anomalous.

Detected anomalies vs. non-anomalous controls

The anomaly-detect data was averaged by subjects and by items and compared to the control, non-anomalous data using paired t-tests. Summary descriptive statistics are detailed in table 4.1. Late measures are significant in the context, pre-critical and critical regions, and earlier measures in the post-critical region. In the following analyses there was not always enough data for all items to be compared, and in those situations the degrees of freedom are reported as being slightly lower than would be expected if all items had been used. Effect sizes are also reported (Cohen's *d*).

Table 4.1: Summary data of anomalous detect, anomalous non-detect, and non-anomalous conditions (mean, standard error) in the context, pre-critical critical, and post- critical regions. Significant paired t-test comparisons between conditions are also indicated by t1 (subjects analysis), and t2 (items analysis).

	First fixation (ms)	First Pass (ms)	First pass regression (%)	Regression path (ms)	Total time (ms)	Number of fixations	Regression-in (%)
Context region							
Detect	287 (15)	449 (34)	20 (3.9)	572 (44)	756 (67)	3.2 (0.27)	38 (6.1)
(sig diffs)					**t2**	**t2**	
Non-Detect	276(14)	400(25)	23 (5.2)	588 (52)	660 (78)	2.7 (0.34)	34 (5.1)
(sig diffs)		**t1**					
Non-Anomalous Detect & Non-anom	276 (8)	449 (24)	21 (3.0)	588 (32)	639 (47)	2.8 (0.23)	28 (3.9)
(sig diffs)						**t1** , **t2**	
Pre-Critical region							
Detect	262	930	19	1327	1859	8.2	63
(sig diffs)	(12)	(54)	(3.3)	(74)	(112)	(0.65)	(4.9)
				t2			**t1** , **t2**
Non-Detect	237	937	27	1611	1821	7.4	40
(sig diffs)	(9)	(67)	(4.4)	(134)	(103)	(0.52)	(4.1)
	t1 , **t2**						
Non-Anomalous Detect & Non-anom	265 (7)	865 (38)	23 (2.7)	1397 (73)	1655 (94)	7.4 (0.57)	46 (3.2)
(sig diffs)						**t1** , **t2**	**t1** , **t2**
Critical region							
Detect	291 (13)	343 (24)	22 (3.5)	480 (31)	718 (95)	2.7 (0.32)	29 (4.4)
(sig diffs)		**t2**		**t2**	**t1** , **t2**	**t1** , **t2**	
Non-Detect	269 (14)	314 (22)	16 (3.1)	415 (32)	443 (42)	1.9 (0.18)	20 (3.3)
(sig diffs)							
Non-Anomalous Detect & Non-anom	278 (8)	328 (13)	17 (2.6)	452 (26)	558 (55)	2.1 (0.17)	20 (2.2)
(sig diffs)					**t1** , **t2**	**t1** , **t2**	**t1** , **t2**
Post-critical region							
Detect	261	757	42	1591	1463	5.7	27
(sig diffs)	(11)	(47)	(5.6)	(145)	(111)	(0.47)	(4.8)
			t1 , **t2**	**t1** , **t2**		**t1** , **t2**	
Non-Detect	255	725	27	1228	1202	4.9	19
(sig diffs)	(6)	(50)	(4.3)	(115)	(79)	(0.27)	(3.4)
						t1	
Non-Anomalous Detect & Non-anom	260 (7)	832 (36)	21 (2.6)	1263 (83)	1279 (70)	5.5 (0.40)	20 (2.8)
(sig diffs)			**t1** , **t2**	**t1**, **t2**	**t1**, **t2**		

Significant subjects and items t-test analysis are indicated where t1 and t2 are position between rows Detect and Non-detect; Non-detect and Non-anomalous; the final row in each region illustrates significant t-test comparisons made between Detected anomalies and Non-anomalous conditions.

Critical region: Significant differences were obtained with number of fixations, total time and regressions-in (borderline for subjects). When an anomaly was detected there was a greater *number of fixations* in the critical region compared to the control (2.7 vs 1.9) which was significant by subjects, $t1(27)=2.6$ $p<0.01$ ($d=0.50$), and by items $t2(20)=2.6$ $p<0.02$ ($d=0.56$) (Figure 4.1 illustrates this along with non-detected cases for easier comparison between conditions). A similar trend is seen with *total time* with significantly more time spent in the critical region when the anomalies were detected (718ms vs 558ms), by subjects $t1(27)=2.3$ $p<0.03$ ($d=0.45$) and by items $t2(18)=4.1$ $p<0.001$ ($d=0.93$). There was also evidence for readers returning to this region when an anomaly was detected shown by a greater percentage of *regressions-in* to this region (29% vs 20%), which approached significance by subjects, $t1(27)=2$ $p<0.06$ ($d=0.37$), but was significant by items $t2(20)=2.4$ $p<0.03$ ($d=0.53$).

Post-critical region: Robust effects were observed with first pass regressions, regression path, and total time in this region. The analysis from the *first pass regressions* measure shows that there were more regressions back within this region when an anomaly was detected compared to the control (42% vs 21%), which was significant by subjects $t1(27)=3.4$ $p<0.002$ ($d=0.64$), and by items $t2(20)=4.4$ $p<0.001$ ($d=0.95$) (see figure 4.2). The *regression path* shows a similar trend (1591ms vs 1263ms), significant by subjects $t1(27)= 2.9$ $p<0.008$ ($d=0.54$), and by items $t2(20)= 2.2$ $p<0.04$ ($d=0.47$) (both results are illustrated in figure 4.2). There was also evidence that readers spent significantly more *total time* in the region when they detected an anomaly (1463ms vs 1279ms), by subjects $t1(27)=2.6$ $p<0.02$ ($d=0.49$), and by items $t2(18)=2.3$ $p<0.04$ ($d=0.52$).

Context and Pre-critical regions: There was a greater percentage of *regressions-in* to the pre-critical region when an anomaly was detected (63% vs 46%), which was

significant by subjects $t1(27)=3.6$ P<0.001 (d=0.67), and by items $t2(20)=4.7$ p<0.001 (d=0.99). There was a greater *number of fixations* when an anomaly was detected in both the context region (3.2 vs 2.8) and in the pre-critical region (8.2 vs 7.4). These differences were significant in the context region by subjects $t1(27)=2$ p<0.05 (d=0.40), and by items $t2(20)=3.1$ p<0.006 (d=0.67), and also in the pre-critical region by subjects $t1(27)=2.1$ p<0.05 (d=0.40), and by items $t2(20)=2.4$ p<0.03 (d=0.53).

In summary, these analyses show a consistent effect on tracking measures in data when an anomaly was detected compared to the non-anomalous condition. This was seen with late measures in the critical region, and with early measures in the post-critical region. Detection resulted in a pattern of regressions back to earlier regions of the text, resulting in a greater number of fixations. These results suggest that detection is not immediate but is slightly delayed until the post-critical region.

Detected anomalies vs. non-detected anomalies

The data from the anomalous condition was separated into instances when participants reported the anomalies and when they failed to detect them. For each item and subject the data were averaged and compared using a paired t-test. There were significant effects in the critical region for total time and number of fixations, but in addition, the regression path measure suggests that readers were slowing down when they first encountered and detected anomalies. In the post-critical region, first pass regressions, regression path, and number of fixations are significant. There was also more regressions-in to the pre-critical region when anomalies are detected.

Critical region: There were significant effects for number of fixations, total time, and regression path (borderline for subjects) in this region. There was a greater *number of fixations* when an anomaly was detected compared to non-detected (2.7 vs 1.9, see

figure 4.1), which was significant by subjects $t1(25)= 2.6$ $p<0.02$ (d=0.51), and by items $t(19)= 2.6$ $p<0.02$ (d=0.58). Detection of anomalies led readers to spend more *total time* in this region, (718ms vs 443ms), which was significant by subjects $t1(24)= 2.8$ $p<0.01$ (d=0.56), and by items $t2(17)= 4.6$ $p<0.001$ (d=0.99) (see figure: 4.1). The *regression path* measure showed that detected anomalies were read more slowly compared to non-detected anomalies (480ms vs 415ms), which approached significance by subjects, $t1(25)= 1.9$ $p<0.07$ (d=0.36), but was significant by items $t2(19)=3.5$ $p<0.003$ (d=0.77).

Post-critical region: First pass regressions, regression path, and number of fixations showed significant effects. There were a greater number of *First pass regressions* within this region when an anomaly was detected compared to missed (42% vs 27%; see figure: 4.2), which was significant by subjects $t1(27)=2.1$ $p<0.04$ (d=0.40), and by items $t2(19)=2.8$ $p<0.01$ (d=0.63). There was also evidence that readers were slowing down in this region when they detected an anomaly, as shown by *regression path* (1591ms vs 1228ms), which was significant by subjects $t1(27)= 2.4$ $p<0.03$ (d=0.45), and by items $t2(19)= 2.2$ $p<0.04$ (d=0.49). There was also a greater *number of fixations* within this region when an anomaly was detected (5.7 vs 4.9), which was significant by subjects $t1(27)= 2.8$ $p<0.01$ (d=0.53), and by items $t2(19)= 2.2$ $p<0.04$ (d=0.50).

Pre-critical region: There were more *regression-in* to the pre-critical region when an anomaly was detected (63% vs 40%), which was significant by subjects $t1(27)$ 4.3 $p<0.001$ (d=0.82), and by items $t2(19)= 4.2$ $p<0.001$ (d=0.94). The *number of fixations* followed a similar trend with a greater number of fixations made in the pre-critical region when an anomaly was detected (8.2 vs 7.4). This was not significant by subjects $t1(27)= 1.5$ $p<0.2$, but was significant by items $t2(19)= 2.7$ $p<0.02$ (d=0.60).

The pattern of effects observed with these analyses closely reflects the trends that were observed in the anomaly detect / non-anomalous comparisons. Significant effects with late measures were obtained in both comparisons with total time and number of fixations in the critical region. However, this comparison provided some evidence to suggest that detection resulted in immediate disruption, as evidenced by the regression path measure. As with previous comparisons, there was evidence for early disruption in the post-critical region, evidenced by the first pass regressions and regression path. Detection triggered re-reading of earlier sections of the text, with more regressions back in to the pre-critical region.

Non-detected anomalies vs. non-anomalous controls

This comparison took the data from the anomalous condition when participants had failed to report anomalies, and compared this to the non-anomalous control condition (as before, the non-detect data had been averaged per subject and item). In this comparison there appeared to be few differences in the data. The only consistent significant effects were observed in the **pre-critical region** with the *first fixation* measure (see figure 4.3 which includes all three conditions for ease of reference). This measure suggested that readers who failed to report anomalies (i.e. non-detect) made shorter initial fixations in this region compared to the non-anomalous control (237ms vs 265ms), which was significant by subjects $t1(26)= -2.5$ $p<0.02$ (d=0.49), and by items $t2(20)= -2$ $p<0.05$ (d=0.44). Also, in the **context region** the *first pass* measure suggested that less time was spent reading this region when anomalies were missed, (400ms vs 449ms), which was significant by subjects $t1(27)= 2$ $p<0.05$ (d=0.39), but not significant by items $t2(20)= 1.2$ $p<0.2$.

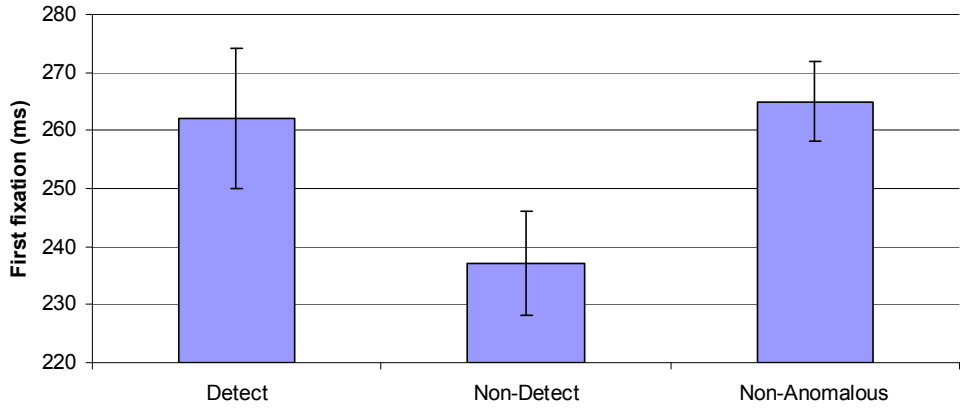


Figure 4.3: First fixation in the pre-critical region separated into detect, non-detect and non-anomalous

In summary, as with Experiment 2, there were few reliable effects with this comparison. There was a noticeable lack of effects in the critical region. Borderline differences were observed in the context and pre-critical regions. These few findings follow the same trend in terms of non-detection being related to shorter or fewer fixations and faster reading, compared to the control non-anomalous condition[5]. Apart from this there is little consistency in the measures providing these results, in the corresponding subject and item analyses, and in the level of significance observed. It is difficult to confidently interpret these results, but it is possible that readers who miss anomalies are doing so because they are generally engaged in shallower processing of the text.

In summary

Overall, this experiment has successfully replicated the pattern of results observed in Experiment 2. The omnibus analyses and follow-up t-test comparisons again showed that when an anomaly is detected there is disruption in the eye movement data. This is

[5] One interpretation of this is that participants were skipping this region, and so failure to detect was the result of not reading anomaly-relevant information. This was investigated by analysing the rates of skipping in all regions and conditions. No significant differences were found in the rates of skipping between missed anomalies and non-anomalous conditions which discounts this explanation.

characteristically demonstrated with late measures in the critical region, and from early measures in the post-critical region. These results suggest that detection is not immediate. When an anomaly is detected there is evidence for re-reading the target sentence. As before there are few differences observed with the non-detect and non-anomalous control, especially within the critical region.

This comparison between the instances when readers failed to report the presence of a semantic anomaly in text and the control non-anomalous condition has failed to show any consistent difference. This suggests that when an anomaly is missed there is no disruption to processing, and no effects to report. This is, of course, a problematic conclusion because it means accepting the null hypothesis. One way of gathering additional support for the null hypothesis is by looking at the level of power observed in the anomaly-detect versus non-anomalous comparison, and using this to determine the likelihood of obtaining an effect in the anomaly-missed versus non-anomaly comparison. This will be discussed in the following section.

The observed power of the anomalous detect vs. non-anomalous comparison to the anomalous non-detect vs. non-anomalous comparison

There is little difference between the trials when anomalies were missed and those of the non-anomalous control, and certainly no evidence in the missed cases for the major effects found in the detect cases. This is true numerically and statistically. However, the claim of no effect relies on accepting the null hypothesis. This in turn raises the question of the power of the experiment to detect effects, should they actually be there. The observed power of the detected anomaly data versus the non-anomalous controls was calculated for the measures showing effects in the anomaly detected versus non-anomalous comparison, and the results are presented in table 4.3. Clearly, some of the

Table 4.3: Selected Power and standard error (*SE*) data from Experiment 3. *SE* for number of fixations is in absolute numbers, for total time and regression path in ms, and for first-pass regressions in percentages. The table shows the five measures cited as evidence for an effect of anomalies detected on tracking performance, indicating region. For each statistic (*t*1 and *t*2), *SE* and observed power for anomaly detect versus control case is given. In the final column, the *SE* is given for the anomaly missed versus control comparison. The crucial observation is that *SE*'s in this latter case are lower, and so one might expect to find an effect of the size observed for detected anomalies at the alpha level given in the text in cases where the anomaly is missed, with a confidence indicated by the observed power indicated.

Region	Measure	*SE* Det vs Control	Power Det vs Control	*SE* Miss vs control
Critical	Total time			
	*t*1	67	0.72	63
	*t*2	79	0.99	44
	Number of fixations			
	*t*1	0.24	0.80	0.20
	*t*2	0.68	0.79	0.16
Post-Critical	First pass regressions			
	*t*1	6.1	0.95	4.2
	*t*2	6.5	0.99	5.1
	Regression path			
	*t*1	121	0.86	94
	*t*2	411	0.67	100
	Total time			
	*t*1	81	0.82	62
	*t*2	167	0.70	108
Context	Number of fixations			
	*t*1	0.21	0.64	0.32
	*t*2	0.31	0.90	0.17
Pre-critical	Regression-in			
	*t*1	4.7	0.96	4.0
	*t*2	5.8	0.99	5.8
	Number of fixations			
	*t*1	0.40	0.64	0.52
	*t*2	1.4	0.75	0.40

effects are more robust than others, but for the most robust effects (regressions out of the post-critical region and into the pre-critical) the power was greater than .95. Table 4.3 also shows the standard errors of the anomaly detected versus non-anomaly comparisons, and the corresponding standard errors for the anomaly missed versus non-

anomaly comparisons. The standard errors in the latter comparisons are smaller in every case. It seems reasonable to claim that the power to detect effects in the anomaly missed versus non-anomaly condition is at least as good as the observed powers observed in the anomaly detect versus non-anomaly comparison. For instance, the power to detect regressions out from the post-critical to the pre-critical region would be .95. It is noteworthy that the probability of detecting *at least one* effect on the basis of this power data is extremely high. The fact that not a single effect found in the detect cases was present in the non-detect cases is thus very unlikely to be a result of low power. These power analyses and comparisons support the interpretation that when readers failed to detect an anomaly the lack of effect was not the insensitivity of our measures, or a lack of power to detect any effects, but is most probably due to there being no effect actually present.

Conclusions & Ways Forward

By and large Experiment 3 has successfully replicated the pattern of effects observed in the preliminary eye tracking study (Experiment 2). The materials and procedure in Experiment 3 were more tightly controlled. These controls included standardising the number of words per region across items, and task instructions that emphasised reading for comprehension rather than anomaly-spotting. These changes have had little impact on the detection rates. The anomalies were detected on average 49.7% of the time, virtually identical to the detection rates observed in experiment 2 (46%). This rate of detection was also ideal for the purposes of comparing anomaly detect and non-detect data.

The omnibus analyses illustrated the time course of anomaly detection. The effects that were reported included significant effects in all regions with late measures, but early measures were only observed in the post-critical region. This suggested that anomaly

detection was not immediate, but delayed. If detection had occurred when the anomalous word was initially encountered then it would have been reflected in early measures in the critical region. Instead, the first sign of disruption is in the post-critical region. In support of this interpretation there have been similar findings reported with easily detectable pragmatic anomalies (Braze, Shankweiler, Ni & Palumbo, 2002), and anomalous noun phrases (Daneman, Lennertz, & Hannon, 2007). This is strong evidence that processes involved in semantic analysis are not completed exhaustively before progressing on to the next word. Why these effects are observed in the post-critical region rather than the critical region will be discussed further in Chapter 7.

A different possibility emerged, however, once the data was separated into cases of detect and non-detect. While the general pattern of effects in these comparisons mirrored the omnibus analyses, there was limited evidence that anomalous words were detected immediately. This was observed in the regression path measure in the critical region, where it appeared that readers were slowing down when they detected an anomaly. This is just one result with one measure, and as such obviously needs replication, but if this is reliable it would provide some evidence in favour of immediate and exhaustive semantic analysis.

A final question was concerned with anomalies that were missed, and whether or not there was evidence for unconscious detection in these cases. What is apparent from all the analyses reported here is that disruption was only ever observed when anomalies were consciously detected. There was no evidence that semantic anomalies were registered by the system without conscious detection. Therefore, there appear to be few discernable differences between cases where anomalies were present but missed, and non-anomalous comparisons. This is strong evidence for shallow semantic processing. The nature of shallow processing in this sense, may be that there has been either a

failure to fully retrieve the meanings of the critical words, or alternatively, that the meanings have not been successfully integrated into the discourse representation (although if meanings have been retrieved we may reasonably expect to see some impact on eye movements). Global fit theory predicts that word meanings may not always be retrieved when words have a strong fit to the global context (Sanford & Garrod, 1998). This is exactly what we have with hard-to-detect semantic anomalies. So, for example, participants may not notice that "victims are not sentenced", precisely because *victims* has such a good fit to a courtroom scenario. Anomalies may have been missed, therefore, because they successfully passed a scenario-relevant check and the individual meanings of words have not been retrieved.

There was some limited evidence to suggest that anomalies were missed because participants were reading less attentively. For example, initial first fixations in the pre-critical region appeared to be shorter when anomalies were missed, compared to the control condition (Experiment 2 also reported a similar effect in the same region with the number of fixations measure). However, the effect size was small in this case which suggests that this is not a particularly reliable result. Furthermore, there is no similar effect observed when detected and non-detected anomalies are compared which would also be expected if this difference was reliable. Furthermore, comparing non-detected and control data from Experiments 2 and 3 there is little consistency with eye movement measures across the two experiments to be confident that these are reliable effects. Also, after an inspection of the skip rates in all regions, there was no evidence that detected and missed anomalous and non-anomalous cases were being read in any way different. Finally, both Van Oostendorp & de Mul, (1990), and Reder & Kusbit (1991), inspected reading times for detected and non-detected anomalies and failed to find any differences in the two conditions. It seems reasonable to conclude, therefore, that without stronger and more reliable evidence, anomalies were not missed because

participants were reading faster and that these effects may in fact just be spurious. This may have been due to the process of partitioning data into instances of detection and non-detection. Partitioning data in this way may have resulted in unequal cell sizes as different subjects contributed different proportions to the overall total of each in some cases.

In summary, there are reliable effects observed when anomalies are consciously detected. The time course of anomaly detection appears to be that detection is slightly delayed rather than immediate. However, in contrast to this, there was some evidence from this experiment to suggest that detection did occur immediately. This requires further investigation. Also, there was no evidence that missed anomalies were unconsciously detected. In fact, there appear to be few differences between missed anomalies and non-anomalous controls. A potential way of investigating this issue further is by manipulating the task difficulty. If the detection task can be manipulated to be easier or harder, we may be able to observe processing differences that result in detection or failure to detect. Experiment 4 attempted to do this in an eye-tracking paradigm by manipulating the sentential load of sentences in an anomaly detection task.

Chapter 5: Manipulating processing load with anomaly detection in an eye tracking study

Experiments 2 and 3 clearly showed that when anomalies were detected there was significant disruption to the eye movement record. However, similar effects were not observed in cases when anomalies went undetected by participants. Because these anomalous words have gone unreported it seems reasonable to assume that they have been shallowly processed. That is, either the meanings of these words have not been fully retrieved, or they have not been integrated into the discourse. If either of these processes had occurred then they would have been reflected in the eye movement record, even in situations when participants were not consciously aware of anomalies. The question of what is actually happening when anomalies are not detected will be explored in future studies. However, in Experiment 5, the focus is on borderline cases of detected and undetected anomalies. If detection rates can be manipulated, so that anomalies may be more or less likely to be detected, then it may be possible to observe different styles of processing in the eye movement data associated with detection and non-detection. In Experiment 5 we manipulate overall detection rates by presenting anomalies in high and low processing load sentences.

Sentential processing load has been shown to influence participant's detection of text changes and anomaly detection. As discussed in Chapter 1 Sanford et al. (2006) reported that increased syntactic and referential load decreased detection of word changes in a text change paradigm. Also, Experiment 1 showed that detection rates could be influenced by the overall complexity of critical sentences, such that increased sentence complexity decreased rates of anomaly detection. Furthermore, Glenberg, Wilkinson & Epstein (1982) reported that the detection of contradictory information was greater in shorter texts (one paragraph) compared to longer texts (three paragraphs),

even when participants were told to expect contradictions and to specifically detect them. These studies suggest that increased processing load will affect rates of anomaly detection.

Processing load was manipulated in the present study by inserting an extra phrase prior to the critical word (see Warren & Gibson 2002; Sanford, Sanford, Filik & Molle 2005). It was hypothesised that more complex sentences would result in decreased rates of detection. This should increase the number of cases where anomalies go unreported, which should permit us to explore the nature of the boundary conditions between detected and undetected anomalies.

Experiment 4

Method

Design and materials

A given material was designed so that it could appear in one of four conditions: high memory load anomalous, high memory load non-anomalous, low memory load anomalous and low memory load non-anomalous. Through participants' responses, the data recorded for anomalous materials could be classified as detect or non-detect.

Materials

There were 26 experimental items which were adapted from the previously reported anomaly detection studies. The anomalous case was achieved in the same way as before, with the manipulation of the prior context affecting the status of the target critical word. The test items were re-written so that the anomalous terms were contained in a sentence that either required a high or low memory load to process.

Memory load was manipulated by inserting an additional phrase into the critical sentence, so that a high memory load sentence contained seven additional words than a low memory load, and introduced a new character, event or situation (in the example below "Sunday School"). The additional words were contained in the target sentence for the high memory load condition, and were placed in the introductory sentence for the low memory load condition. This meant that the overall length of the passage was held constant, as was the total number of words prior to the anomalous term. For example,

	LOW LOAD	HIGH LOAD
Introductory sentence	Recently some non-denominational schools have banned the telling of religious stories **that have been popular at Sunday School**.	Recently some non-denominational schools have banned the telling of religious stories.
Target sentence	The story of Jesus *on the cross / leaving the tomb* during the *resurrection* has been banned first.	The story of Jesus *on the cross / leaving the tomb* ,**which is a popular Sunday school story**, during the *resurrection* has been banned first.
Final sentence	Many church leaders are very angry.	Many church leaders are very angry.

In the above example, words in bold are the extra information, which may appear in the first sentence (low load), or in between the context manipulation and the critical word (high load). Words in italics are the context manipulation which would determine whether the target critical word, *resurrection* would be anomalous or not.

58 filler items were added to the test materials. These were also 3 sentences in length. 21 of these contained easy-to-detect role violations, for example the last sentence in an item about the Western Hebrides reads, "Many happy tourists travel to watch the *milkshakes*.", where milkshakes is obviously anomalous. Also, 8 items contained homophone substitutions, for example, "Gloria combed her *hare*..." (instead of hair), which were added to provide more variety and distraction from the target items. 29

filler items containing no anomalies or homophone substitutions were also included. This meant that the overall composition of items was 50% with anomalies and 50% without.

For a given participant, the test materials were presented in one of the four experimental conditions, i.e. as either anomalous or non-anomalous and either in the high or low memory load versions. Four files were produced so that each participant would be exposed to only one version of each given item.

During presentation the text appeared over 4 lines, and anomalies were never presented at the start or end of any line.

Participants

24 undergraduate psychology students from the University of Glasgow were tested, 4 male, 20 female. They were paid either £6 or participated for course credit.

Procedure

A Generation 5.5 Fourward Technologies Dual Purkinje Image eye-tracker (with an angular resolution 10 minutes of arc) was used. Text was displayed on a computer monitor approximately 80 cm from the participant giving ~4 characters/degree of visual angle. Gaze location was monitored every millisecond. A bite-bar and head rest minimised head movements. The tracking procedure was explained to participants at the start, and they were instructed to read for normal comprehension. A calibration procedure was completed at the start, and calibration was checked at the start of each trial. A fixation spot ensured that when the text appeared participants were looking at the start of the text.

The procedure was identical to the previous eye tracking study, where participants were asked to read the short stories in a manner supporting normal comprehension. They were informed that each passage would be followed by a comprehension question and that responses would be either "yes" or "no". Their main aim, it was explained, was to answer these questions correctly. After this had been explained, they were also informed that there might occasionally be anomalies in the text. Anomalies were defined as whole words that were out of context for some reason. To illustrate this they were given some examples, including the *Moses* illusion. They were asked to inform the experimenter whenever they noticed any such anomalies. A short practice block of four items (2 with anomalies, 2 without) was presented before the calibration and experimental stage, and participants were given the opportunity to ask questions. Emphasis was again placed on reading normally and answering the questions correctly.

Participants were positioned comfortably on the bite bar, and held a response button in each hand. After they had read a story they pressed either response button to progress to the comprehension question. They then responded by pressing either the left button for "yes" or the right for "no". If an anomaly had been detected they 'knocked' on the table, the experimenter turned the tracking beam off, and the participant came off the bite bar and explained what they thought the anomaly was. Participants always answered the question first and afterwards explained detected anomalies. The experimenter made a note of all anomalies that participants detected. The participant then went back on to the bite bar and the calibration procedure was repeated. After the tracking stage had been completed participants were debriefed about the nature of the study. They were then asked to complete a multiple-choice knowledge check questionnaire, to ensure that they understood the anomalies. They were also encouraged to make any comments about any aspects of the anomalies on the sheet and to the experimenter verbally.

Regions of analysis

Two regions were chosen for analysis, they were the critical region (which contained the target word which may be either semantically anomalous or not) and the post-critical region (which was the remainder of the sentence following the target word – this was held constant across all items at four words in length). The pre-critical region was not analysed because it varied in length across conditions. Since this region separated the context and critical region, it was decided that processing of the context region may also have been affected and so this region was not analysed either. The total number of words prior to the target word was constant across all conditions. For example,

	LOW LOAD	HIGH LOAD
Critical region	Recently some non-denominational schools have banned the telling of religious stories that have been popular at Sunday School. The story of Jesus *on the cross / leaving the tomb* during the *RESURRECTION*	Recently some non-denominational schools have banned the telling of religious stories. The story of Jesus *on the cross / leaving the tomb* ,which is a popular Sunday school story, during the *RESURRECTION*
Post-critical region	has been banned first	has been banned first.

Results

Question answering

The comprehension questions were answered correctly 92% of the time.

The final multiple-choice quiz was answered correctly over 98 % of the time.

Detection rates

The overall detection rate was 31%, which is lower than in the two previous tracking studies (46% and 49.7%). The average detection rate in the low load condition was

34% and 30% in the high load condition. There was no statistically reliable difference between high and low memory load detection rates.

Eye-tracking analysis

Fixations of less than 80ms were combined with adjacent fixations within one character position, and remaining fixations of less than 80ms were excluded from the analysis. Fixations of over 1200ms were also excluded. Data that contained two or more regions with no data in first pass, which may indicate tracker loss, was removed from the analysis. This affected less than 3.2% of the data. Four measures are reported that are normally taken to indicate early processing. *First fixation*, *first pass*, *first pass regressions*, and *regression path*. Also, three measures normally associated with later processing were included in the analysis, *number of fixations*, *regressions-in*, and *total time*.

The verbal reports were used to classify the data from the anomalous condition into anomaly detect and non-detect. A 2 x 2 ANOVA was carried out to compare high and low memory load and anomalous vs. non-anomalous conditions. Then, data were collapsed over load (i.e. combining high and low memory load data) and a one-way ANOVA comparing anomaly detect, non-detect and non-anomalous conditions was performed. Post hoc t-test comparisons between anomalous detect and non-anomalous, anomalous detect and non-anomalous, and non-detected anomalous and non-anomalous conditions were also carried out.

High and low memory load in anomalous and non-anomalous conditions

A 2 x 2 ANOVA compared high and low memory load in anomalous and non-anomalous conditions. There was only one reliable main effect observed for load and that was in the post-critical region with the number of fixations measure.

Post-critical region: There was a greater *number of fixations* made in the high load compared to the low load condition and this was reliable by subjects, $F1(1,23)= 5.97$ $p <0.023$, and items $F2(1,25) = 5.2$ $p <0.03$ (see figure 5.1), and this trend was observed in both anomalous and non-anomalous conditions. There was no interaction between memory load and anomaly. However there does appear to be an additive effect of load, with increased load similarly resulting in more fixations in both conditions.

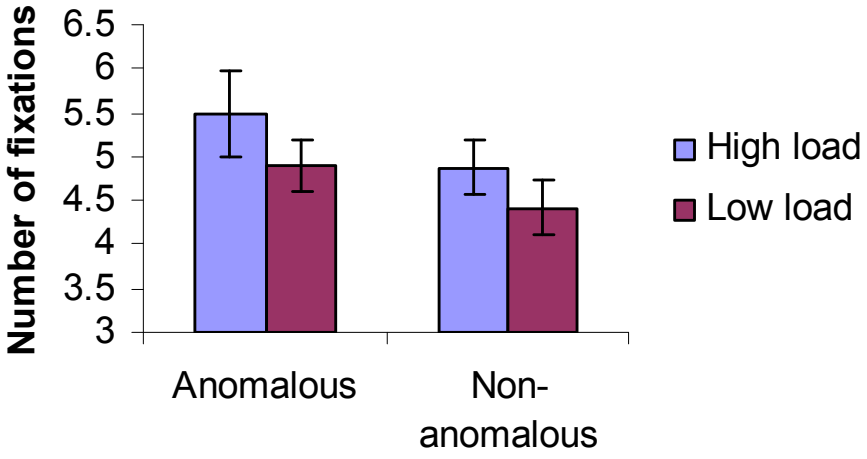


Figure 5.1: Average number of fixations by condition in the post-critical region

In summary, there was only one reliable effect of load in the eye movement data. There was a greater number of fixations in the post-critical region when anomalies were presented in high load versions compared to low load. In the next analysis, data were

collapsed over load (i.e. high and low combined) because there was a limited amount of data in each condition and also because there was little evidence for load having an effect. The data were then compared using one-way ANOVAs (with 3 levels: anomaly detect, non-detect, and non-anomalous).

Omnibus analyses: comparing anomaly detect, non-detect and non-anomalous

Using the verbal reports of participants, the data from the anomalous condition was separated into anomaly detect and non-detect. Once the data was separated into detect and non-detect, it was averaged by subjects and items as reported in previous analyses. Summary statistics are detailed in table 5.1.

Critical region: Two measures showed significant effects in this region, number of fixations and regressions in. The *number of fixations* varied significantly depending on whether an anomaly was present and detected, present and missed, or absent (see figure 5.2). This was significant by subjects $F1(2,46)= 3.62$ $p<0.035$, and items $F2(2,40)= 8.10$ $p <0.001$. *Regressions in* to the critical region was also significant by subjects $F1(2,46)= 4.3$ $p <0.019$, and items $F2(2,40) = 7.9$ $p <0.001$ (see figure 5.3).

Post-critical region: There were significant effects observed with first pass regressions, number of fixations, total time, and regressions in to this region. Significant effects were found with *first pass regressions* (see figure 5.4), by subjects $F1(2,46) = 7.63$ $p <0.001$, and items$F2(2,40) = 6.03$ $p <0.005$. Also, the *number of fixations* made in the post-critical region was significant by subjects $F1(2,46) = 6.34$ $p <0.004$, and items $F2(2,40)= 7.26$ $p <0.002$. The difference in *total time* spent in the region also approached significance by subjects $F1(2,46)= 2.85$ $p <0.068$, but was significant by items $F2(2,40)= 3.88$ $p <0.029$. There were significant effects observed with

regressions in to the post-critical region, which was significant by subjects $F1(2,46)= 3.8$ $p < 0.03$, and items $F2(2,40)= 3.2$ $p < 0.05$.

In summary, these analyses have demonstrated a significant pattern of disruption in the critical and post-critical regions of the text due to the presence of an anomaly that was either detected or missed, or absent. The data were then analysed using paired t-tests to permit three pairwise comparisons as with the previous studies. These compared first anomalous items when participants reported the anomalies with the control non-anomalous condition; secondly, anomalous items when these were detected with instances of non-detection; and finally, the anomalous non-detected items were compared to the control non-anomalous condition.

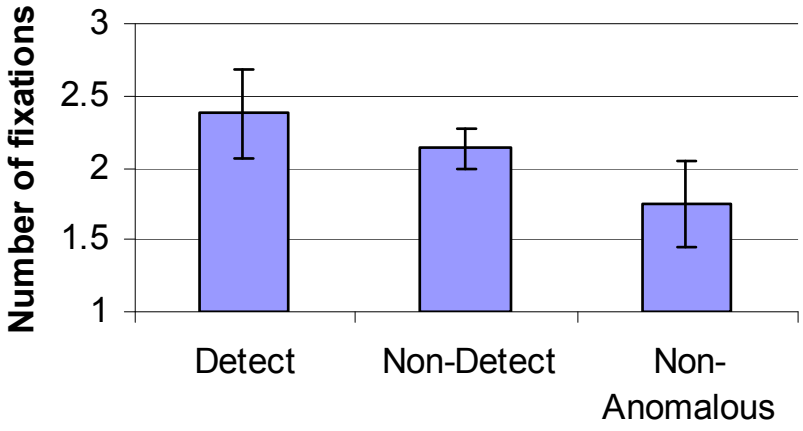


Figure 5.2: Number of fixations per condition in the critical region

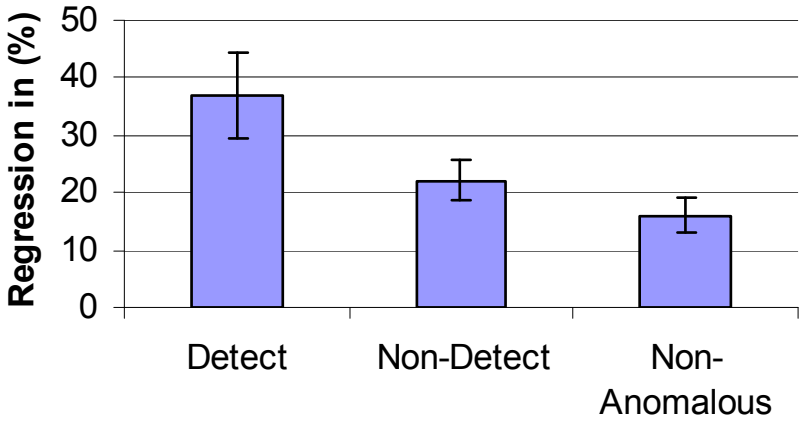


Figure 5.3: Regressions in per condition in to the critical region

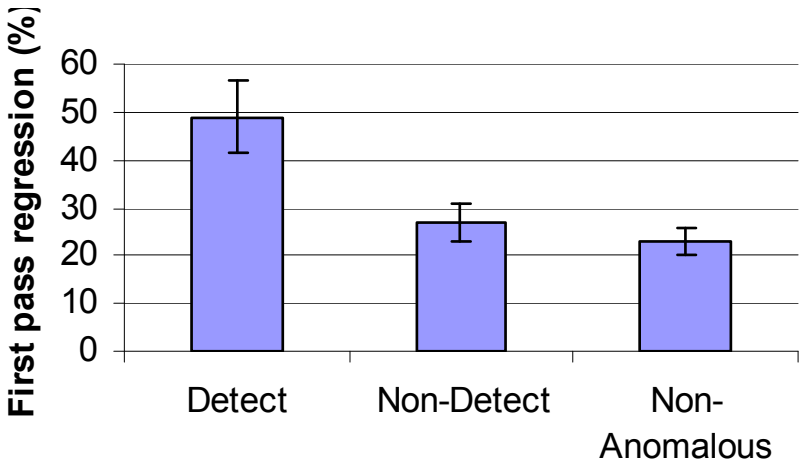


Figure 5.4: First pass regressions per condition in the post-critical region

Detected anomalies vs. non-anomalous controls

The data from the anomalous condition when anomalies have been detected were compared to the control versions. This comparison should highlight the effect that detection has on eye movements and fixations.

Critical region: There were a greater *number of fixations* in this region when an anomaly was detected compared to the control condition (detect =2.4, non-anomalous =

1.8), which was significant by subjects $t1(23)= 2.4$ $p<0.03$, and items $t2(20)= 3.96$ $p<0.001$. There was also a greater percentage of *regressions in* to the region when an anomaly was detected (detect = 37%, non-anomalous = 16%), which was significant by subjects $t1(23) = 2.3$ $p<0.03$, and items $t2(20)= 3.7$ $p<0.002$.

Post-critical region: A significant effect was observed with *total time* (detect = 1188ms, non-anomalous = 912ms), by subjects $t1(23)= 2.03p<0.054$, and items $t2(20)= 2.3$ $p<0.034$. There was also a significant effect found with *number of fixations* (detect =6.1, non-anomalous = 4.7), by subjects $t1(23)= 2.9$ $p<0.007$, and items $t2(20)= 2.8$ $p<0.01$. There was significant effect observed with *first pass regressions out* (detect = 50%, non-anomalous = 23%), by subjects $t1(23)= 3.01$ $p<0.006$, and items $t2(20)= 2.99$ $p<0.007$. *Regressions in* to this region was significant by participants (detect = 23%, non-anomalous= 13%), $t1(23)= 2.2$ $p<0.04$, and approaching significance by items $t2(20)= 2.0$ $p<0.058$.

Detected vs. non-detected anomalies

These analyses compared data from the anomalous condition separated into instances when the anomalies were reported by participants and when they were not.

Critical region: There were more *regressions in* to this region when an anomaly was detected compared to missed (detect = 37%, non-detect = 22%), which was borderline significant by participants, $t1(23)= 1.8$ $p<0.08$, but significant by items, $t2(20) = 2.8$ $p<0.01$.

Post-critical region: There was a greater *number of fixations* made in this region when an anomaly was detected compared to missed (detect = 6.1, non-detect = 5.0), which was significant by subjects $t1(23)= 2.23$ $p<0.04$, and by items $t2(20)= 2.9$ $p<0.009$.

There were significant effects observed with *first pass regressions out* (detect = 50%, non-detect = 27%), by subjects $t1(23)= 2.8$ $p<0.01$, and by items $t2(20)= 2.7$ $p<0.01$.

Non-detected anomalies vs. non-anomalous controls

These analyses used instances in the anomalous condition when participants failed to detect the anomalies compared to the non-anomalous version. This case directly tests whether failing to detect an anomaly leads to tracking patterns that are different from cases where an anomaly is not present.

Critical region: There were a greater *number of* fixations made in this region when an anomaly was missed (non-detect = 2.2, non-anomalous = 1.8), which was significant by participants $t1(23)= 2.8$ $p<0.01$, but not by items, $t2(25)= 1.7$ $p<0.1$.

Post-critical region: There was evidence that readers slowed down in this region with the *regression path* measure when an anomaly was missed (non-detect = 1183ms, non-anomalous = 975ms), which was significant by participants, $t1(23)= 2.7$ $p<0.01$, and approaching significance by items, $t2(25)= 1.9$ $p<0.059$.

Table 5.1: Summary data from anomalous detect, anomalous non-detect, and non-anomalous conditions (mean standard error) in the critical and post- critical regions

	First fixation (ms)	First Pass (ms)	First pass regression (%)	Regression path (ms)	Total time (ms)	Number of fixations	Regression -in (%)
Critical region							
Detect	206	248	20.2	369	493	2.4	37.2
	(19)	(24)	(4.9)	(44)	(85)	(0.30)	(7.4)
(sig diffs)							**t2**
Non-Detect	193	245	21.2	411	418	2.1	22.1
	(15)	(20)	(2.7)	(58)	(39)	(0.14)	(3.5)
(sig diffs)						**t1**	
Non-Anomalous	214	242	26.4	411	378	1.8	16.1
	(9)	(14)	(2.6)	(24)	(38)	(0.14)	(3.0)
Detect-Non-Anom (sig diffs)						**t1**, **t2**	**t1**, **t2**
Post-Critical Region							
Detect	219	532	49.5	1453	1188	6.1	23.1
	(15)	(63)	(7.4)	(231)	(118)	(0.54)	(3.8)
(sig diffs)			**t1, t2**			**t1, t2**	
Non-Detect	193	651	27.3	1183	994	5.0	14.4
	(9)	(36)	(3.8)	(72)	(67)	(0.38)	(2.6)
(sig diffs)				**t1**			
Non-Anomalous	198	635	23.0	975	912	4.7	12.6
	(9)	(28)	(2.8)	(59)	(58)	(0.27)	(2.3)
Detect-Non-Anom (sig diffs)			**t1, t2**		**t1, t2**	**t1, t2**	**t1**

Significant subjects and items t-test analysis are indicated where t1 and t2 are position between rows Detect and Non-detect; Non-detect and Non-anomalous; the final row in each region illustrates significant t-test comparisons made between Detected anomalies and Non-anomalous conditions.

In summary, these analyses show that anomaly detection causes some disruption. Anomaly detection results in increased regressions to, and more fixations made in, the critical region. In the post-critical region readers began to slow down, as evidenced by first pass regressions and regression path, resulting in increased total time and number of fixations made within this region. Unlike Experiments 2 and 3 there was some evidence to suggest that when readers failed to report anomalies they were more likely

to fixate in the critical region (subjects only), and slow down in the post-critical region (regression path measure, subjects only).

Observed power of the anomalous detect vs. non-anomalous comparison to the anomalous non-detect vs. non-anomalous comparison

As argued in Experiment 3, one way of determining the likelihood of not gaining a significant difference in the missed anomalies / non-anomalous comparison is by looking at the observed power for the anomaly detect / non-anomalous analyses and comparing the standard errors for the mean differences between each pair of analyses. The results are summarised in table 5.2 and in all cases the standard error in the anomaly missed / non-anomalous comparisons were lower than in the anomaly detect / non-anomalous comparisons. The observed power in the anomaly detect / non-anomalous comparisons were on the whole quite high, with *t*2 in number of fixations and regressions in for the critical region, and both *t*1 and *t*2 with the first pass regression measure in the post-critical region, all reporting power greater than .90. These results again suggest that the general lack of significant differences in the anomaly missed / non-anomalous comparison was not due to a lack of power.

Conclusions

The aim of Experiment 4 was to investigate the on-line processing differences between anomaly detection and non-detection. This was done by manipulating sentential load and it was expected that the differences between high and low load would differently affect the eye movement data. The overall detection rate was 31% and this was considerably lower than either Experiment 2 (46%) or Experiment 3 (49.7%), which suggested that sentential load did indeed affect overall detection rates. Inspection of

high and low load detection rates, however, revealed no significant differences between them.

Table 5.2: Table of standard errors and observed power for all major results in the anomalous detect / non-anomalous comparisons, and the associated standard errors in the anomaly missed / non-anomalous comparisons.

Region	Measure	SE (Detect)	Power	SE (Missed)
Critical	Number of fixations			
	• *t*1	0.26	0.74	0.14
	• *t*2	0.31	0.99	0.21
	Regressions in			
	• *t*1	8.9	0.74	3.8
	• *t*2	6.4	0.97	3.7
Post-Critical	Total time			
	• *t*1	135.8	0.63	62.4
	• *t*2	145.2	0.71	86.4
	Number of fixations			
	• *t*1	0.48	0.89	0.24
	• *t*2	0.76	0.85	0.28
	First pass regressions			
	• *t*1	8.8I	0.90	4.2
	• *t*2	6.6	0.89	4.6

However, it does seem reasonable, given the low detection rates, to assume that the load manipulation had generally made the task of anomaly detection harder. This increased difficulty was also reflected in the eye movement data. Significant effects were observed in the post-critical region, where there were a greater number of fixations made in the high load condition. This affected both the anomalous and non-anomalous conditions in the same way, which suggested that sentential load had an additive effect

in both conditions. This means that the materials in the high load condition were harder to read overall, but this was not affected by the presence or absence of an anomaly.

A specific focus of this study was whether increased task difficulty would reveal processing differences between detected and non-detected anomalous conditions. There was marginal evidence for the unconscious detection of missed semantic anomalies. In the crucial comparison between missed anomalous and non-anomalous data, there were more fixations made on the critically anomalous word when it was missed compared to the control condition. This disruption also appeared to persist into the post-critical region, where effects with the regression path measure suggested that participants were reading more slowly when anomalies were missed. This is admittedly slim evidence for unconscious anomaly detection. However, the effects are in the regions we would expect, and appear in the measures that we would expect, based on the results from detected trials. So, this does raise the intriguing possibility that there may be some registration of semantic anomalies within the processing system despite the fact that an anomaly has not been consciously detected. This issue obviously requires further investigation before firmer conclusions can be made.

The time course of anomaly detection, and the pattern of disruption caused by anomaly detection, was also clearly seen in the data. When anomalies are consciously detected there are effects in the critical region with late measures, and in the post-critical region with both early and late measures. This again suggested that the time course of detection is not immediate, but slightly delayed. There was no evidence here, unlike Experiment 3, that anomaly detection occurred when first encountering the anomalous word. This supports the view that processes involved in semantic analysis are not complete and exhaustive. This general pattern of results has now been replicated over three experiments and strengthens our confidence in their reliability.

Chapter 6: Incidental anomaly detection: Participants eye movements without forewarning of semantic anomalies

Partitioning the reading data into detected and non-detected cases has revealed major differences between the two conditions. In the present chapter, the same anomalies were used to investigate eye movement behaviour when there was no explicit instruction to report anomalies. In many experiments using anomalies as a way of probing processing, checks on overt detection are not used (e.g. Braze, Shankweiler, Ni, & Palumbo, 2002). Would the effects of anomalies on tracking with the present materials appear when no explicit instructions are given to report them? By removing the instruction to look for anomalies, it is possible that there will be a shift in strategy-driven depth of processing such that fewer anomalies will be noticed. There is some evidence to suggest that experimental task instructions can influence anomaly detection rates. Thus, Van Jaarsveld, Dijkstra & Hermans (1997) reported higher detection rates when task demands emphasised accuracy over speed of response. In a similar vein, Kamas, Reder & Ayers (1996) reported that participants in a single task procedure (just detect anomalies) outperformed participants who were in a dual task procedure (detect anomalies *and* answer questions). However, in both studies, participants were forewarned that passages contained semantic anomalies. Without forewarning, anomalies may go unnoticed by readers, especially when the anomalies have a good global fit to the context and so are harder to detect. It is thus possible that the monitoring instruction increases the likelihood of detection over what would normally occur in reading. This is an interesting question in its own right. However, it should be noted that if instructions do indeed influence detection rates, this does not invalidate our conclusions regarding differences between effects resulting from detection and failure to detect.

An eye-tracking study is presented here in which participants' eye movements were monitored while they read short stories, some of which contained semantic anomalies. Participants were not forewarned of the existence of these anomalies. A post-tracking anomaly detection questionnaire was administered to gauge if detection had occurred on-line.

Experiment 5: Incidental anomaly detection in the eye-movement data

Method

Design and Materials

This was a within-subjects design. Each experimental item was designed so that it could appear in either an anomalous or non-anomalous condition. A post-eye-tracking questionnaire determined whether participants had detected anomalies on-line or not. These responses permitted anomalous data to be classified as either detect or non-detect. On the basis of these classifications, within participants analyses (one-way 3 level ANOVAs) and pairwise comparisons were carried out.

There were 26 experimental items, each of which was produced as a non-anomalous version and an anomalous version. These materials were controlled for overall length of the passages (most passages were standardised at 32 words long), and length within each of the six regions was also controlled. The construction of the materials was essentially identical to those reported in chapter 4. For example:

> A North American jumbo jet was forced at gunpoint to land in Canada. The authorities 1/ {*negotiated / communicated*} 2/ with the scared and desperate 3/ hostages 4/ and calmed them down.5/ The siege lasted for two days. 6/

1. *Introduction region* – this introduced the story's theme and included the whole of the first sentence.
2. *Context region* - this region manipulated the anomalous nature of the later critical word.
3. *Pre-critical region* - this region linked the context and critical region.
4. *Critical region* - this was the critical word which was either anomalous or not.
5. *Post-critical region* - this region included all the words up to the end of the sentence.
6. *End region* - the end region is the final sentence.

The experimental items were randomly distributed amongst 78 fillers. The fillers were composed of 26 easy to detect anomalies and 52 non-anomalous stories, so that the overall composition was 50% anomalous, 50% non-anomalous. Two files were constructed. In one file, half of the experimental materials appeared in the anomalous condition, and half in the non-anomalous condition. In the second file, the half that were in the non-anomalous condition appeared in the anomalous condition, mutatis mutandis. Half of the participants saw file 1 and half saw file 2. Comprehension questions that required a 'yes' or 'no' response were asked on half of all stories presented, but this included *all* experimental items. During presentation the text appeared over 4 lines, and anomalies were never presented at the start or end of any line.

Participants

21 participants were recruited from the undergraduate population of the University of Glasgow. They were paid £6 for their participation. None had participated in previous anomaly studies.

Procedure

A Generation 5.5 Fourward Technologies Dual Purkinje Image eye-tracker (with an angular resolution 10 minutes of arc) was used. Text was displayed on a computer monitor approximately 80 cm from the participant giving ~4 characters/degree of visual angle. Gaze location was monitored every millisecond. A bite-bar and head rest minimised head movements. The tracking procedure was explained to participants at the start, and they were instructed to read for normal comprehension. A calibration procedure was completed at the start, and calibration was checked at the start of each trial. A fixation spot ensured that when the text appeared participants were looking at the start of the text.

Participants were asked to read the short stories in a manner supporting normal comprehension. Their main aim, it was explained, was to read for normal comprehension. A short practice block of four items was presented before the calibration and experimental stage, and participants were given the opportunity to ask questions. Emphasis was again placed on reading normally and answering the questions correctly. Participants were positioned in a comfortable position on the bite bar, and held a response button in their hand. After participants had read a story they pressed the response button to progress.

After the tracking stage participants were asked to complete a multiple-choice questionnaire. This was to determine whether or not they had detected the anomalies during the eye-tracking stage. All anomalous experimental items were presented in full and the anomalous word printed in bold to aid explanation. Participants responded by ticking one of three boxes, the first if they had detected the anomaly during the eye-tracking phase, a second box if they had only just noticed the anomaly (while reading

the questionnaire), and a third box to tick if they did not understand the anomaly. After completing the questionnaire the experimenter discussed the participants' responses with them and they were encouraged to provide any additional information / comments on the stories

Regions of analysis

The experimental items were split into six regions for the purposes of data analysis:

> A North American jumbo jet was forced at gunpoint to land in Canada. The authorities 1/ {*negotiated / communicated*} 2/ with the scared and desperate 3/ **hostages** 4/ and calmed them down.5/ The siege lasted for two days. 6/

As in the previously reported studies, the introduction and end regions are not reported in the formal analysis because they are large regions and served only to introduce or tie-up the stories. The four remaining regions are reported and these are the context (region 2), pre-critical (region 3), the critical region (region 4), and post-critical region (region 5).

Results

Question answering results

The comprehension questions were answered correctly 90% of the time.

Detection rates

39% of the experimental items were identified as anomalous in the post-experimental; questionnaire. This compares favourably to the previous studies. All detection rates to date are summarised in table 6.1. In comparison to the preliminary (Experiment 2) and second (Experiment 3) eye-tracking studies, the detection rate observed here is slightly

lower, but is also higher than Experiment 4 that manipulated sentential load. However, because participants were required to re-read the anomalies, and the anomalies were clearly pointed out to them, participants may have been influenced by a hindsight bias, and as such the detection rate could be an overestimate (see later discussion).

	Detection rates
Experiment 2: Preliminary investigation	46%
Experiment 3: Detecting anomalies	49.7%
Experiment 4: Load manipulation	31%
Experiment 5: Incidental detection	39%

Table 6.1: Summary of detection rates in all eye-tracking studies

Eye-tracking analysis

Fixations of less than 80ms were combined with adjacent fixations within one character position, and remaining fixations of less than 80ms were excluded from the analysis. Fixations of over 1200ms were also excluded. Data that contained two or more regions with no data in first pass, which may indicate tracker loss, was removed from the analysis. This affected less than 4% of the data. Four measures are reported that are normally taken to indicate early processing, *first fixation duration*, *first pass*, *first pass regressions*, and *regression path*. Also, three measures normally associated with later processing were included in the analysis, *number of fixations*, *regressions-in*, and *total time*.

The post-tracking questionnaire was used to classify the data from the anomalous condition into anomaly detect and non-detect. One-way ANOVAs were carried out that compared performance on missed anomalies, detected anomalies, and the control non-anomalous condition. Follow-up t-tests were also carried out comparing the three conditions as before, and these are reported in subsequent sections. It was hypothesised

that incidental anomaly detection should show a similar pattern of disruption in the anomaly detect data compared to both the control non-anomalous and anomaly miss data.

Omnibus analyses of anomaly detect, non-detect and non-anomalous data

Detection was assessed via a post-tracking questionnaire that allowed participants to indicate which semantic anomalies they had detected during the main reading task. These responses were used to partition anomalous data into detect and non-detect cases. One-way ANOVAs were performed that compared non-detected anomalies, detected anomalies, and the control non-anomalous condition. The main results are reported below and follow-up t-test comparisons are reported in the following sections. There was one main effect found with the regression path measure (participants only) in the critical region. There were no effects observed in the context, pre- and post-critical regions.

Critical region: There was only one main effect with the *regression path* measure which was significant by subjects $F1(2,34)= 3.30$ $p<0.049$ but not by items $F2(2,48)=0.88$ $p<0.4$ (non-sig.) (see figure 6.1) No other analyses were significant.

The data were then analysed with a series of paired t-tests. The detected anomalies were compared to the non-anomalous controls, the detected and non-detected anomalous data were compared, and finally, the non-detected anomalous data and non-anomalous controls were also compared, as in previous experiments.

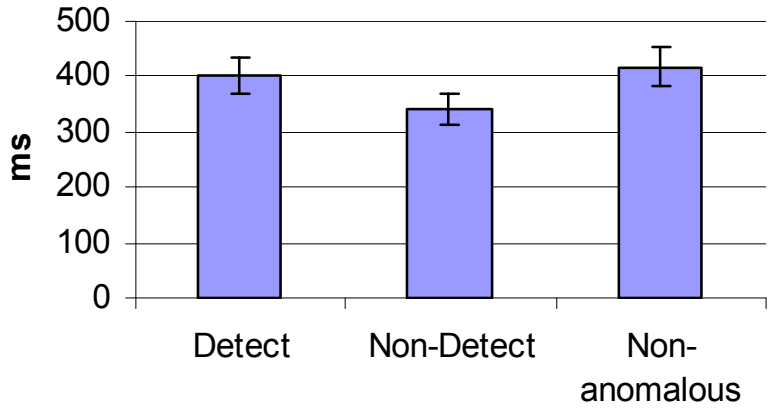


Figure 6.1: Regression path (ms) in the critical region by condition

Detected anomalies vs. non-anomalous controls

These analyses should illustrate the disruptive effects of anomaly detection on the eye movement data. Data was classified as being detected or non-detected based on the post-tracking questionnaire responses and compared to the non-anomalous condition. There were significant effects observed with first pass and the first fixation measures in the critical region.

Critical region: There was a significant difference by participants between detected anomalies and non-anomalous data with the *first pass* measure, (detect = 350ms, non-anomalous = 293ms), which was significant by subjects $t1(18)= 2.07$ $p<0.05$, but not by items $t2$ $p<0.1$ns. There was also a significant effect found by participants with *first fixation duration* which was longer for detected anomalies than non-anomalous controls (detect = 278ms, non-anomalous = 261ms), $t1(18)= 2.2$ $p<0.04$, but again was not significant by items $t2$ $p< 0.2$ ns. Both measures indicate that when an anomaly was

Table 6.2: Summary statistics (mean, standard error) from anomalous detect, anomalous non-detect, and non-anomalous conditions in the context pre-critical critical and post-critical regions with significant differences for subject (t1) and item (t2) analyses indicated

	First fixation (ms)	First Pass (ms)	First pass regression (%)	Regression path (ms)	Total time (ms)	Number of fixations	Regression-in (%)
Context region							
Detect	256 (14)	352 (27)	10.3 (3.7)	431 (41)	459 (56)	1.93 (0.26)	8.2 (3.4)
Non-Detect	259 (14)	414 (37)	11.6 (2.9)	482 (43)	545 (52)	2.23 (0.20)	11.8 (2.6)
Non-Anomalous Detect-Non-Anom	275 (10)	401 (19)	14.0 (2.1)	508 (24)	519 (36)	2.02 (0.13)	12.9 (2.9)
Pre-Critical region							
Detect	256 (15)	974 (84)	8.5 (2.8)	1289 (103)	1304 (108)	5.48 (0.44)	27.9 (6.1)
Non-Detect	260 (13)	1036 (59)	9.1 (2.8)	1204 (89)	1359 (114)	5.74 (0.60)	22.0 (6.3)
Non-Anomalous Detect-Non-Anom	258 (10)	952 (62)	15.5 (3.0)	1236 (80)	1354 (108)	5.57 (0.44)	20.4 (4.3)
Critical region							
Detect	278 (13)	350 (27)	8.8 (3.3)	403 (33) t1	404 (41)	1.52 (0.15)	8.4 (2.7)
Non-Detect (Sig diffs)	287 (18)	304 (18)	4.2 (1.9) t1	339 (28)	379 (29)	1.41 (0.11)	9.9 (2.7)
Non-Anomalous Detect-Non-Anom (sig diffs)	261 (10) t1	293 (14) t1	8.9 (2.2)	418 (33)	355 (30)	1.43 (0.11)	12.5 (2.7)
Post-critical region							
Detect	252 (10)	785 (99)	18.8 (6.0)	1152 (128)	1008 (96)	4.09 (0.38)	5.7 (2.1)
Non-Detect	265 (15)	769 (53)	24.2 (5.5)	1067 (77)	1002 (91)	4.09 (0.33)	4.6 (1.6)
Non-Anomalous Detect-Non-Anom	244 (10)	708 (44)	17.2 (3.8)	1000 (93)	935 (80)	3.89 (0.30)	6.0 (1.9)

- t1 / t2 between pairs of mean scores for detect and non-detect; non-detect and non- anomalous represent significant differences between pairs And t1 / t2 in rows labelled Detect-Non-Anom represent significant differences between mean scores in that region for that measure

detected readers spent more time in this region. These also suggest that detection occurred as soon as the anomaly was encountered.[6]

Detected vs. non-detected anomalies

Data from the anomalous condition, separated into instances of detection and non-detection, were then compared. Detection was again determined by responses to the questionnaire. These analyses should indicate if detection and non-detection affected eye movements differently. There was only one significant effect by participants observed in the ***critical region*** with *regression path*, (detect = 403ms, non-detect= 339ms), $t1(17)= 2.48$ $p<0.024$, which again was not significant by items $t2(24)=0.99$ $p<0.3$ns. This suggests that participants spent longer reading the critical word when it was identified as anomalous.

Non-detected anomalies vs. non-anomalous controls

The final analyses compared instances when participants failed to detect anomalies to non-anomalous controls. Non-detection was determined by responses given to the post-tracking questionnaire. There was only one significant effect reported with *first pass regressions out of* the ***critical region***. There was a lower proportion of regressions when an anomaly was missed (non-detect = 4.2%, non-anomalous = 8.9%) compared to the control. This was significant by participants, $t1(20)= 2.09$ $p<0.049$, and approached significance by items, $t2(25)= 1.94$ $p<0.06$. This suggests that there was less re-reading, or careful reading, of the critical word when an anomaly was missed compared to a control condition.

[6] The lack of a similar effect in earlier regions suggests that this effect is not a reflection of faster reading which leads to missed anomalies.

Summary of results

There were few differences between anomalous detect, non-detect, and non-anomalous conditions. There was only one main effect found in the omnibus analyses (participants only), where the regression path measure showed an effect when an anomaly was present and detected, present and non-detected or absent. Follow-up comparisons suggested that there was a significant difference between detected and non-detected anomalous conditions, with detected anomalies read more slowly. Because regression path is normally taken as an index of early processing, this also suggests that detection is occurring immediately in the critical region, rather than delayed as has been reported previously. When data for detected anomalous was compared to non-anomalous conditions, there were significant effects reported for first pass and first fixation in the critical region (participant analyses only). The average reading time / length of first fixation in both was longer when anomalies were detected. Again these results suggest that readers detected these anomalies immediately. Finally, there were significantly fewer first pass regressions in the critical region when anomalies were missed compared to the control condition. The overall pattern of effects indicates that detection occurred quickly and was confined to the critical region.

Comparing the main effects reported in Experiment 3 to those found in Experiment 5: a question of power

In Experiment 3, when participants were forewarned about the presence of semantic anomalies, there was a clear pattern of effects observed between detected anomalies and non-anomalous controls. These effects were reported in Chapter 4 and the observed power, and mean standard errors, for these effects were used to estimate the likelihood of not finding a significant effect with anomalies that were not-detected (compared to control versions). The following argument was made: if the standard errors from non-detected anomalies/non-anomalous comparisons were lower than the detected anomalies/non-anomalous comparisons, then the likelihood of not finding a significant effect in the non-detected cases is at least equal to that observed in the detected cases. A similar argument is presented in this section, which compares the main effects reported in Experiment 3 to Experiment 5. The significant effects reported in Chapter 4 between detected anomalous/non-anomalous controls are detailed in table 6.3, along with standard error and observed power. The standard errors for the same comparisons from Experiment 5 are also detailed. It can be seen that for all effects bar one (regression in to the pre-critical region) the standard errors in Experiment 5 are lower than in Experiment 3. This suggests that the lack of significant effects in the main study is not due to a lack of power, or lack of sensitivity in the measures. Rather there were in fact no significant differences in the data.

Table 6.3: Comparison of standard errors and power from main effects reported in Experiment 3 in comparisons between detected anomalies and non-anomalous controls to the corresponding analyses in Experiment 5

		Experiment 3 (detected anomalous vs non-anomalous)		Experiment 5 (detected anomalous vs non-anomalous)
Region	Measure	SE	Power	SE
Critical	Total time			
	• t1	67	0.72	31
	• t2	79	0.99	38
	Number of fixations			
	• t1	0.24	0.80	0.12
	• t2	0.68	0.79	0.13
Post-Critical	First pass regressions			
	• t1	6.1	0.95	3.8
	• t2	6.5	0.99	5.8
	Regression path			
	• t1	121	0.86	108
	• t2	411	0.67	106
	Total time			
	• t1	81	0.82	78
	• t2	167	0.70	68
Context	Number of fixations			
	• t1	0.21	0.64	0.18
	• t2	0.31	0.90	0.18
Pre-critical	Regression-in			
	• t1	4.7	0.96	5.3
	• t2	5.8	0.99	6.5
	Number of fixations			
	• t1	0.40	0.64	0.26
	• t2	1.4	0.75	0.46

Discussion and further comparative analyses

The aim of Experiment 5 was to investigate the influence of task instructions on anomaly detection rates and eye movement data. Participants were not forewarned that anomalies were in the text, nor were they aware that they were relevant to the experiment. A post-tracking questionnaire was administered to gauge if participants had detected anomalies during the eye-tracking phase. The responses from the

questionnaire were used to separate the anomalous data from the anomalous condition into detected and non-detected.

The overall rate of incidental anomaly detection was 39%. This was lower than the rates reported in Experiment 2 (46%) and 3 (49.7%). Lower detection rates would be expected where the anomalies are hard-to-detect and when participants were not expecting them. It is possible that participants adopted different reading strategies when the task demands were different. So, if instructions emphasise detection, readers may adopt a deeper processing strategy resulting in higher detection rates. If anomalies are not expected, processing may be shallower resulting in fewer detections overall.

There is supporting evidence that task instructions can modulate rates of anomaly detection (Kamas, Reder, & Ayers, 1996; van Jaarsveld et al. 1997). For example, van Jaarsveld et al (1997) manipulated task instructions that emphasised either the speed or accuracy of responses in an anomaly detection task. They noted a substantial drop in detection rates when the task instructions were to respond quickly (18.3% from 32.9%). The drop in detection reported in Experiment 5, compared to Experiments 2 and 3, is more modest in comparison. However, at 39%, detection still seems fairly high, especially since participants were not expecting anomalies and is very similar to previously reported rates of detection. As such, the effects observed in Experiments 2, 3, and 4 may reflect the task demands, which emphasises detection and reporting of anomalies, rather than just detection. This will be considered further in Chapter 7. Alternatively, it is possible that the 39% detection rate is an overestimate of the true rate of on-line detection. We believe that this may be so because the accuracy of responses provided in the post eye tracking questionnaire may have been impeded by both memory failure and hindsight bias.

Memory inaccuracies may have occurred because participants read a total of 103 stories during the eye tracking phase of the experiment, but only 26 of these were experimental items. In the post-tracking questionnaire these 26 items were re-presented. Participants were expected to recall them, and also whether they had detected an anomalous word at that time. This was a very demanding task and participants may have struggled to perform this task accurately. To aid memory retrieval and explanation of the anomalies, the experimental items in the questionnaire were re-presented whole and the anomalous word printed in bold. However, this may have made the task of accurately remembering anomalies more difficult as participants may have been influenced by hindsight bias. Hindsight bias is the feeling of overconfidence in one's own knowledge after the correct answer has been made public (e.g., Slovic & Fischhoff 1977). For an example in relation to our items, if asked the question, "did you know that victims are not sent to prison?" most would say that they did, but then still fail to notice such an anomaly in a story under normal reading conditions! So, by directly presenting the answer to participants in the questionnaire, this may have increased their confidence that they had actually noticed the anomaly originally. This bias, coupled with memory inaccuracies, could have in turn inflated the detection rate. In fact, participants may have responded that they had noticed the anomaly in the tracking phase for various different reasons, such as, (a) they just remembered reading the passage (& not specifically detecting the anomaly), or (b) if they felt that they *should* have noticed the anomaly, or even (c) if at the time of reading, they just thought that *something* was wrong, without actually being able to identify it. Given these concerns it is possible that the detection rate is an overestimate of what had actually happened at the time of reading.

In the case of anomalies that went undetected, there was no evidence for system registration of the anomalous word. If this had occurred then a similar pattern of effects

would have been observed as those reported in detected anomalous and control comparisons. Instead, as in previous experiments, there were few signs to indicate that missed anomalies were processed differently from controls. The only significant effect reported was that there were fewer first pass regressions made in the critical region when anomalies were missed compared to controls. This might indicate a superficial reading strategy in missed cases, however we would argue against this interpretation. If participants had been reading superficially in cases where anomalies were missed, and this was the reason that they were missed, then we would have expected to see effects in other regions of the text, especially in regions prior to the critical word as well. Also, if a superficial strategy was the cause of missing anomalies, not only should there have been more differences observed when missed cases were compared to controls, but also when compared to detected anomalies. There were no significant differences observed in these comparisons. Therefore, it seems safe to conclude that when anomalies were missed, there was no evidence for registration of the anomaly, nor was there any evidence that failure to detect was due to inattentive processing.

The apparent time course of incidental anomaly detection was very different from that found in Experiments 2, 3 and 4, where the task instruction was to detect and report anomalies. Incidental anomaly detection resulted in longer initial fixations and longer first pass reading times in the critical region compared to controls. Detected anomalies also resulted in longer reading times in comparison to missed anomalies in the same region, as evidenced by the regression path measure. However, characteristically, the results in Experiments 2, 3, and 4 were that there were significant effects with late measures in all regions, and early measures in the post-critical region only. These had been interpreted as reflecting that anomaly detection was slightly delayed and caused severe disruption outside the critical region as well as within. However, in Experiment

5, anomaly detection appeared to occur immediately, and caused no disruption outside of the critical region.

While the possibility of immediate detection is entirely feasible, the lack of any disruption beyond the critical region was unexpected. This is particularly important in comparison to the eye tracking studies already reported in this thesis where disruption has been consistently reported outside the critical region. This has also been the case in other eye tracking anomaly detection studies (Braze, et al., 2002; Ni, et al., 1998; Daneman, et al., 2006). For example, Braze et al. recorded participants' eye movements while they were reading pragmatic anomalies, such as "The cats won't usually *bake* the food we put on the porch", where the word *bake* is used anomalously within the context of the sentence. Similar to the experiment reported here, participants were neither forewarned that anomalies were in the text, nor that these were relevant to the experiment. Braze et al. reported that in regions following a pragmatic anomaly there was a gradual increase in regressive eye movements which reached a maximum at the end of the sentence. They inferred that these eye movements reflected anomaly detection, and that detection caused progressive disruption throughout the remainder of the sentence. While the pragmatic anomalies might be considered to be easier to detect than our hard-to-detect anomalies, the experimental task is very similar. Even under conditions where they did not clearly separate out instances where anomalies had been detected from instances where they were not, they recorded significant effects in the eye movement data. It was decided therefore, to re-analyse the present data simply by comparing all the data from the anomalous condition (detected and non-detected) to the control comparisons. The reason for doing this was two-fold. Firstly, because Braze et al. reported clear effects with simple global analyses, similar effects may be observed in our data if analysed in the same way. Secondly, as was discussed above, because there were serious concerns over the accuracy of the responses in the post-tracking

questionnaire, global comparisons may be more appropriate. Global comparisons were carried out by comparing anomalous (with detect and non-detect combined) to non-anomalous data, in each region and for each measure.

Global Analyses of Anomalous to Non-Anomalous Conditions in Experiments 5 and 3

The anomalous data in Experiment 5 were compared to the non-anomalous data in a series of paired t-tests for each measure per region (descriptive statistics are summarised in table: 6.4). There were few significant effects observed, however this time they did extend beyond the critical region. In the **critical region** the *first pass* measure approached significance for items only (anomalous = 281ms, non-anomalous = 225ms), $t1(20)$- 1.5 $p<0.2$ (non-sig.), $t2(25)= 1.9$ $p<0.064$. In the **post-critical region** *first pass* again approached significance by items only (anomalous = 626ms, non-anomalous= 534ms), $t1(20)=1.1$ $p<0.3$ (non-sig.), $t2(25)= 1.9$ $p<0.06$. *Total time* also approached significance by participants in this region, (anomalous = 768ms, non-anomalous = 489ms), $t1(20)=1.9$ $p<0.068$, $t2(25)= 0.2$ $p<0.9$ (non-sig.). So, even when anomalous and non-anomalous conditions were compared globally, there were few significant effects. However, the few that were observed all suggested that the anomalous condition resulted in more disruption. Furthermore, disruption now appeared to extend beyond the critical region into the subsequent post-critical region suggesting that disruption is not confined to the critical region only. This study was conducted under conditions where participants had not been forewarned of the presence of anomalies and it is possible that clearer effects would be observed from global analyses when participants had been instructed to detect and report anomalies. Alternatively, global analyses of anomalous and non-anomalous conditions may have the effect of obscuring effects in the data. To investigate this, the same analyses were performed on the data collected in Experiment 3.

The data from Experiment 3 (reported in Chapter 4) were re-analysed comparing anomalous and non-anomalous conditions. The data from this particular experiment was chosen because 22 of the experimental items were identical to the items used in the present experiment. The main effects that had been reported in Experiment 3 were that in the critical and post-critical regions, more time was spent reading, and more fixations were made, when an anomaly was detected. Also, in the post-critical region similar effects were observed with regression path and first pass regressions out. However, global analyses of anomalous and non-anomalous data from Experiment 3 revealed only one significant effect. That was in the **post-critical region**, where a significant effect with *first pass regressions* was found. Here there were more regressive eye movements when an anomaly was present (anomalous = 34%, non-anomalous = 26%), $t1(27)= 2.9$ $p<0.008$, $t2(25)= 2.6$ $p<0.02$. Descriptive statistics for anomalous and non-anomalous conditions are summarised in table 6.5.

In summary. What is apparent from these two further global analyses is that simply comparing anomalous and non-anomalous data can, in some circumstances, obscure effects. Without taking into consideration conscious awareness, as in overt anomaly detection, we run the risk of missing important effects in a data set. This is demonstrated in the re-analyses of the data from Experiment 3. Global comparisons of anomalous to non-anomalous conditions revealed few differences in the data. This could have been interpreted as a reflection of the fact that hard-to-detect anomalies were not detected on-line because they had little apparent impact on eye movement data. However, when anomalous data was partitioned into instances when items were or were not detected, a different picture emerged, with effects reported in different text regions and with different measures. It would appear, therefore, that combining detect and non-detect data resulted in an effect of them 'cancelling' each other out.

From the analysis of anomalous vs. non-anomalous conditions, regardless of detection, it also seems as though the request to report anomalies had little impact on the data. But, if conscious detection is taken into account, a different pattern emerges. Of course, given the questions raised over the validity of the post-tracking questionnaire we can only draw tentative conclusions, but it does seem that a change in experimental instructions can affect the processing strategy adopted by participants and a change in the rate of anomaly detection. Evidence to support this claim is offered both in the lower rates of detection reported here and in the results of Kamas et al., (1996), and van Jaarsveld et al., (1997).

The time course of incidental anomaly detection appeared to be immediate in the present study, rather than delayed as reported in previous chapters. It is obviously important to know which items are detected when materials are hard-to-detect anomalies. However, to gather this data it is necessary that participants be forewarned of the presence of anomalies. This in turn affects the eye movement data. On the other hand, if participants are not forewarned of upcoming anomalies they may not notice them. Therefore, the task demands may affect reading strategy in that the request to detect anomalies may lead to careful reading, while no explicit request to detect may lead to a more shallow reading strategy.

Table 6.4: Summary data of anomalous and non-anomalous global analysis (mean, standard error) in the context, pre-critical, critical and post- critical regions with significant differences for subject (t1) and item (t2) analyses indicated

	First fixation (ms)	First Pass (ms)	First pass regression (%)	Regression path (ms)	Total time (ms)	Number of fixations	Regression -in (%)
Context region							
Anomalous	238 (28)	363 (37)	13 (2.8)	424 (45)	454 (49)	1.9 (0.17)	12.3 (24)
Sig Diff							
Non-Anomalous	247 (28)	347 (33)	17 (2.6)	434 (41)	431 (47)	1.7 (0.14)	13.5 (3.0)
Pre-Critical region							
Anomalous	201 (22)	701 (106)	10 (2.4)	903 (148)	962 (160)	5.4 (0.49)	20.0 (3.8)
Sig Diff							
Non-Anomalous	211 (22)	719 (94)	15 (3.0)	990 (114)	1073 (162)	5.5 (0.43)	21 (4.1)
Critical region							
Anomalous	244 (22)	281 (21)	9 (2.5)	387 (34)	337 (33)	1.1 (0.10)	13.3 (3.0)
Sig Diff		t2					
Non-Anomalous	200 (23)	225 (27)	15 (3.3)	288 (41)	308 (37)	1.1 (0.10)	15.2 (3.7)
Post-critical region							
Anomalous	198 (23)	626 (61)	20 (4.1)	698 (120)	768 (100)	3.9 (0.31)	6.6 (1.5)
Sig Diff		t1			t1		
Non-Anomalous	200 (21)	534 (67)	16 (3.7)	733 (100)	489 (85)	3.8 (0.30)	5.3 (1.7)

Table 6.5: Summary data of anomalous and non-anomalous global analysis for Experiment 3 (mean, standard error) in the context, pre-critical, critical and post-critical regions with significant differences for subject (t1) and item (t2) analyses indicated

	First fixation (ms)	First Pass (ms)	First pass regression (%)	Regression path (ms)	Total time (ms)	Number of fixations	Regression -in (%)
Context region							
Anomalous	276 (8)	435 (21)	25 (3.1)	584 (34)	706 (55)	2.6 (0.2)	36 (4.1)
Sig Diff							
Non-Anomalous	279 (7)	466 (24)	26 (3.1)	595 (29)	696 (54)	2.7 (0.2)	30 (3.8)
Pre-Critical region							
Anomalous	256 (9)	932 (36)	21 (2.5)	1407 (69)	1814 (91)	7.5 (0.5)	51 (3.5)
Sig Diff							
Non-Anomalous	265 (7)	847 (36)	21 (2.4)	1390 (65)	1718 (99)	7.7 (0.6)	47 (3.3)
Critical region							
Anomalous	275 (9)	320 (16)	24 (3.1)	435 (24)	584 (55)	2.0 (0.3)	30 (3.0)
Sig Diff							
Non-Anomalous	271 (6)	313 (8)	23 (2.8)	454 (27)	581 (52)	1.9 (0.2)	26 (2.7)
Post-critical region							
Anomalous	270 (6)	791 (38)	34 (3.1)	1430 (114)	1367 (91)	5.4 (0.4)	23 (2.8)
Sig Diff			t1,t2				
Non-Anomalous	258 (6)	820 (32)	26 (2.8)	1391 (96)	1289 (78)	5.6 (0.4)	20 (2.8)

Chapter 7: Detection and non-detection of semantic anomalies as reflected in ERP measures

The absence of effects for undetected anomalies in Experiments 2, 3 and 4 leads to the question of whether we register undetected anomalous words at all at any level. One approach to this question is to examine dynamic measures of brain activity, such as ERPs measured from EEG data. While, inevitably, there will be some brain activity associated with the processing of even undetected anomalies, here we attempt to see what activity takes place that is typically associated with semantic processing. In this chapter we report an ERP study where participants were asked to detect and identify visually presented semantic anomalies embedded in two-sentence stories. The same logic is followed as before: after each trial, participants have the opportunity to report detecting an anomaly, or to indicate that they did not notice one. This task permits comparisons between the ERP waveforms for detection, non-detection, and baseline conditions. The empirical questions explored are; what type of waveform do we observe when hard-to-detect anomalies are reported by participants, compared to non-detection and baseline conditions, and are there any observable effects when participants fail to detect anomalies compared to a control condition? We begin with a review of both traditional and newly emerging findings in the ERP semantic anomaly literature, and consider in detail what we would expect to find with hard-to-detect semantic anomalies.

Language-sensitive components in ERPs

The electroencephalogram (EEG) is a non-invasive neuroimaging technique that provides a real-time measurement of neural activity. It measures the summed post-synaptic potential of groups of neurones. When these potentials are analysed they are

time-locked to a sensory, motor or cognitive event which is assumed to elicit this activity. However, the raw EEG signal is not sensitive enough to record subtle changes of mental activity in single trials, and therefore an average waveform from multiple trials is calculated which increases the strength of the signal associated with the cognitive event, and decreases any noise in the data associated with background or non-event related activity. This measurement is called an event-related potential (ERP) and the resultant waveform contains a series of positive and negative deflections. The waveform may be described, and classified, by characteristics such as its *latency*, which is the time point at which the wave reaches its peak; its *polarity*, whether or not the waveform is positive or negative deflection; and *scalp distribution* (or topography) of the neural activity. This has permitted researchers to investigate the temporal properties of cognitive processes, such as language comprehension.

N400s elicited by semantic anomalies

In a seminal study, Kutas & Hillyard (1980) identified the N400, an ERP component highly related to semantic processing. In a sentence reading task, words were presented serially on a computer screen and sentences ended in either a semantically predictable or incongruent way, for example, [1]

[1] He spread the warm bread with *socks* (or butter)

They observed that easy-to-detect semantic violations, such as a non-edible item (*socks* in [1]) referred to as food, produced an exaggerated negative going waveform, beginning at approximately 200msec after stimulus onset and peaking at 400msec. The scalp distribution was mostly posterior (more pronounced over parietal, posterior temporal, and occipital, rather than frontal sites), and larger and more prolonged over the right than left hemisphere. Sentences that ended with a congruous word elicited a

more positive (or less negative) going waveform instead. In fact, a negative deflection (N400) occurs with all content words, but is exaggerated in anomalies. Thus the difference between the magnitude of the N400 in response to *butter* and *socks* in [1] is the N400 *effect*.

The N400 effect is not only elicited by semantically anomalous words in a terminal position, but also by words that appear in the middle of the sentence (Kutas & Hillyard 1983). Furthermore, it is not restricted to anomalies. The N400 has been demonstrated to be sensitive to expectancy, as defined as Cloze probability (Kutas & Hillyard 1980). Cloze is an off-line technique, where the expectancy of a word is assessed using a sentence completion task. The N400 amplitude is inversely related to the probability that a word will be used in a Cloze task. Expectancy, however, is not the same thing as contextual constraint. A sentence can highly constrain an upcoming word; however the cloze probability may still be independently high or low. So, for example, the sentence, "the paint turned out to be the wrong …" is likely to be finished with the word *colour*. The word *colour* is constrained by the context, and would also have a high cloze probability. However, an equally acceptable word, such as *shade*, could also be used, not violate the context, but has a much lower cloze probability. Through independent manipulation of both factors Kutas and colleagues (Kutas & Hillyard 1984; Kutas, Lindamood & Hillyard 1984) demonstrated that N400 effects were not due to violating the expectation of a non-presented word, but appeared instead to represent the ease with which a word can be integrated into an unfolding sentence. Subsequent research verified that the N400 is in fact a default response, with most open-class words eliciting it, with the amplitude and latency of the waveform modulated by the experimental manipulation (Kutas & Van Petten 1994).

The N400 is not only sensitive to contextual constraints in sentences, but is also sensitive to the semantic relations between pairs of words. This has been demonstrated with paradigms such as lexical decision and category judgement tasks. In lexical decision tasks participants have to decide if a letter-string constitutes a real word or not, and in category judgement tasks whether a word is an example from a particular category. Words which are semantically unrelated to the prime generally elicit an N400 component, compared to semantically related word pairs (e.g. Holcomb 1988; Heinze, Muente, & Kutas 1993). This effect has been observed with auditory priming (Holcomb and Neville 1990), and with line drawings that take the place of a final word, which, within the sentence, depict a semantically congruent or incongruent figure (Kutas & Van Petten 1990). The N400 has also been shown to be sensitive to repetition, with a reduced N400 amplitude for both words and whole sentences that are repeatedly presented (Besson et al 1992; see Kutas & Van Petten 1994), and to the class of words, so that open-class words (nouns, verbs, adjectives, *-ly* adverbs) elicit larger N400s than closed-class words (pronouns, prepositions, articles etc.) (Van Petten & Kutas, 1991).

At the other extreme, discourse-semantic N400 effects have been demonstrated by Van Berkum, Hagoort, & Brown (1999). They presented participants with sentences that were either congruous or incongruous, as determined by the prior context. So, for example in [2] we expect the mouse to move quickly rather than slowly.

[2]Context Sentence: The cat entered the room suddenly, startling a mouse which had found a bit of cheese in the corner.

Critical Sentence: The mouse {*quickly / slowly*} returned to its hole.

Van Berkum and colleagues reported that the incongruous verb, *slowly,* elicited a larger N400 than *quickly*, when preceded by the prior context. When these sentences were tested in isolation, however, both sentences elicited equivalent waveforms. Beyond discourse-based effects, N400 effects have been demonstrated with statements that contain information that violates real-world knowledge. Hagoort, Hald, Bastiaansen, & Petersson (2004) presented their participants with statements, such as, "Dutch trains are *white*", when they are in fact *yellow*. The word *white* elicited an enhanced N400 compared to *yellow*.

The amplitude of the N400 in response to an individual word is, therefore, modulated by the context in which it appears, be that a single word, sentence, or discourse. It has been consistently demonstrated that words that are either expected or semantically related elicit a smaller amplitude, compared to unexpected or unrelated words. The default assumption, most common to a large part of the studies, is that the N400 reflects processes of integration into context, so that a word with a strong fit to the local context would elicit a smaller N400 response, compared to a word with a poor contextual fit (Kutas & Federmeier 2000; Rugg & Doyle 1994).

In sum, the consensus is that the N400 reflects the ease of integrating new semantic information into the current context. The majority of work demonstrating this has used experimental materials containing clear semantic violations (e.g. spreading bread with *socks*). The critical words in these experiments clearly violate contextual constrains within the sentence, and may be described as a poor fit to the context. However, many of the classic semantic illusions, (e.g. *Moses* and *Survivors*) are difficult at least in part just because they are NOT poor fits to the overall context. In fact these semantic anomalies have a good global fit to the context. So, for example, Moses has a generally good fit in a statement concerning an old-testament biblical story. Likewise, the word

survivors is a word generally expected in a disaster-type scenario. Because these words have a high contextual relevance they are more likely to be processed shallowly, according to Sanford & Garrod's (1998) global-fit theory, and are therefore more likely to go undetected as being anomalous.

While the N400 amplitude appears to reflect the ease with which a word can be integrated into the overall context, the materials commonly used to elicit N400 effects are words with obvious contextual violations at all levels, including the global level. It is open to investigation, therefore, whether or not hard-to-detect semantic anomalies will elicit a similar N400 effect. In fact, it is possible that hard-to-detect semantic anomalies will evoke a P600 waveform and not an N400, as we shall argue later. Such an ERP pattern has been recently reported for a range of materials, including semantic reversal anomalies (Kolk et al 2003); thematic role violations (Kuperberg et al 2003); and animacy violations (Nieuwland & Van Berkum 2005). Such findings suggest that with hard-to-detect semantic anomalies we may find no evidence of an N400, but may in fact find a P600 effect instead.

P600s elicited by semantic anomalies

The P600 is a large positive waveform with a centroparietal distribution, normally found in a 500-800ms time window, peaking at approximately 600ms post-stimulus onset. It is generally considered to be a syntax-relevant component because it has been elicited by a number of different syntactic violations, for example, subject-verb agreement (Hagoort, Brown & Grootheson 1993), verb inflections (Frederici, Pfeifer, & Hahne 1993; Gunter, Stowe & Mulder 1997), case inflections (Neville, Nicol, Barss, Forster & Garrett 1991), incorrect pronoun inflections (Coulson, King, & Kutas 1998), violations of phrase structure (Hahne & Frederici 1999; Frederici et al 1993; Neville et

al 1991), and it is also observed in non-canonical sentences, such as garden path sentences (Osterhout & Mobley 1995), and grammatically complex sentences (Kaan, Harris, Gibson, & Holcomb 2000). However, there is now increasing evidence that the P600 does not only reflect syntactic processing; but it may also be elicited by semantic violations (Hoeks, Stowe & Doedens, 2004; Kim & Osterhout, 2005; Kolk & Chwilla, 2007; Kolk, Chwilla, van Herten, & Oor, 2003; Kuperberg, 2007; Kuperberg, Sitnikova, Caplan & Holcomb, 2003; Niewland & Van Berkum, 2005; see Kuperberg, 2007, for a review). For example, Kuperberg, et al (2003) reported P600s with simple unambiguous sentences where a critical verb was used in a semantically inappropriate way with a preceding inanimate noun phrase. Thus, in [3a], the final verb was semantically incongruous, since eggs do not eat (a violation of the animacy selection restriction for *eat*). In contrast, in [3b] there is no violation of the animacy rule, as in [3b].

[3a] Every morning at breakfast the eggs would eat …

[3b] Every morning at breakfast the boys would plant …

In [3a] eggs are inanimate and so cannot be doing the eating and sentences such as these elicited a strong P600 effect but no N400 effect. In contrast when the final verb was a semantically incongruous word and no animacy violation [3b] there was a strong N400 effect, but no P600, effect observed. Note that the verb *eat* has a good global fit to the breakfast "situation", while *plant* does not have a good fit (this was true of all their materials). This is consistent with the idea that the N400 might reflect global contextual fit, which is good in [3a] but poor in [3b].

Our borderline-detect materials could be considered similar to Kuperberg's in that there is a strong fit between the anomalous word and the global situational context. However,

while Kuperberg's materials present a gross violation in respect of animacy, and so the anomalies are probably always easily detected (though this is just conjecture – no test of detectability was employed by Kuperberg). Our materials have a more complex mismatch with thematic roles (as in "sentencing *victims*"), and so are harder to detect. Therefore, it is possible that an N400 will not be found in our data.

P600 effects have also been demonstrated in response to sentences describing implausible events. For instance, Kolk, Cwilla, van Herten, & Oor (2003) reported a centroparietal P600 with semantic reversal anomalies. These anomalies are syntactically correct unambiguous sentences, which describe implausible events, such as "The cat that from the mice *fled*" (this is semantically anomalous because mice are more likely to run away from cats). Van Herten, Kolk & Chwilla (2005) replicated these results with sentences such as, "the fox that hunted the poacher", and controlled the grammatical number for agents and themes (hunters and poachers) and again observed no N400 but a pronounced P600 to critical verbs (in Dutch sentences the agent and themes precede the critical verb).

Using a somewhat different approach, Nieuwland & Van Berkum (2005) illustrated the role of context in influencing the ERP concomitants of semantic processing. They reported a P600 effect and no N400 to animacy violations which were embedded within a larger discourse. They asked participants to listen to short stories that contained two characters and a scenario-relevant inanimate object (e.g. a male and a female tourist and a suitcase). At one point in the story one animate character was replaced by the inanimate object (e.g. the woman carried on a conversation with the suitcase). On encountering the anomalous word *suitcase,* a large positive deflection, beginning in the 500-600ms time range, with a peak latency within 900-1100ms, and with a centro-parietal distribution, was elicited. There was no evidence of an N400. They argued that

the animacy violation was not detected immediately (because there was no N400), and that this was evidence for a temporary semantic illusion. However, Kuperberg (2007) argues for a more conservative interpretation because participants did detect the anomalies, and that the P600 may reflect a 'processing cost' associated with anomalies, rather than a "temporary neural semantic illusion".

Similarly, Nieuwland & Van Berkum (2006) demonstrated that the power of the discourse can over-ride animacy violations. They used cartoon-like stories where inanimate objects were the central agents of the story, for example "peanuts falling in love". They hypothesised that as the cartoon-like context was established, any initial processing problems associated with animacy violations would be eliminated. They demonstrated that the initial animacy violation elicited an N400, however this effect was attenuated as the story unfolded. They argued that local semantics were overruled by a strong discourse context. The power of the discourse context in over-ruling local semantic anomalies could also be quickly overturned. In stories establishing inanimate objects as central characters (e.g. peanuts falling in love), if a phrase was introduced later in the discourse that contradicted the stories overall cartoon-like context (but was appropriate in respect of real-world expectations, such as peanuts being "salted"), an N400 effect was elicited.

Kuperberg (2007) reviewed and summarised the literature on P600 effects with semantically anomalous materials. She concluded that a P600 will be evoked when certain noun phrase and verb selection restriction constraints are violated, for example with animacy violations. Other situations leading to the occurrence of a P600 are; (a) when there is a strong semantic association between a verb and its arguments (which may lead to a temporary neural semantic illusion); (b) task manipulations such as where participants are asked to make acceptability judgements; and, (c), the presence of a

biasing context that is powerful enough to override anomalous local semantics (similar to the argument we are developing here). Kuperberg concludes that the P600 reflects the repair processes that arise when an inconsistency is detected between semantic and syntactic sentence processes.

There is strong evidence, therefore, to suggest that hard-to-detect semantic anomalies may not elicit an N400, but may instead evoke a P600. However, this may only be apparent in situations where the critical word is **consciously detected** by participants. In the semantic anomaly studies discussed above, detection appears to be either assumed, or the materials can safely be assumed to normally permit detection, even if this detection is slightly delayed (as in Nieuwland & van Berkum's 2005 *temporary* semantic illusion). With the hard-to-detect semantic anomalies examined in this thesis, eventual detection may not occur at all. In the eye-tracking studies reported previously, post-test debriefing sessions confirmed that participants had not been aware of the presence of anomalies when they were not reported (see also similar observations by Barton & Sanford 1993). What, therefore, can we expect to observe in situations where anomalies are not reported by participants? Based on the analyses presented in the previous chapters using eye-tracking measures, the simple prediction is that there will be no differences in non-detect and non-anomalous conditions. However, this lack of differences may be the result of experimental insensitivity rather than there being no effect present (although there is some evidence that eye-tracking is sensitive enough to do this, see below). Alternatively, detection may occur on some level of processing, although this does not reach conscious awareness. In support of this hypothesis there is evidence from a diverse range of experimental paradigms, including eye-tracking, visual change detection, and attentional blink experiments.

Detection without conscious awareness

Daneman, Rheingold & Davidson (1995) reported intriguing effects for homophone error detection in an eye-tracking paradigm. In their study participants were required to proofread a short story containing a number of homophone and orthographically matched non-homophone errors (e.g. *hair* spelt as *hare*, and *bored* as *board*). The eye movement data suggested that homophone and non-homophone errors were equally disruptive, compared to the correct target words, and that there was no difference in the initial processing time for homophone and non-homophone errors. However, the behavioural data (in this case a button response to indicate that an inconsistent word had been detected) revealed significant effects with homophone detection substantially lower than non-homophone detection rates. Just as Daneman et al. found no behavioural effect, while recording disruption in the eye movement data, we suggest that our participants who do not detect (behaviour) an anomaly and who show no effect of detection on the eye movement records, may nevertheless show a different ERP for undetected anomalies.

Secondly, visual change detection studies have also provided evidence that changes may be unconsciously processed even in the absence of overt detection. Change detection studies have repeatedly demonstrated how difficult it can be to detect changes in the visual world. This effect is referred to as change blindness, and may be defined as the "failure to become explicitly aware that a change is or was taking place" (Thornton & Fernandez-Duque (2002), p.100). Change detection studies present two versions of a visual stimulus (separated by a short inter-stimulus mask), and participants are required to report whether or not there is a change in some aspect of the scene in the second stimulus. Some researchers have modified this procedure whereby the two stimuli are repeatedly alternated, termed a 'flicker paradigm', so that the temporal

properties of changes can be investigated as well. This technique has permitted researchers to assess the influence that aspects of the visual scene, for example stimulus features, or psychological attributes, such as attention, have in modulating rates of change detection (see Simons, 2000, for a review).

However, there have also been a number of reports of participants failing to report changes, while the data (including detection rates, reaction time, and eye movement measures) suggests that some effect of change has in fact occurred in the processing system. For example, Fernandez-Duque & Thornton (2000) used a modified visual change detection task whereby two objects were presented, one of which had a feature changed across presentations. The participant's task was to identify the object which was changed. In cases where participants were not certain of which object had changed they were asked to guess. Performance was above chance level in situations where they were asked to guess, which suggests implicit change detection. Williams & Simon (2000) measured how long it took participants to decide that no change had been made to a complex object. They reported that it took longer for participants to incorrectly decide that no change had occurred (i.e. failed to detect a change) compared to a no-change comparison. Again, this suggests that the un-detected change is influencing visual perception. Furthermore, eye movement studies have demonstrated that unreported object changes resulted in longer re-fixation, compared to no-change conditions (Hollingworth, Schrock, & Henderson 2001; Hollingworth, Williams, & Henderson 2001). Finally, in an ERP flicker change detection paradigm Fernandez-Duque, Grossi, Thornton, & Neville (2002) identified what they termed an 'implicit marker' for change detection in cases where participants failed to detect changes, compared to no change control situations. In their visual change detection experiment, the participant's task was to report a change in a flicker presentation of a complex scene (500ms alternate presentations of a picture, separated by a 300ms blank screen, with

flickers occurring up to a maximum of 40 times). Their attention was directed at either a central fixation spot or towards the site of a previously detected change. At the same time they also monitored for a second change in the scene. They observed that in cases where a change occurred but went unreported, that there was a bilaterally distributed deflection over anterior sites within a 240-300ms time window, when compared to a no-change situation when participants were actively searching for a change. This neural activity, they suggested, may have reflected implicit change detection without explicit awareness.

Thirdly, evidence for unconscious semantic processing was reported by Vogel, Luck & Shapiro (1998). In a rapid serial visual presentation task (RSVP) they reported an N400 effect to semantically incongruous words presented within a time window termed the 'attentional blink'. This refers to a short time period after an initial stimulus has been perceived during which subsequent processing is suppressed. In an RSVP task an initial context word may be consciously identified by participants, however a suitable period of time must elapse before participants will reliably identify a subsequent presented word (the term attentional blink is used analogously to eyeblink). Vogel et al employed semantically related or incongruous word pairs (e.g. doctor – nurse, vs doctor – chicken) presented amongst random consonant strings in an RSVP paradigm. They observed that semantically incongruous words presented within the 'attentional blink' period (and hence not consciously reportable by participants) elicited an N400 effect. This, they argued, provides evidence for unconscious semantic processing. Similar effects have also been reported by Sergent, Baillet, & Dehaene (2005).

In Experiment 6, we investigated the difference in ERP waveforms between easy-to-detect (globally incoherent) anomalies and hard-to-detect (globally relevant) semantic anomalies, and compared detected and non-detected anomalies in the hard-to-detect set.

It was predicted that easy-to-detect anomalies would elicit a classic N400 waveform, whereas hard-to-detect anomalies were expected to evoke no N400, but rather a late positivity. Since participants were requested to immediately report detected anomalies, a comparison was possible between overt detection and non-detection. In cases where anomalies have not been detected, compared to non-anomalous controls there may be no differences observed in the data, which would support the previously reported eye tracking studies. Alternatively, the evidence may support the interpretation that implicit detection has occurred in the absence of conscious detection (as suggested by homophone error detection, change detection, and attentional blink studies).

Experiment 6

Method

Participants

27 participants took part in the study which was carried out in a single session lasting approximately two and half hours. All were right-handed and were native English speakers with no diagnosed reading disorders. All had normal or corrected-to-normal vision and were paid £15 for participation.

Materials

For an ERP study, many more items are required per cell of the design than is the case for eye-tracking. As a rule-of-thumb based on the experience of many other ERP investigations measuring the N400, an ideal figure is 40 readings per design cell (see Van Berkum et al. 2004) (as compared to 6-10 per cell for eyetracking). For this reason, many more anomaly materials were created. Experimental items were adapted

from various sources. The majority had been developed by ourselves and had been extensively pre-tested. Additional items were modified from published anomaly research (Reder & Kusbit 1991; Bredart & Modolo 1988). Necessarily, many of these relied on general knowledge rather than on purely semantic information.

The items were written so that each would have an introductory sentence that established the context. The second sentence was 17 words long and the critical word was always the 13th word position[7]. Whether or not the critical word was anomalous or not was achieved by manipulating a prior context word or phrase within the critical sentence. The context manipulation and target words were separated by 5 words. *For example:*

First sentence (presented whole)	A pay dispute between lorry drivers and their employer reached a crisis in negotiation, even the professional mediators seemed dejected.
Context either anomalous or not **target item**	After five days of discussion the *Government - union* rejected outright the final conciliatory **pay-offer** and halted the talks.

It would be anomalous for the government to reject the pay-offer (instead they would be making the pay-offer) and so readers who detected this item would report, "pay-offer" as anomalous. When the context was changed to union the target word pay-offer is appropriate in the context and would not be identified as anomalous. The context was manipulated by changing one word where possible, but this was impossible for some of the items where more words were changed. However anomalous and non-anomalous

[7] We were confident that this would not result in participants predicting the anomaly based on the debriefing sessions from the prior eye-tracking studies. These studies also placed the anomaly in the same position and participants were explicitly asked as part of the debriefing session whether they were predicting where the anomaly would appear. Neither in those studies, nor in the present ERP study, did any participant report that they were aware of where the anomalies appeared.

versions always contained the same number of words and the number of words between the context manipulation and the critical word was always five.

A total of 135 experimental items were used. Three files were constructed, each containing 90 of the materials in the anomalous condition and 45 in the non-anomalous condition. On the hopeful assumption that 50% of the anomalies would be detected, this would place 1/3 of the observed data into each of the three experimental categories (detect, non-detect, non-anomalous), thus optimizing statistical comparisons. By rotation, over all three files, each material would occur in each of the conditions.

To these materials, 40 fillers were added with obvious anomalies, and 45 non-anomalous controls, in each case consisting of 2 sentences. The anomalous word in these semantic incongruent fillers were placed at various points throughout the second sentence, but never in the same place as the experimental items (i.e. 17th word). An example semantically incongruent filler is:

> Jenny decided to spring clean her house. She washed her floors with a bucket of **mud** and the old floorboards came up sparkling clean.

This made a total of 220 stories in each file. Thus, overall, 59% of the materials contained an anomaly of some sort, and 41% did not.

Procedure

Participants were seated in a testing booth 80cm away from a computer monitor. A chin rest was used to minimise head movements. A 3-button response box was used to control text presentation and make responses. A small microphone and camera permitted communication between participant and experimenter. Participants were

asked to read the stories for normal comprehension. When they were ready to begin the participants pressed a 'proceed' button on the box (the middle button) and the complete first sentence was presented on the screen. When participants had read the sentence they pressed the same button to proceed to the second sentence. The second sentence was presented one word at a time.

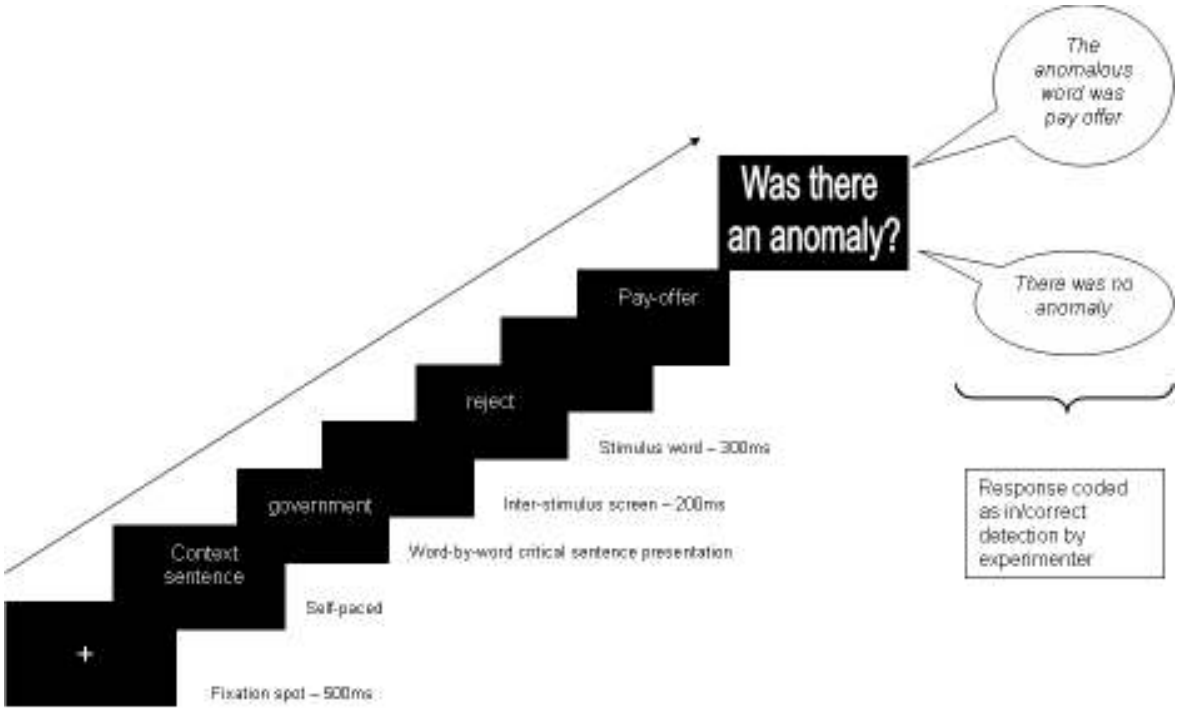


Figure 7.1: Presentation order of screens as seen by participants

This began with a fixation point presented for 500 ms. Words were presented for a standard 300 ms with an inter-stimulus time period of a standard 200 ms (see Hald et al. 2006; Van Berkum et al. 1999; Van Herten et al. 2005 for similar parameters). All text was presented in a white font on a black background. Two other response keys were used, one marked 'OK' and the other 'No Sense' (these were swapped around, so for half the participants OK was the left button and for the remainder it was the right button). Participants were asked to press the 'No Sense' button as soon as they detected an anomaly. If there was no anomaly they pressed the 'OK' button at the end of the

sentence. A screen prompt then asked them to verbally report their response to the experimenter. They did this for all trials. If there was an anomaly detected they were also required to identify and explain it. The experimenter recorded what the participant said and whether or not they were correct. There were ten blocks of 22 stories. They also completed an initial practice block of 8 stories. After the ERP recording a short multiple choice questionnaire (MCQ) was delivered that checked participants understood the anomalies. If participants did not understand any anomaly, the item was not included in the analysis. However, this did not happen: all responded that they understood each item.

EEG recording

A BIOSEMI Active-Two amplifier system was used for continuous recording of electroencephalographic (EEG) activity from 72 Ag/AgCl electrodes over midline electrodes Fpz, AFz, Fz, FCz, Cz, CPz, Pz, POz, Oz, and Iz, over the left hemisphere from electrodes IO1, Fp1, AF3, AF7, F1, F3, F5, F7, F9, FC1, FC3, FC5, FT7, C1, C3, C5, M1, T7, CP1, CP3, CP5, TP7, P1, P3, P5, P7, PO3, PO7, O1, two nonstandard positions PO9' and O9' which were located at 33% and 66% of the M1-Iz distance, respectively, and from the homologue electrodes over the right hemisphere. EEG and EOG recordings were sampled at 256 Hz. The online reference electrode was the Biosemi Common Mode Sense (CMS) electrode (see http://www.biosemi.com/faq/cms&drl.htm for details). Off-line, all EEG channels were recalculated to a linked mastoid reference for direct comparison with the studies in the literature. Trials containing blinks were corrected using a dipole approach (BESA Version 5.1.6). Automatic artifact detection software (BESA) was run and trials with non-ocular artifacts (drifts, channel blockings, EEG activity exceeding ± 120 μV) were

automatically discarded. The epoch started 200 ms prior to the onset of the critical word and lasted for a total duration of 1,200 ms.

Data Analysis.

For the artifact-free trials, the signal at each electrode site was averaged separately for each experimental condition, time-locked to the onset of the critical word. Before the measurement of ERP parameters EEG and EOG activity was band-pass filtered (0.1-25 Hz, 6 dB/oct). The ERP waveforms were aligned to a 100-ms baseline immediately prior to the onset of the critical word. Mean ERP amplitudes were measured in typical time intervals for P1 (100-130ms), N1 (150-190 ms), P2 (200-300 ms), N400 (300-500 ms) and P600 (500-900 ms). Early effects are not commonly reported in language comprehension ERP literature, however P1 effects have been reported with effects of word length (Hauk & Pulvermüller, 2004), and the N1 and P2 with lexical processing (Hauk & Pulvermüller, 2004; Sereno, Posner & Rayner, 1998; for a review see Sereno & Rayner 2003).

ERP amplitudes at midline electrodes (Fz, FCz, Cz, CPz, Pz) were analysed separately from data recorded over lateral electrode sites. Lateral electrode sites were pooled to form regions of interest (ROIs) as recommended for the analysis of high-density electrode arrangements (Dien & Santuzzi, 2005). The electrodes were divided along a left-right dimension, an anterior-to-posterior dimension, and a dorsal-ventral dimension. The six ROIs over the left hemisphere were: left-anterior-ventral (AF7, F7, FT7, F5, FC5), left-anterior-dorsal (AF3, F3, FC3, F1, FC1), left-central-dorsal (C3, CP3, C1, CP1), left-central-ventral (TP7, T7, C5, CP5), left-posterior-ventral (PO9', O9', P7, PO7, O1), and left-posterior-dorsal (P3, PO3, P1, P5); six homologous ROIs were defined for the right hemisphere (see figure 7.2).

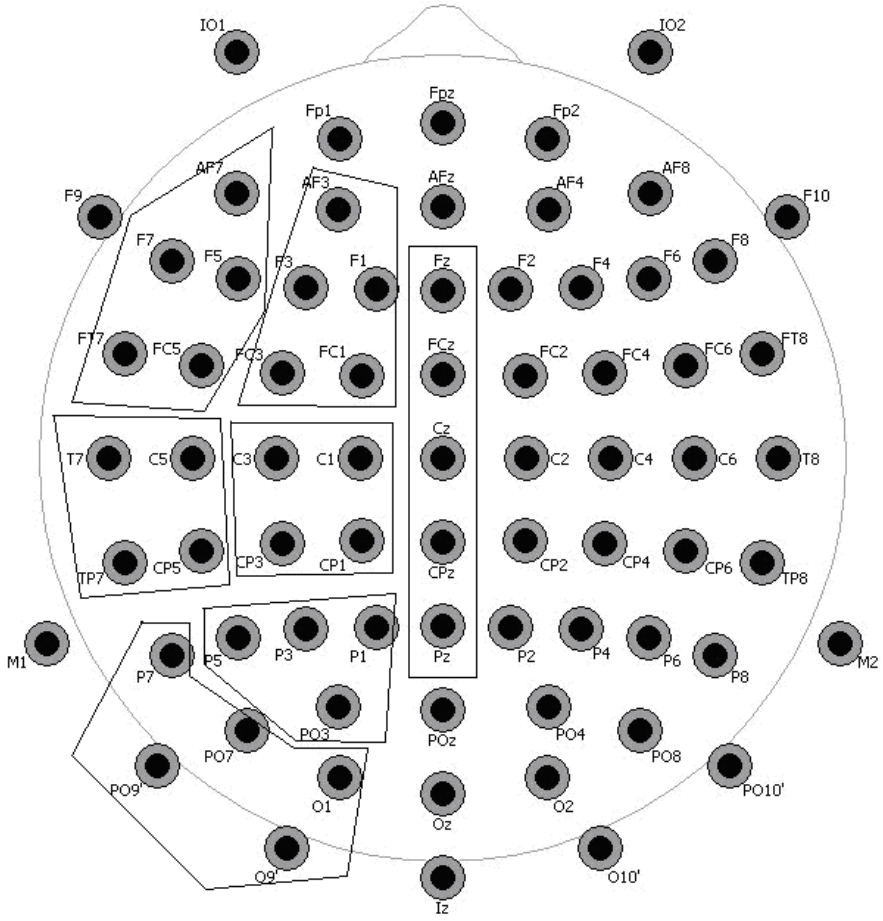


Figure 7.2: Arrangement of electrodes with lateral regions of interest (ROI) and midline electrodes included in statistical analyses (only the left hemisphere lateral ROI are illustrated however there were homologous right hemisphere sites as well). There were six lateral ROIs: left-anterior-ventral, left-anterior-dorsal, left-central-ventral, left-central-dorsal, left-posterior-ventral, left-posterior-dorsal.

Results

Detection rates

On average, participants correctly detected semantic anomalies at a rate of 73%. The average rate of detection for the easy-to-detect fillers was 98%.

Statistical analyses.

Statistical analyses were performed by means of repeated measures analyses of variance (ANOVA)[8]. For the analysis of ERP amplitude data recorded from midline electrodes, an ANOVA was carried out with the variables **condition** (non-anomalous vs. incongruent; and non-anomalous vs. anomalous detect vs. non-detect, respectively), and **electrode** (Fz, FCz, Cz, CPz, Pz). For the analysis of ERP deflections maximal over lateral electrode sites, an ANOVA was performed with variables **condition**, **hemisphere** (left, right), **anterior-posterior** (anterior, central, posterior), and **verticality** (ventral, dorsal).

The data were analysed in five separate comparisons. The initial comparison investigated whether a classic N400 effect was evident in the data when readers were presented with 'obvious' semantically incongruent words. The data from the control non-anomalous condition was used as a comparison to investigate this further, hence, the variable condition has the two levels non-anomalous and incongruent word. The remaining analyses used the data from the hard-to-detect experimental items. ERP recordings from the semantically anomalous condition were separated into detect and non-detect (these were coded by the experimenter as detect or non-detect based upon participants' verbal responses). The procedure was exactly the same as that used in the previously reported eye-tracking studies. An initial omnibus ANOVA (3 levels condition variable) compared anomalous detect, non-detect and non-anomalous conditions in the time-periods of interest. Planned additional analyses compared anomalous detect to non-detect, anomalous detect to non-anomalous, and anomalous non-detect to non-anomalous controls.

---

[8] The Huyhn-Feldt correction was applied. This is less conservative than the Greenhouse-Geiser correction, and so reduces the chances of making a type 2 error,

Is there an N400 for easy-to-detect anomalies?

This analysis investigated whether easy-to-detect anomalies would produce an N400 effect within the current task. Data from filler items that were easily detectable as semantically incongruous (referred to as the easy to detect condition) was compared to that from the control non-anomalous condition. All results are summarized in table 7.1, and summary statistics for all conditions in table 7.2. Grand average waveforms are presented in figure 7.3b and difference waveforms in 7.3c.

100-300ms In both midline and lateral recording sites the ANOVA for the N400 condition there were no main effects or significant interactions within the time windows of 100-130ms, 150-190ms, or 200-300ms time windows, with the exception of a Condition x Hemisphere x Ant-Pos interaction, $F(2, 50) = 4.3$, $p < 0.03$, which indicated a tentatively larger P1 asymmetry over posterior sites for the incongruent than non-anomalous condition (see discussion).

The N400 window: 300-500ms In the 300-500ms time window the midline analysis revealed a significant main effect of condition ($F(1,25)= 6.1$ $p<0.02$), indicating a 2 μV more negative-going ERP in the incongruent compare to the non-anomalous condition (2.6 vs. 4.8 μV), consistent with an N400 effect. The analysis of ERP amplitudes at lateral ROIs confirmed the main effect of condition ($F(1,25)= 9.5$ $p<0.005$), which was more pronounced over frontocentral than posterior ROIs, $F(2, 50) = 4.97$, $p < 0.01$, and this interaction was further modulated by verticality (condition x anterior / posterior x dorsal / ventral), $F(2, 40) = 5.3$, $p < 0.008$. Figures 7.3a presents the topographic figure for this analysis and it can be seen that there is a broadly distributed negative potential within the time epoch consistent with an N400 effect.

Table 7.1: ANOVA analyses for all midline and lateral comparisons comparing semantically incongruous and non-anomalous conditions per epoch

Midline recordings	Df	100-130	150-190	200–300	300-500	500-900
Condition	1,25	ns	ns	ns	F=6.1 p<0.02	F=5.8 p<0.02
Electrode	4,100	F=8.8 p<0.0002	ns	ns	ns	F=25.2 p<0.0001
Condition * electrode	4,100	ns	ns	ns	ns	F=12.1 p<0.0001
Lateral recordings	Df	100-130	150-190	200–300	300-500	500-900
Condition	1,25	ns	ns	ns	F=9.5 p<0.005	F=5.34 p<0.03
Hemisphere	1,25	F= 10.4 p<0.004	ns	F=8.9 p<0.006	ns	F=5.94 p<0.02
Anterior / Posterior	2,50	F=47.1 p<0.0001	F=16.4 p<0.0003	F=28.0 p<0.0001	ns	F=26.9 p<0.0001
Dorsal / Ventral	1,25	F=9.6 p<0.005	F=84.2 p<0.0001	F=124.0 p<0.0001	F= 66.8 P<0.0001	F=125.6 p<0.0001
Cond * hemi	-	ns	ns	ns	ns	ns
Cond * Antpos	2,50	ns	ns	ns	F=5.0 p<0.02	F=28.8 p<0.0001
Hemi * antpos	2,50	F= 11.6 p<0.0001	F= 71.8 p<0.003	F=4.3 p<0.03	ns	ns
Cond * dove	1,25	ns	ns	ns	ns	F=14.8 p<0.0007
Hemi * dove	-	ns	ns	ns	ns	ns
Antopos * dove	2,50	F=33.6 p<0.0001	F=28.7 p<0.0001	F=51.3 p<0.0001	ns	F=10.8 p<0.0003
Cond * hemi * antpos	2,50	F=4.3 p<0.03	ns	ns	ns	ns
Cond * hemi * dove	-	ns	ns	ns	ns	ns
Cond * antpos * dove	2,50	ns	ns	ns	F=5.3 p<0.008	ns
Hemi * antpos * dove	2,50	F=5.7 p<0.006	ns	F=6.6 p<0.005	F=4.1 p<0.03	F=5.2 p<0.02
Cond * hemi * antpos * dove	-	ns	ns	ns	ns	ns

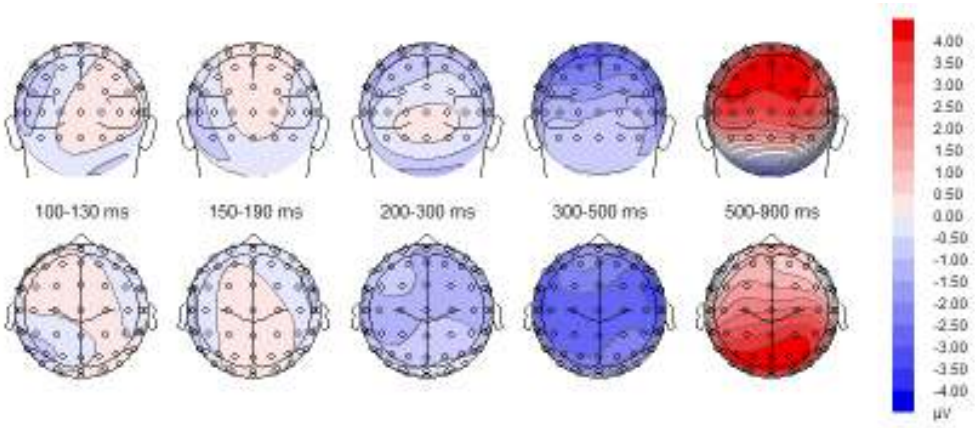
Cond = condition; Hemi = hemisphere; Antpos = anterior – posterior; Dove = dorsal – ventral

The P600 window: 500-900ms In the later 500-900ms time window, for midline sites there was a main effect of condition ($F(1,25)= 5.8$ $p<0.02$), with a larger positivity for incongruent than for non-anomalous sentences (9.5 vs. 6.7 μV; see table 7.2). The significant Condition x Electrode interaction, $F(4, 100) = 12.1$, $p < 0.0001$, indicated this effect to be strongest over centroparietal electrodes (Figure 7.2). Again, the analysis of mean ERP amplitudes at lateral sites corroborated these findings by revealing a significant main effect of condition ($F(1,25)=5.3$ $p<0.03$, and a Condition x Ant-Pos interaction, $F(4, 100) = 28.8$, $p < 0.0001$. The condition effect was stronger over dorsal than ventral lateral sites as indicated by the Condition x Verticality interaction, $F(1, 25) = 14.8$, $p < 0.0007$, and over right centroparietal ROIs as indicated by the Hemisphere x Ant-Pos x Verticality interaction, $F(2,50) = 5.2$, $p < 0.02$. Figure 7.3a illustrates that semantically incongruent words compared to the non-anomalous condition elicited a more positive-going waveform mostly over posterior sites.

So, the easy-to-detect anomalies show clear evidence of an N400, and also evidence for a later positivity. This will be discussed later.

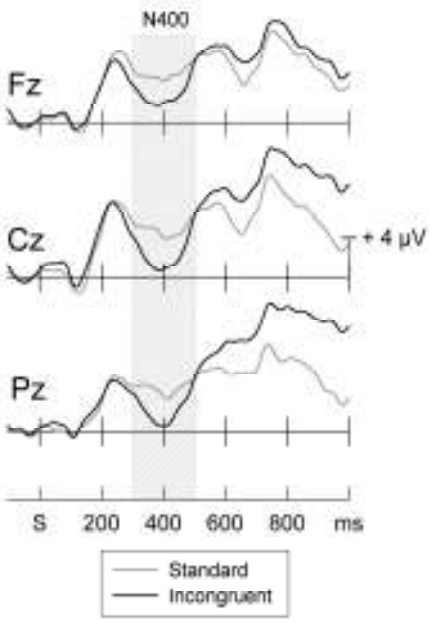
Table 7.2: Summary of mean amplitude (standard deviation) per condition in midline and lateral recording sites

Midline	N	Non-anomalous	Anomaly detect	Anomaly Non- detect	N400
100 -130	130	0.9 (2.5)	0.9 (1.9)	0.9 (2.9)	-0.5 (2.7)
150 – 190	130	2.4 (2.8)	2.6 (3.1)	2.2 (3.1)	2.9 (3.6)
200 – 300	130	6.0 (2.9)	6.2 (3.5)	4.5 (3.7)	5.3 (4.8)
300 – 500	130	4.8 (3.0)	4.9 (4.4)	3.2 (4.1)	2.6 (5.5)
500 - 900	130	6.7 (3.2)	8.3 (5.3)	5.0 (5.6)	9.5 (7.0)
Lateral					
100 -130	312	0.4 (2.5)	0.5 (2.4)	0.6 (2.8)	0.5 (2.6)
150 – 190	312	0.8 (2.9)	0.7 (2.9)	0.8 (3.1)	0.8 (3.3)
200 – 300	312	3.7 (3.1)	3.8 (3.3)	2.9 (3.7)	3.1 (4.0)
300 – 500	312	3.3 (2.7)	3.3 (3.6)	2.4 (3.7)	1.3 (4.4)
500 - 900	312	5.1 (3.1)	6.2 (4.7)	3.9 (5.0)	7.2 (6.4)



7.3a

Figure 7.3a illustrates topographic maps for semantically incongruent nouns minus control non-anomalous conditions per each time epoch relative to critical word onset. The negative polarity is demonstrated in the 300-500ms epoch and can be seen to be broadly distributed. This is followed by a positivity in the 500-900ms window which appears to be concentrated over posterior sites



7.3b

Figure 7.3b illustrates the grand average waveforms for semantically incongruent and standard conditions for three representative electrodes (Fz, Cz, Pz). The shaded region represents the 300-50ms time window where the incongruent conditions can be seen diverging from the standard condition.

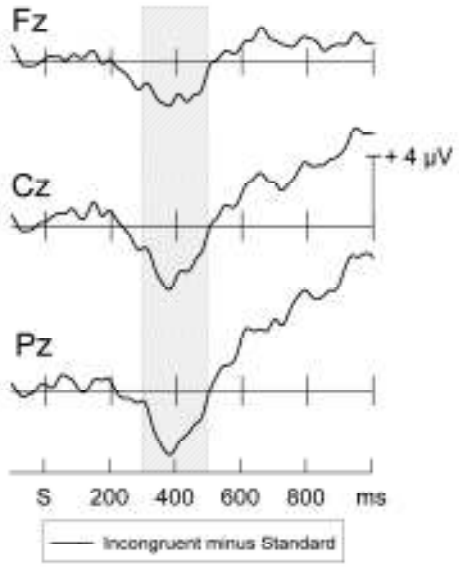


Figure 7.3c illustrates this analysis in the form of a difference waveform which more clearly illustrates the extent of the negative deflection in the 300-400 window, and the following positivity.

Hard-to-detect anomaly analysis: Anomalous detect / non-detect / non-anomalous

The remaining analyses were concerned with the borderline, hard –to-detect anomalies, and the pattern of analyses follows the pattern established in the eye-tracking work. The omnibus analysis compared detected anomalies, undetected anomalies, and non-anomalous control materials. This analysis indicates whether there are effects amongst the three conditions. Table 7.3 summarises all ANOVA results for the omnibus analysis. Post hoc paired comparisons are displayed in tables 7.4 (anomaly detect compared to anomaly non-detect), 7.5 (anomaly detect compared to non-anomalous), and 7.6 (non-detected anomalies compared to non-anomalous). These tables are presented at the end of the chapter to avoid too much disruption to reading. Figure 7.4 illustrates the grand average waveforms for these three conditions.

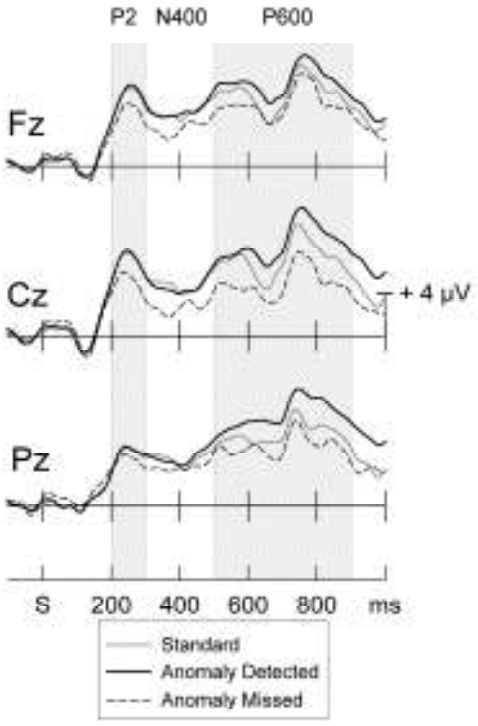


Figure 7.4: Grand average waveforms for three representative electrodes from midline sites (Fz, Cz, Pz), illustrating non-anomalous (standard), anomalous detected and anomalous missed conditions.

100-130ms The analysis of mean ERP amplitudes (100-130 ms) for lateral ROIs indicated an initial positivity (P1) that was larger over right than left hemisphere, and dorsal compared to ventral, occipito-parietal electrodes, $F(2, 50) = 5.0$, $p < 0.01$. There was no reliable effect for conditions as a main effect, $F < 1$, or in interaction with topographic factors, all Fs < 1.5, ps > .24, with the exception of the Condition x Ant-Pos x Verticality interaction, $F(4, 100) = 3.8$, $p < 0.009$. However further post-hoc comparisons revealed no significant condition effects over posterior-lateral sites where P1 was maximal.

150-190ms The following negativity (N1), peaking at about 170 ms, was most pronounced over left parieto-occipital electrodes as indicated by the Ant-Pos x Hemisphere interaction, $F(2, 50) = 5.5$, $p < 0.02$, and over posterior ventral compared to dorsal ROIs, (2, 50) = 31.2, $p < 0.0001$. Whilst the main effect for condition was again not significant, there was a reliable Condition x Ant-Pos x Verticality interaction, $F(4, 100) = 2.9$, $p < 0.04$. However, there were no significant condition effects over posterior electrodes where N1 was maximal.

200-300ms As can be seen in figure 7.5, the ERP was more positive-going for anomaly detect than anomaly non-detect conditions for the time intervals from 200 to 900 ms. It can be seen in the topographic map (figure 7.6) that this effect is broadly distributed, but appears to be more pronounced over central and parietal midline sites (see difference waveforms illustrated in figure 7.5). This impression was corroborated by the ANOVA of mean ERP amplitudes, firstly, during the 200-300 ms interval over midline electrodes, which revealed a significant main effect of condition, $F(2,50) = 4.5$, $p < 0.02$. Post-hoc comparisons revealed no amplitude difference between non-anomalous and anomaly-detect conditions (6.0 vs. 6.2 μV), $F < 1$. ERP amplitude was significantly less positive, however, in the anomaly-missed condition (4.5 μV) than in

the anomaly detected condition (6.2μV), $F(1, 25)=8.9$, $p < 0.006$, and approached significance when compared to the

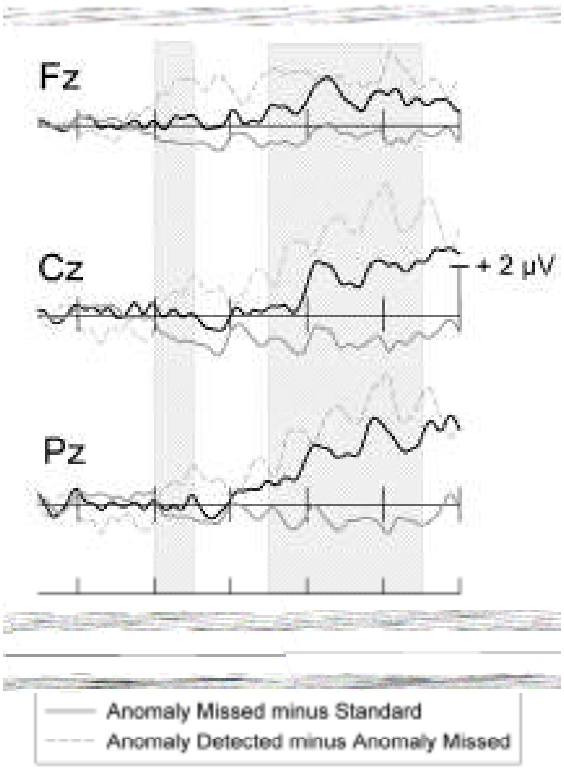


Figure 7.5: Difference waveforms for three representative electrodes (Fz, Cz, Pz) for anomalous detected minus non-anomalous (standard), anomalous missed minus standard, and anomalous detected minus anomalous missed.

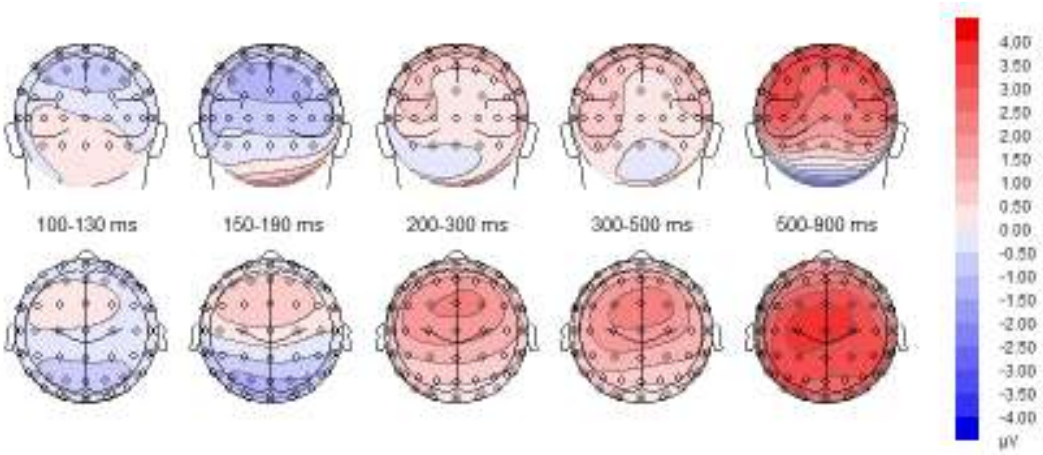


Figure 7.6 Topographic maps of ERP difference waverforms for each time window of interest relative to onset of critical word in detected anomaly minus missed anomaly conditions

non-anomalous condition $F(1, 25)=4.1$, $p < 0.055$. The analysis of ERP amplitudes over lateral ROIs revealed a Condition x Verticality interaction, $F(2,50) = 4.3$, , $p<0.02$, due to a more positive ERP over dorsal rather than ventral ROIs for detected than non-detected anomalies, $F(1, 25) = 10.3$, $p < 0.004$. In this early time window there is a reliable divergence in the amplitudes for detected (displaying a positive-going waveform) and missed (more negative-going) anomalies. This will be discussed later.

N400 window: 300-500ms The analysis of mean ERP amplitudes for midline sites during the 300-500 ms interval revealed a similar significant main effect of condition, $F(2,50) = 4.2$, $p < 0.03$. In addition, the Condition x Electrode interaction was also significant, $F(8, 200)= 2.3$, $p < 0.05$, indicating a larger condition effect over frontocentral electrodes**.** Further comparison revealed a reliable difference between the anomaly-detected versus the anomaly-missed condition as indicated by a significant main effect of condition, $F(1, 25)= 11.0$, $p < 0.003$. Furthermore, the comparison between non-detect and non-anomalous conditions once again approached significance, $F(1, 25)= 4.0$, $p <0.056$ (Figure 7.4). However, the Condition x Electrode interaction revealed a more positive ERP waveform over frontocentral electrodes for the non-anomalous versus the anomaly-missed condition, $F(4, 100) = 3.6$, $p < 0.02$. In this time window there is still a reliable difference in the amplitude differences between detected and missed anomalies reported in the previous time period. However, there also appears to be a difference between missed anomalies and non-anomalous controls, with a more negative-going waveform for anomalies that are missed (see figure 7.7 for a topographic illustration, and figure 7.5 for how this difference appears in respect of three midline electrodes).

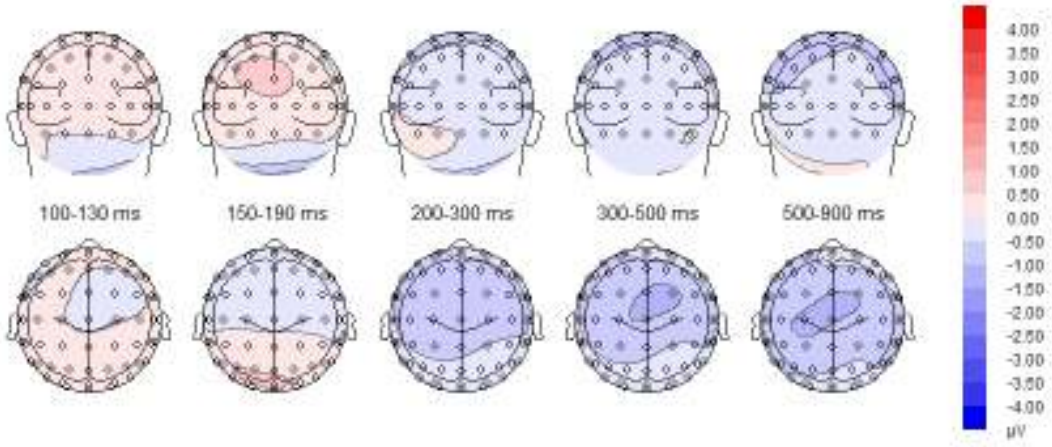


Figure 7.7 Topographic maps of ERP difference waverforms for each time window of interest relative to onset of critical word in missed anomaly minus non-anomalous conditions

It is evident from topographic maps that the condition effect was maximal over midline electrode sites, and that the analysis of ERP amplitudes over lateral ROIs produced weaker condition effects. Nevertheless, there was a significant effect of condition between anomaly detect and non-detect $F(1,25)=4.4$ $p<0.05$ (with a higher mean amplitude for detected than missed anomalies, 3.3 vs. 2.4 µV respectively), and a reliable interaction between Condition x Ant-pos x Verticality $F(2,50)= 3.4$ $p<0.04$. The implications of this will be considered in the discussion.

P600 window: 500-900ms The analysis of mean ERP amplitudes for midline sites during the 500-900 ms interval revealed a similar significant main effect of condition, $F(2,50) = 8.1$, $p < 0.002$. In addition, the Condition x Electrode interaction was also significant, $F(8, 200)= 2.7$, $p < 0.02$, indicating a larger condition effect over central electrodes**.** Further comparison revealed a reliable difference between the anomaly-detected versus the anomaly-missed condition as indicated by significant main effect of condition, $F(1, 25)= 22.1$, $p < 0.0001$, with a greater positivity for anomaly detects compared to non-detects (8.3 vs. 5.0 µV respectively). The Condition x Electrode interaction revealed that this effect was stronger over central electrodes for the non-

anomalous versus the anomaly-missed condition, $F(4, 100) = 2.8$, $p < 0.04$. Finally, it was only the anomaly detect condition that elicited a stronger positivity as compared to the non-anomalous condition (8.3 vs. 6.7 μV), $F(1, 25) = 4.5$, $p < 0.05$. In this time window there is still a reliable difference between anomalies that were detected and missed (see figures 7.6 for a topographic illustration and 7.5 for difference waveforms), however, a reliable difference between detected anomalies compared to the standard non-anomalous condition was also revealed. This demonstrates that anomaly detection results in a late positivity, especially over central and parietal areas (see figures 7.8 for a topographic illustration and 7.5 for difference waveforms).

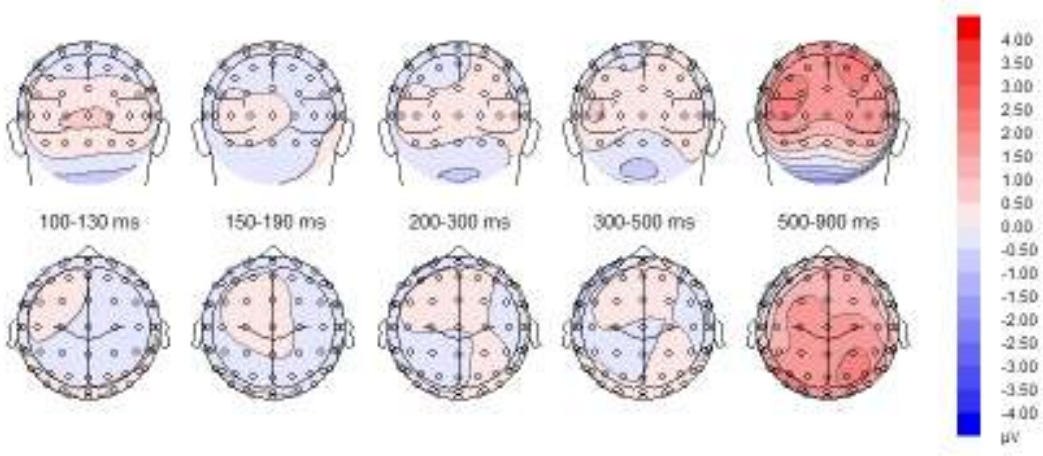


Figure 7.8 Topographic maps of ERP difference waverforms for each time window of interest relative to onset of critical word in detected anomaly minus non-anomalous conditions

In the analysis of ERP amplitudes over lateral ROIs there was a main effect of condition, $F(2,50) = 6.4$, $p < 0.005$, and a Condition x Ant-Pos interaction, $F(4, 100) = 5.7$, $p < 0.0003$, and Condition x Verticality, $F(2,50) = 11.1$, $p < 0.003$ were also significant. As with midline sites, post hoc tests revealed significant effects for condition between anomalous detect and non-detect, F=13.9 p<0.001 (6.2 vs 3.9 μV respectively), and between anomalous detect and non-anomalous F=4.4 p<0.05 (6.2 vs.

5.1μV respectively), but with no reliable difference between non-detect and non-anomalous conditions.

In summary, easy-to-detect, semantically incongruous words elicited a classic N400 response within the current paradigm. This consisted of a significant negativity over central sites within the 300-500ms time-window, which then reversed into a significant positivity over parietal sites within the 500-900ms time epoch (see discussion). The omnibus ANOVA compared hard-to-detect anomalous items that were detected and non-detected, and the control non-anomalous conditions. Main effects for condition in midline sites were found from 200ms onwards. Post hoc comparisons revealed that there were significant differences between anomalous detect and non-detect in all three time windows (200-300ms, 300-500ms, and 500-900ms). However, significant differences between anomalies that were detected and the control condition were confined to the later time window of 500-900ms, consistent with a P600 effect. Similar effects were found in the lateral sites, however, these were less pronounced than in the midline recordings. Finally, anomalies that were undetected by participants, compared to the non-anomalous condition, came close to significance in both time windows between 200-500ms, with a more negative-going waveform for non-detected anomalies.

Discussion

ERP recordings were made while readers were presented with short stories containing semantic anomalies. The participant's task was to report the semantic anomalies. ERP recordings were separated into instances when anomalies were detected and instances when they were not for easy and hard-to-detect semantic anomalies. Data was then averaged across these conditions and compared. There were four main findings following from the analysis. First, easy-to-detect semantic anomalies elicit a classic

N400 result. Secondly, hard-to-detect semantic anomalies evoked a different waveform dependent on whether or not they were detected. Thirdly, detected anomalies elicited a late positivity. Fourthly, non-detected anomalies appeared to elicit a small negative-going waveform, compared to detected and non-anomalous controls.

Easy-to-detect semantic anomalies elicited a clear N400 effect demonstrating that N400 effects could be detected within the current paradigm. This is important to show because it underscores the difference found with easy and hard-to-detect semantic anomalies. Easy-to-detect anomalies are items where there is a clear violation of fit to global as well as local context, and participants have little difficulty in detecting them. So, for example Kutas & Hillyard (1980) used items such as, "he spread the warm bread with *socks*," whereas in the current study we used items such as, "she cleaned the floors with *mud.*" In both examples, the critical words, *socks* and *mud*, violated the contextual constraints set up by the sentences (be they things that are edible or things that are used for cleaning), and in both cases these materials evoked an N400 waveform. The N400 is explained as the response to a word that does not easily integrate into the unfolding sentence or context. These results, therefore, successfully replicate commonly reported findings within the ERP literature with these types of materials. There was also an interaction between condition x hemisphere and verticality within the P1 time window. The P1 is known to reflect sensitivity to word length (Hauk & Pulvermüller 2004), which was a factor not controlled in this comparison, and post-hoc analysis did reveal a significant difference in word length between the easy-to-detect and control word comparisons, however this does not invalidate the later N400 effect. Further analysis of this condition also revealed that the N400 was followed by a late positivity. Such late positivities have been reported in the literature (see Munte, Heinze, Matzke, Wieringa, & Johannes 1998; Osterhout & Mobley 1995). Munte et al suggest that a P600 waveform in response to a semantic violation may be one that has been generally under-

reported in much of the research interested in the N400. They argue that many studies of semantic violations and the N400 (for example, Kutas & Hillyard 1983) confine their analysis to a time window relevant for observing N400 effects, but do not analyse data within a relevant latency range to observe P600 effects.

Hard-to-detect semantic anomalies did not elicit an N400 response but a P600 waveform instead. This effect is similar to those found with other types of semantic violations as discussed earlier (Hoeks, Stowe & Doedens, 2004; Kim & Osterhout, 2005; Kolk & Chwilla, 2007; Kolk, Chwilla, van Herten, & Oor, 2003; Kuperberg, 2007; Kuperberg, Sitnikova, Caplan & Holcomb, 2003; Niewland & Van Berkum, 2005; see Kuperberg, 2007, for a review). These reports have challenged the 'traditional' view of the P600 as an index of syntactic violation (Hagoort, Brown & Grootheson, 1993; Frederici, Pfeifer, & Hahne, 1993; Gunter, Stowe & Mulder, 1997; Neville, Nicol, Barss, Forster & Garrett, 1991; Coulson, King, & Kutas, 1998; Hahne & Frederici, 1999; Osterhout & Mobley, 1995; Kaan, Harris, Gibson, & Holcomb, 2000).

Van Herten, Kolk, & Chwilla (2005) argue that the P600 in response to semantic violations reflects a process of reanalysis when conflict has arisen between a semantic-based plausibility strategy and a syntactic parsing algorithm. They used semantic reversal anomalies, such as "the cat that fled from the mice" where a plausibility strategy assumes that the cat was chasing the mice, however this (wrong) interpretation clashes with the less-expected situation of mice chasing cats which would be the output from a more carefully parsed sentence. For them, the P600 represents re-processing that serves to check the accuracy of the original analysis.

Kuperberg (2007) argues for a similar interpretation of the P600, but also points out that the P600 is more likely to be observed in situations where there are strong semantic

associations between words in the sentence. She conceives the P600 to represent not a reanalysis of the sentence, but rather representing a continuing analysis of semantic and syntactic processes. She also makes the observation that an N400 effect may be attenuated, and a P600 evoked instead, when there is some contextual support for the anomalous word. This may occur, she argues, if there is a strong enough semantic association to over-ride syntactic analysis, or if the context introduces enough syntactic complexity to bias readers to construct a representation based on semantic association rather than syntactic analysis.

In relation to our study, with hard-to-detect semantic anomalies, the critical word has a strong association with the preceding context, and it is therefore not surprising that we have observed a similar effect with our data. It seems probable that the P600 partly reflects reprocessing due to task demands that emphasise the detection and reporting of anomalies. In comparison to the previously reported eye-tracking studies in this thesis, different effects were observed in the eye movement data when instructions emphasised detection and retrieval (significant effects with late measures was observed in all key regions of the text, and with early measures in the post-critical region, in Experiments 2,3, and 4) to when instructions emphasised normal comprehension and with no forewarning that anomalies would be in the text (effects were then confined to the anomalous word and only with early measures in Experiment 5). It is possible that this difference reflects the task demands rather than just the disruptive influence of anomaly detection. In the present study a late positivity was observed in both easy and hard-to-detect anomalies which might then reflect reprocessing due to the task demands in both instances. However, whether or not the P600 is a reflection of reprocessing due to the demands of the task, or due to the conflict between semantic and syntactic representations, or even the continued processing of combined semantic and syntactic analysis is open for future investigation.

One assumption that much of the previous anomaly research has made is that participants always detect the anomalies, or at least, detection occurs even if it is unconscious. For example, Nieuwland & van Berkum (2005) reported what they referred to as a *temporary* semantic illusion, because their participants identified their anomalies eventually (as evidenced by a P600 response). However, Moses-type semantic illusions are of interest because they illustrate that gross semantic anomalies can go completely undetected by participants. A critical question explored in the current study is whether or not this distinction between detection and non-detection of semantic anomalies has a different effect on the data.

Data from hard-to-detect semantic anomalies was partitioned out in to instances where the anomalies were detected and where they went unreported. It was clear that from 200ms onwards that there was a difference in the waveforms for the two conditions. Whilst anomaly detection led to a positive-going waveform, non-detection was more negative-going. Therefore, there appears to be a difference in the nature of processing from a fairly early stage that will eventually lead to either successful detection or failure to detect. This early difference in the waveforms also suggests that detection is more immediate than the effects reported in the earlier eye tracking studies would suggest. For example, in Experiment 3 (reported in Chapter 4) significant effects with early measures (which would indicate immediate anomaly detection) were only observed in the post-anomaly region. This was interpreted as reflecting delayed anomaly detection. In the present experiment, however, it is apparent that processes leading to detection occurred rapidly. These differences may in part be due to differences in the presentation of materials. In the current study target sentences are presented one word at a time and displayed for 300ms, whereas in Experiment 3 the sentences are presented whole. Therefore, a word-by-word presentation may permit processes dealing with

lexical access and integration to be completed more fully compared to a whole sentence presentation where eye movements have rapidly progressed to the next text region.

A clearer idea of how detection and non-detection affected processing was seen when these conditions were compared to the control non-anomalous condition. Detection resulted in a clear P600 type effect supporting the findings of Kuperberg, and Nieuwland & van Berkum reported above. However, non-detection did not result in the same pattern. Based on eye tracking results reported in previous chapters, it was expected that there may not be a difference observed in this comparison. Instead, a small significant effect was found in the 300-500ms time window between non-detect and non-anomalous conditions, with non-detection eliciting a more negative-going waveform (see figure 7.4). This may be interpreted in one of three ways. It might reflect that the semantic violation has been detected (as evidenced by an N400 type effect), but for some reason did not reach conscious awareness (perhaps further processing is suppressed in response to contextual factors). However, this seems unlikely when the topography of this response is compared to typical N400 responses. The N400 is commonly observed over central and parietal sites, whereas this effect has been observed over frontal and central sites instead. Secondly, this effect may in fact be a type I error, where a small biased data set has produced a seemingly significant effect on this occasion. One reason for considering this point is that detection rates were substantially higher in this study compared to the previously reported eye tracking studies (here it was 73%, whereas the highest eye tracking study was 49.7%). Another reason for being more conservative in interpreting these results is that this effect has yet to be replicated. An attempt was carried out recently in our lab where an auditory version of this study reported slightly lower detection rates for the anomalous condition at 68.4 % (which increased the amount of non-detect data) and reported a similar pattern of effects for all comparisons, except for this critical significant difference between non-

anomalous and non-detection conditions. This is illustrated in figure 7.9 which, compared to figure 7.4, shows no apparent differences at Fz and Cz when an anomaly was missed compared to the control condition.

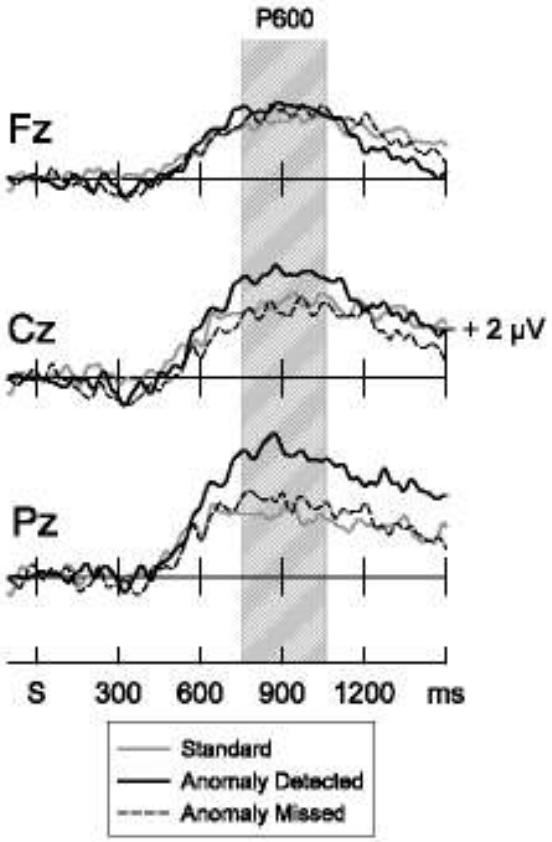


Figure 7.9: Grand average waveforms for three representative electrodes from midline sites (Fz, Cz, Pz), illustrating non-anomalous (standard), anomalous detected and anomalous missed conditions in auditory anomaly detection.

Thirdly, this effect might in fact be the result of participant's lack of confidence in the accuracy of their responses. Evidence to support this argument is provided by Eimer & Mazza (2005) who carried out an ERP visual change detection study (participants were presented with either 2 or 4 faces, 1 of which occasionally changed on second presentation), and for each trial participants rated how confident they were that their response (detect change / no change) had been accurate. They also analysed their data in a similar manner to the present study, and separated their data into detected and

undetected changes, and compared the ERP signature to no-change conditions. They reported that undetected changes evoked a more negative-going waveform within the P3 time window (500-700ms) compared to no-change conditions. Furthermore, they reported a reliable effect of confidence, so that when participants were highly confident of their response a more positive-going waveform was observed in this time-window, however, when participants had subjectively rated their confidence low for their response, a more negative-going waveform was observed instead. Unfortunately, they did not report any analysis between 350-500ms which would allow a more direct comparison to the present study. Neither did they observe any interaction between detection / non-detection and confidence (they suggested that this may have been due to the complexity of the experimental task, and such an interaction would be observed in a more simplified task). However, their results do raise the possibility that the more negative-going waveform observed when anomalies were missed may in fact reflect low response confidence, rather than any indication of implicit detection or the artefact of a restricted sample. The modulation of confidence on ERP data appears to be an interesting avenue to pursue in future work.

One important implication of this analysis is that conscious detection of these types of materials cannot be assumed, and analysis that does not partition data out into conscious detection and non-detection runs a serious risk of missing important effects in the two conditions, and misrepresenting their data.

Overall, this study has demonstrated that hard-to-detect semantic anomalies do not elicit an N400 response, which has traditionally been associated with semantic violations. Instead, a P600 response was reported which supports existing literature with semantic thematic violations where there is some contextual support for the incongruous target word and a strong semantic association between the anomalous word and the context.

However, this was only observed in situations where the anomaly was consciously detected by participants. In the anomalous condition, when anomalies were detected compared to non-detected, there were differences observed in the waveforms from 200ms onwards. Detection resulted in a late positivity, whereas non-detection appeared to be more negative (or less positive). These differences were also apparent when these conditions were compared to a control condition.

Table 7.3: Omnibus ANOVA analyses: Non-anomalous, anomalous detect and non-detect in midline and lateral recordings in all time epochs

Midline recordings	Df	100-130	150-190	200–300	300-500	500-900
Condition	2,50	ns	ns	F=4.5 p<0.02	F=4.2 p<0.03	F=8.1 p<0.002
Electrode	4,100	F=16.1 p<0.0001	ns	ns	F=4.0 p<0.01	F=7.6 p<0.0001
Condition * electrode	8,200	ns	ns	ns	F=2.3 p<0.05	F=2.7 p<0.02
Lateral Recordings	Df	100-130	150-190	200–300	300-500	500-900
Condition	2,50	ns	ns	ns	ns	F=6.4 p<0.005
Hemisphere	1,25	F=7.6 p<0.01	ns	F=15.6 p<0.0006	F=5.5 p<0.03	F=11.4 p<0.002
Anterior / Posterior	2,50	F=48.2 p<0.0001	F=13.2 p<0.0009	F=18.8 p<0.0001	ns	F=6.9 p<0.003
Dorsal / Ventral	1,25	F=18.8 p<0.0002	F=97.1 p<0.0001	F=163.8 p<0.0001	F=79.4 p<0.0001	F=77.9 p<0.0001
Cond * hemi	-	ns	ns	ns	ns	ns
Cond * Antpos	4,100	ns	ns	ns	ns	F=5.7 p<0.0003
Hemi * antpos	2,50	F=4.8 P<0.01	F=5.5 p<0.02	F=3.6 p<0.05	ns	ns
Cond * dove	2,50	ns	ns	F=4.3 p<0.02	F=4.1 p<0.03	F=11.1 p<0.0003
Hemi * dove	-	ns	ns	ns	ns	ns
Antopos * dove	2,50	F=31.1 p<0.0001	F=31.2 p<0.0001	F=44.7 p<0.0001	F=3.3 p<0.05	F=7.9 p<0.002
Cond * hemi * antpos	-	ns	ns	ns	ns	ns
Cond * hemi * dove	-	ns	ns	ns	ns	ns
Cond * antpos * dove	4,100	F=3.8 p<0.009	F=2.9 p<0.04	ns	ns	ns
Hemi * antpos * dove	2,50	F=5.0 p<0.01	ns	F=5.0 p<0.01	F=4.6 p<0.02	F=6.1 p<0.008
Cond * hemi * antpos * dove	-	ns	ns	ns	ns	ns

Cond = condition; Hemi = hemisphere; Antpos = anterior – posterior; Dove = dorsal - ventral

Table 7.4: Paired comparison between Anomaly Detect / Non-detect in midline and lateral recording sites

Midline recordings	Df	100-130	150-190	200–300	300-500	500-900
Condition	1,25	ns	ns	F=8.9 p<0.006	F=11.0 p<0.003	F=22.1 p<0.0001
Electrode	4,100	F=11.7 p<0.0003	ns	ns	F=4.7 p<0.009	F=7.4 p<0.0003
Condition * electrode	4,100	ns	ns	ns	ns	F=3.6 p<0.01
Lateral recordings	**Df**	**100-130**	**150-190**	**200–300**	**300-500**	**500-900**
Condition	1,25	ns	ns	F=4.7 p<0.04	F=4.4 p<0.05	F=13.9 p<0.001
Hemisphere	1,25	F=7.5 p<0.01	ns	F=17.3 p<0.0003	F=4.4 p<0.05	F=11.6 p<0.002
Anterior / Posterior	2,50	F=38.6 p<0.0001	F=10.2 p<0.002	F=14.7 p<0.0004	ns	F=6.8 p<0.004
Dorsal / Ventral	1,25	F=14.4 p<0.0009	F=80.7 p<0.0001	F=134.6 p<0.0001	F=52.1 p<0.0001	F=56.5 p<0.0001
Cond * hemi	-	ns	ns	ns	ns	ns
Cond * Antpos	2,50	ns	ns	ns	ns	F=7.4 p<0.005
Hemi * antpos	2,50	F=4.5 p<0.03	F=4.3 p<0.04	ns	ns	ns
Cond * dove	1,25	ns	ns	F=10.3 p<0.004	F=14.2 p<0.0009	F=31.7 p<0.0001
Hemi * dove	-	ns	ns	ns	ns	ns
Antopos * dove	2,50	F=29.7 p<0.0001	F=27.3 p<0.0001	F=34.8 p<0.0001	ns	F=4.7 p<0.02
Cond * hemi * antpos	-	ns	ns	ns	ns	ns
Cond * hemi * dove	-	ns	ns	ns	ns	ns
Cond * antpos * dove	2,50	F=7.2 p<0.002	F=3.4 p<0.05	ns	F=3.4 p<0.04	F=3.6 p<0.04
Hemi * antpos * dove	2,50	F=3.5 p<0.04	ns	ns	ns	F=4.5 p<0.02
Cond * hemi * antpos * dove	2,50	ns	ns	F=3.3 p<0.05	ns	ns

Cond = condition; Hemi = hemisphere; Antpos = anterior – posterior; Dove = dorsal - ventral

Table 7.5: Paired comparisons between Anomaly Detect / Non-anomalous in midline and lateral recording sites

Midline recordings	Df	100-130	150-190	200–300	300-500	500-900
Condition	1,25	ns	ns	ns	ns	F=4.5 p<0.05
Electrode	4,100	F=17.8 p<0.0001	ns	ns	ns	F=7.2 p<0.0001
Condition * electrode	-	ns	ns	ns	ns	ns
Lateral recordings	Df	100-130	150-190	200–300	300-500	500-900
Condition	1,25	ns	ns	ns	ns	F=4.38 p<0.05
Hemisphere	1,25	F=7.0 p<0.01	F=4.3 p<0.05	F=14.5 p<0.0008	F=4.1 p<0.05	F=10.7 p<0.003
Anterior / Posterior	2,50	F=62.2 p<0.0001	F=17.8 p<0.0002	F=29.8 p<0.0001	ns	F=10.2 p<0.0003
Dorsal / Ventral	1,25	F=17.2 p<0.0003	F=77.7 p<0.0001	F=152.7 p<0.0001	F=102.7 p<0.0001	F=118.8 p<0.0001
Cond * hemi	-	ns	ns	ns	ns	ns
Cond * Antpos	2,50	ns	ns	ns	ns	F=8.1 p<0.002
Hemi * antpos	2,50	F=5.2 p<0.01	F=6.1 p<0.01	ns	ns	ns
Cond * dove	1,25	ns	ns	ns	ns	F=11.1 p<0.004
Hemi * dove	-	ns	ns	ns	ns	ns
Antopos * dove	2,50	F=44.2 p<0.0001	F=31.2 p<0.0001	F=43.0 p<0.0001	ns	F=6.6 p<0.003
Cond * hemi * antpos	-	ns	ns	ns	ns	ns
Cond * hemi * dove	-	ns	ns	ns	ns	ns
Cond * antpos * dove	-	ns	ns	ns	ns	ns
Hemi * antpos * dove	2,50	F=5.6 p<0.006	ns	F=8.3 p<0.001	F=7.0 p<0.003	F=8.0 p<0.002
Cond * hemi * antpos * dove	-	ns	ns	ns	ns	ns

Cond = condition; Hemi = hemisphere; Antpos = anterior – posterior; Dove = dorsal - ventral

Table 7.6: Paired comparisons between Anomaly non-detect / Non-anomalous in midline and lateral recording sites

Midline recordings	Df	100-130	150-190	200–300	300-500	500-900
Condition	1,25	ns	ns	F=4.1 p<0.055	F=4.0 p<0.056	ns
Electrode	4,100	F=12.03 p<0.0002	ns	ns	F=4.7 p<0.004	F=5.1 p<0.0009
Condition * electrode	4,100	ns	ns	ns	F=3.6 p<0.02	F=2.8 p<0.04
Lateral recordings	**Df**	**100-130**	**150-190**	**200–300**	**300-500**	**500-900**
Condition	1,25	ns	ns	ns	ns	ns
Hemisphere	1,25	F=5.3 p<0.03	ns	F=10.7 p<0.003	F=5.8 p<0.02	F=9.5 p<0.005
Anterior / Posterior	2,50	F=39.1 p<0.0001	F=9.4 p<0.0004	F=11.9 p<0.002	ns	F=3.3 p<0.05
Dorsal / Ventral	1,25	F=15.9 p<0.0005	F=99.6 p<0.0001	F=157.9 p<0.0001	F=71.8 p<0.0001	F=58.7 p<0.0001
Cond * hemi	-	ns	ns	ns	ns	ns
Cond * Antpos	2,50	ns	F=3.8 p<0.05	ns	ns	ns
Hemi * antpos	2,50	F=3.9 p<0.04	F=5.0 p<0.02	F=3.9 p<0.04	ns	ns
Cond * dove	1,25	ns	ns	ns	ns	ns
Hemi * dove	-	ns	ns	ns	ns	ns
Antopos * dove	2,50	F=17.1 p<0.0001	F=26.7 p<0.0001	F=41.9 p<0.0001	F=6.3 p<0.005	F=11.1 p<0.0002
Cond * hemi * antpos	-	ns	ns	ns	ns	ns
Cond * hemi * dove	-	ns	ns	ns	ns	ns
Cond * antpos * dove	-	ns	ns	ns	ns	ns
Hemi * antpos * dove	2,50	F=4.4 p<0.02	ns	F=3.2 p<0.05	F=3.2 p<0.05	F=4.1 p<0.03
Cond * hemi * antpos * dove	-	ns	ns	ns	ns	ns

Cond = condition; Hemi = hemisphere; Antpos = anterior – posterior; Dove = dorsal - ventral

Chapter 8: Summary and Conclusions

The aim of this thesis was to investigate the nature of shallow processing. Shallow processing, as it has been defined here, refers to the notion that the contributions of syntactic and semantic processes in language comprehension may not be carried out fully, and the resultant representation of a text may be underspecified. Understanding the nature of shallow processing is of theoretical significance because it poses a serious challenge to the orthodox view of language comprehension, where the assumption has often been made that these processes are completed fully and automatically. However, as was reviewed in Chapter 1, there is a substantial body of evidence that demonstrates the ubiquitous nature of shallow or incomplete processing in language comprehension. Since shallow processing appears to be so common (see Sanford & Sturt 2002; Ferreira et al. 2002) it can not be considered as "degenerate" (MacDonald et al. 1994), and it is therefore of theoretical importance to understand the nature of shallow processing, as well as what factors may modulate depth of processing in sentence comprehension.

The best illustration of shallow processing is, in our view, when readers fail to notice semantically anomalous words or phrases in text. The term semantic anomaly is used to refer to instances when an individual word or phrase is used incorrectly, normally within a highly constraining context. The semantic anomalies utilised in this thesis are of a unique kind and we have described them as borderline-detect, or hard-to-detect, semantic anomalies, because they often go unnoticed by the reader. At the same time these anomalies apparently do not disrupt processing, as reported by off-line measures. The failure to detect the anomalous word can be taken as strong evidence for shallow processing because it suggests that processes dealing with lexical recovery and / or semantic integration have not been completed fully. These materials are, therefore, ideal stimuli for investigating the nature of shallow processing.

The studies reported in this thesis use on-line methodological techniques, namely eye-tracking and evoked potentials to investigate shallow processing. These techniques were used because they permit the investigation of the temporal properties associated with shallow processing within an anomaly detection paradigm. Participants were asked to report detected anomalies so that comparisons could be made between consciously detected anomalies, undetected anomalies, and non-anomalous controls. This allowed us to address the questions outlined at the start of this thesis: what is the time course of anomaly detection? What are the processing differences (if any) between detected anomalies, undetected anomalies, and non-anomalous controls? Is there evidence for system registration of semantic anomalies that are not consciously reported? Do factors such as task demands and processing load modulate detection rates and processing style? And, specifically in relation to ERPs, are hard-to-detect semantic anomalies processed differently from easy-to-detect anomalies?

Experiments 2 and 3 demonstrated that anomaly detection was not immediate. Significant effects were observed with early measures, which would reflect immediate detection, but only in the post-critical region. This was taken as evidence that detection is slightly delayed rather than immediate, thus challenging the assumption that semantic analyses are completed exhaustively. When data from missed anomalies were compared to a non-anomalous control condition, there were no observed reliable effects. This was taken as evidence that missed anomalies were not detected unconsciously.

Experiments 1 and 4 investigated whether processing load would modulate rates of anomaly detection. Processing load was considered to be a possible factor that would modulate depth of processing, as evidenced in rates of anomaly detection. This is because language processing is assumed to rely on limited computational resources. Under conditions where the task difficulty is high, fewer resources may be available to

devote to processing, and this may lead the processor to adopt a less efficient, or shallow processing strategy. This, in sum, would result in fewer anomalies being detected. Experiment 1 demonstrated in an off-line study that sentences containing extra parenthetical information did reduce the overall rate of reported anomalies. Experiment 4 also used extra parenthetical information to manipulate load. In this experiment the extra information was placed within the critical sentence containing the anomalous word in the high load version, or earlier in the text in the low load condition. This had the effect of reducing detection rates overall, however no reliable differences were observed between high and low load conditions. There was, however, evidence in the eye movement data that high load increased reading difficulty (as evidenced by an increased number of fixations in the post-critical region). Eye movement data again suggested that anomaly detection was delayed until the post-critical region. There was also some evidence to suggest that missed anomalies were unconsciously detected (see later discussion).

Experiment 5 investigated the effect of task demand on anomaly detection. When participants were not forewarned of upcoming hard-to-detect semantic anomalies there was little evidence of detection, as evidenced by disruption in the eye movement data. This may reflect the fact that readers adopt different reading strategies under different task instructions. That is, tasks requiring anomaly detection may result in more careful reading compared to 'normal' reading tasks. Furthermore, it was suggested that if conscious detection is not taken into consideration when analysing eye movement data in an anomaly-detection task, important effects may be lost in the data. This was demonstrated when the data from Experiment 3 was re-analysed without partitioning the anomaly condition into detect vs. non-detect. A comparison between anomalous vs. non-anomalous conditions revealed only one significant difference (there was a higher percentage of first pass regressions in the post-critical region when an anomaly was

present), which would suggest that there was little disruption when an anomaly was present. In contrast to this, when conscious detection had been taken into account, there was evidence for much wider disruption, with significant effects observed with more measures and in more regions of the text.

Experiment 6 utilised ERP methodology to explore on-line anomaly detection. Hard-to-detect semantic anomalies did not elicit the classic N400 waveform commonly associated with easy-to-detect anomalies. Instead a late positivity (P600 type) effect was observed in cases where anomalies were detected. This was not observed in cases where anomalies were not consciously reported. However, there was some evidence to suggest that missed anomalies evoked a more negative-going waveform (however, the scalp distribution of this negativity did not conform to that commonly seen with the N400). The implications of this finding will be discussed later.

The remainder of this discussion has two parts: Firstly, we will consider what is, or is not, happening when readers report, or fail to report, hard-to-detect semantic anomalies. We will also consider what is meant by the term, "shallow processing". Secondly, we will consider the implications for future and related research.

Depth of processing and shallow processing

Non-detection of semantic anomalies provides strong evidence for shallow processing in language comprehension. However demonstrating the existence of shallow processing does not explain the nature of it (Ferreira & Patson 2007). Shallow processing is ubiquitous in language and the dynamic properties which determine when and why we process to a particular depth must be fully delineated (Sanford 2002; Sanford & Sturt 2002; Ferreira et al. 2002; Ferreira & Patson 2007). The first challenge, therefore, is to describe what processes are, or are not, occurring when a

semantic anomaly is missed. Traditional psycholinguistic models (Kintsch & van Dijk 1978; Frazier 1979; Just & Carpenter 1980; Frazier & Rayner 1982; MacDonald et al. 1994) make the assumption that semantic information for lexical items is retrieved from long term memory and then integrated into an unfolding discourse model. Within this framework shallow processing may be the result of inefficient processing at either one of these stages of language processing, that is, inefficient retrieval or integration of lexical information. A further, but unlikely, possibility is that semantic information is both successfully retrieved and integrated, but readers do not become aware of the conflict between the anomaly and the global meaning of the sentence. However, given the general lack of evidence for detection without conscious awareness in the eye-tracking studies, this explanation seems unlikely and so will not be considered further.

When readers fail to notice semantic anomalies, such as "The judge sentenced the *victim* to 10 years", it may be because the relevant semantic information about *victim* (a person harmed by the crime, protected by the law, rights upheld in court, etc) has not been fully retrieved. *Victim* is a highly-relevant word within a court room scenario and is an expected role within this scenario. Therefore the high-relevancy of the word may be sufficient in itself to construct an apparently coherent representation of the story, and this may have the effect of limiting further effortful memory-retrieval processing. This explanation fits the view offered by Sanford & Garrod's (1981; 1998) Scenario Mapping and Focus theory of language processing. They argue that new linguistic information is initially mapped onto situation-relevant knowledge, for example schemas or scripts. This process is fast and passive, akin to a statistical-type analysis which merely establishes the relatedness between lexical items. Importantly, this fast passive process occurs *before* interpretation of the message and so does not represent word meaning, merely a process of checking that an individual word is relevant to a current context. If a word is not relevant it will be detected as scenario-irrelevant and receive

further processing. However, a relevant word may or may not receive such extra processing. Borderline-detect anomalies are, by their nature, words that are highly relevant to the scenario. In line with the Scenario Mapping and Focus theory these words would successfully pass an initial simple statistical association test. To illustrate using our courtroom scenario, we know that in this scenario there are characters such as criminals, victims, judges, and normally there are events such as verdicts and sentences. Once this scenario is activated we know that the likely state of affairs is that the judge is likely to sentence somebody, and we do not have to read who he sentenced in order to (seemingly) understand the event (e.g. "The judge sentenced the _____ to 10 years"). This explanation is also in the spirit of Ferreira's Good Enough approach to language comprehension. Ferreira & Patson (2007) argue that the goal of language comprehension is to establish a representation of a message that is suitable for the task (e.g. to find out what happened next in a story, or to maintain a dialogue etc.). The goal is not to establish a detailed and accurate representation of the message. Therefore a system which checks initially for relevancy (although still capable of later, effortful processing) may produce 'good-enough' representations in some situations.

The ERP literature provides additional support for this interpretation. The N400 is a negative-going waveform associated with the ease of semantic integration. When a word is easily integrated into a context a smaller N400 amplitude is observed, compared to a contextually unsupported word that requires effortful integration (Kutas & Hillyard 1984; Kutas, Lindamood & Hillyard 1984; Van Berkum, et al. 1999; Hagoort et al. 2004; Kutas & Federmeier 2000; Rugg & Doyle 1994). Borderline-detect semantic anomalies failed to elicit a classic N400 effect associated with a clear thematic violation but instead evoked a late positivity. Experiment 6 therefore provides some support for the view that hard-to-detect anomalies are processed differently than easy-to-detect anomalies. Furthermore, the general absence of an N400 effect may suggest that the

semantics of anomalous words are not fully retrieved (however, see later for a contrasting interpretation). Additional support for this view was also provided in Experiments 2 and 4, where the eye movement record failed to show significant effects with early measures in the critical region, which would be expected if the meaning of the anomalous word had been retrieved immediately.

The second possible explanation of shallow processing is that the semantic features for lexical items are successfully retrieved but for some reason these are not integrated into the discourse model. This view seems implausible because it is seems strange that a system which successfully retrieves information should not also use it. However, there may well be circumstances where this may occur. For example, in cases of high cognitive load, meaning may be successfully retrieved but due to limited resources it may not be fully integrated with the discourse model. Some evidence to suggest that this may be the case was observed in Experiment 4 where processing load was manipulated by inserting parenthetical information prior to the anomalous word. The focus in this experiment was on the borderline cases of detected and unreported anomalies. It was hypothesised that if task difficulty modulated detection rates it may reveal processing differences between detected and undetected anomalies. It was observed that when anomalies went unreported by participants they made more fixations on the anomalous word, compared to the control version. Furthermore, they took longer to read the subsequent post-critical region when an unreported anomaly was present compared to the control. While these were weak effects, they do suggest that the presence of an anomaly can cause some disruption even when participants are not consciously aware of the anomaly. Similar eye-tracking results are provided by Daneman, Rheingold & Davidson (1995) who reported significant differences in behavioural responses to homophone and non-homophone detection (participants pressed a button when an error was detected in text), and they found that both types of

errors disrupted eye-movement patterns. That is, even undetected errors led to disruption in the eye movement record. They interpreted the eye movement data as an indication that homophone errors are detected immediately, whereas the behavioural responses suggested that detection did not always reach conscious awareness.

Experiment 6 also provided some evidence to suggest that lexical retrieval without conscious awareness is possible. Experiment 6 utilised an ERP paradigm to investigate the time course of anomaly detection and non-detection. In cases where the anomalies went unreported compared to the non-anomalous controls, a more negative-going waveform was observed between 200 – 500 ms. This time window is commonly associated with the N400 which is thought to represent ease of semantic integration (e.g., Kutas & Federmeier 2000). While an N400 is the default response to open-class words, a word used in a semantically inappropriate context will evoke a larger N400 response. In Experiment 6 an N400-type response was observed in cases where anomalies went unreported by participants which may be interpreted as evidence supporting unconscious system registration of semantic anomalies. Further support for this interpretation, and with this methodology, can also be found in the literature. For example, Fernandez-Duque, Grossi, Thornton, & Neville (2002) reported ERP data based on a visual change detection task, and showed that participants detected a scene change even though they did not report the change, and Vogel, Luck & Shapiro (1998) reported N400 responses to semantically incongruous words that were not consciously perceived by participants (see also Sergent, Baillet, & Dehaene, 2005). Evidence from semantic priming studies also suggests that semantic processing occurs automatically and out with conscious awareness (for a review see Lucas, 2000). Therefore, it may be that semantic retrieval is carried out fully, but that meaning is not fully utilised or integrated into the discourse model for some reason.

Experiments 4 and 6 therefore raise the possibility that detection may occur even without overt conscious detection. However, the effects reported in these studies were weak and so may not be reliable. To make firmer conclusions requires replication of these effects. One replication of Experiment 6 (reported in Chapter 7) failed to find this N400-type effect in cases where anomalies went unreported, and so this N400-type response may not be reliable. However, this issue cannot be resolved from the data presented here. To resolve the issue, we need a technique which allows us to investigate retrieval and integration independently. One such technique which may help resolve this issue is Time Frequency Analysis which is a new approach used in the analysis of EEG and MEG data. This analysis falls outside the scope of this thesis, however it is potentially a fruitful way forward for investigating this issue, and is the focus of our present research, therefore I will describe what time frequency analysis is and its relevance to this research.

Time frequency analysis is a statistical technique that permits the quantification of neuronal synchronization, which can be expressed in terms of amplitude, frequency and phase for each electrode used in the EEG recording. The synchronization of neuronal assemblies is assumed to reflect functional networks that subserve cognitive acts. These networks may be local (spatially located within 2mm) or widely distributed. The neuropsychological interest is in the patterns of synchronous and desynchronous neuronal activity produced by the formation and dissolution of functional networks in relation to cognitive acts such as language comprehension (see for example, Singer 1999; Varela et al. 2001). Furthermore, this technique permits the changing nature of neuronal activity on a single trial-by-trial basis to be investigated thus increasing the amount of data available for analysis. This is different from ERP (and even event-related field (ERFs) with MEG) where a large number of trials are carried out, and then

averaged, to increase the signal-to-noise ratio, which has the effect of attenuating (or even cancelling) neuronal activity that is not precisely time-locked to the eliciting event.

Empirical findings using this technique have demonstrated that language-relevant processes such as memory retrieval and integration processes are associated with power changes in four frequency bands; theta (4-7Hz), alpha (8-12Hz), lower beta (13-18Hz), and gamma (above 30 Hz) (Bastiaansen & Hagoort 2006). These power changes may be illustrated using a time-frequency representation (TFR) which illustrates the percentage power changes at individual electrodes (see figure 8.1 for a simple schematic of this information in respect of semantic retrieval and unification processes).

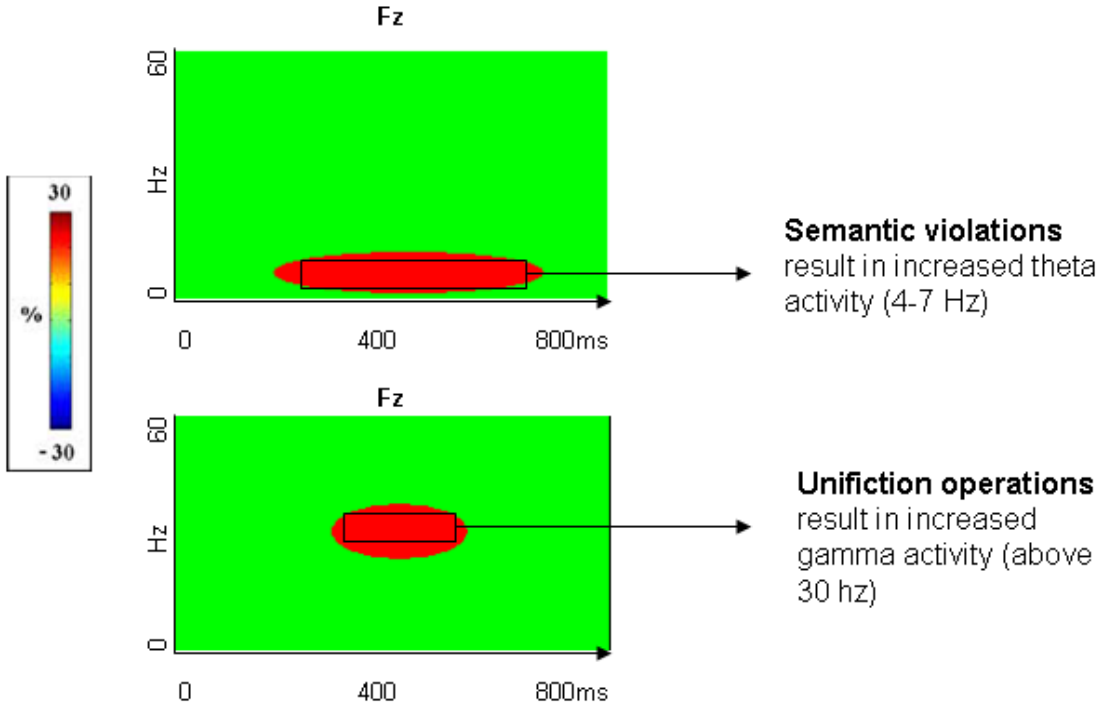


Figure 8.1: A schematic illustrating a representation of a time frequency analysis. The y-axis represents the percentage change in frequency across time (x-axis) for an individual electrode (in this case Fz).

Bastiaansen et al. (2005) reported an increase in the theta frequency band in an experimental task associated with the retrieval of semantic information. They

demonstrated this by comparing the power changes for open and closed class words. Because open class words carry the 'meaningful' content of a sentence, and closed class words provide more syntactic-type information, the differential effect in the theta band was interpreted as reflecting semantic retrieval operations. They further replicated this effect with a lexical decision task where real words elicited a larger theta power response compared to non-words (Bastiaansen 2005). Semantic unification processes, on the other hand, have been shown to modulate synchronization within the gamma band frequency range. Hagoort et al. 2004 (and Hald et al. 2006) presented participants with sentences that violated real-world or semantic expectations, compared to control statements. An example sentence was, "The Dutch trains are *yellow*/*white*/*sour* and very crowded." (Sentences were written in Dutch and presented to Dutch participants). Dutch trains are in fact yellow, so reading the word *white* violates real world knowledge. Also, the word *sour* is a semantically inappropriate word to use within this context. They observed an increase in gamma power in response to the correct word (*yellow*) and to the real-world violation (*white*), but no such response was observed for semantic violation (*sour*). They argued that unification processes were not possible in the semantic violation condition and this was reflected in the absence of gamma activity. Similar findings have been reported with semantically incongruous sentence endings (Weiss & Mueller 2003).

This analytical technique has the exciting potential to explore the nature of shallow processing in relation to the detection and non-detection of semantic anomalies. In my future research I plan to investigate the modulation of power changes within theta and gamma frequency bands in relation to anomaly detection. Specifically, if the semantic information for an anomalous word is retrieved, even when participants fail to report anomalies, then we should observe equal levels of theta compared to control conditions, along with an absence of gamma if the anomalous word cannot be integrated into the

sentence. Alternatively, if the semantic information is not retrieved, we would expect to see a decrease in theta, but with no comparable differences in levels of gamma between non-detected anomalies and control conditions.

Implications for related research

A further issue raised in this thesis has been the importance of conscious awareness in anomaly detection. This is important because many anomaly detection studies have not requested participants to report whether detection has occurred (for example, Kutas & Hillyard 1983; Ni et al. 1998). However, as has been demonstrated in the studies reported here, detection rates vary across experiments, and so even in studies that have used easy-to-detect or pragmatic anomalies, it is reasonable to assume that participants do not detect all experimental anomalies. This raises an important methodological issue, namely that if conscious detection is not taken into consideration when analysing data, important effects may be obscured or 'cancelled' out, as became apparent in Chapter 6.

Recording conscious detection, however, necessarily changes the nature of the experimental task. When participants are not forewarned of hard-to-detect semantic anomalies, effects observed in the eye movement record are substantially reduced (if not lost). One explanation for this is that participants may adopt different reading strategies in response to different task demands. Therefore, in a task requiring anomaly detection, subjects adopt a more careful and thorough processing style than when anomaly detection is not requested. The ERP results reported in Experiment 6 are consistent with this interpretation. That is, in Experiment 6 semantic anomaly detection resulted in a P600-type effect, rather than an N400 waveform commonly associated with semantic incongruity. Similar effects have also been reported with other anomaly detection

studies (Kim & Osterhout 2005; Kolk & Chwilla, 2007; Kolk, et al., 2003; Kuperberg, 2007). The P600 has traditionally been interpreted as reflecting processes dealing with syntactic analysis or reanalysis (Hagoort, et al., 1993). However, other interpretations of the P600 include, that it is a general task response (Coulson, et al., 1998); that it reflects a general 'monitoring process' for misperceptions or errors (Kolk & Chwilla 2007); and, that it is evoked when there is conflict between semantic memory and combinatorial processing streams (Kuperberg 2007). These different interpretations of the P600, argue Nieuwland and Van Berkum (under review), suggest that the P600 is a neuropsychological response that may be modulated by task instruction, which in turn affects the reading strategy adopted by participants. Therefore, if participants are not forewarned of hard-to-detect anomalies, as in Experiment 5, then the P600 response may not be evoked. In fact it would be interesting to consider whether *any* effects at all (P600 *or* N400) would be observed under such conditions.

Finally, Experiment 6 also demonstrates that strong ERP effects are observed when anomalies are consciously detected. Failure to report anomalies is taken as evidence for shallow processing. The question still remains, however, why do readers notice anomalous words and phrases in some situations, and not in others? Anomaly detection may in some situations be in response to focus effects (see Chapters 1 and 2 for a review of empirical evidence). The assumption is that anomalous words in focus are processed more deeply, and hence they are more likely to be detected, than anomalies not in focus. Taken together these studies suggest that focus and conscious awareness may be closely linked, so that focus partly determines what information is within conscious awareness, in which case this information is likely to receive preferential processing, resulting in higher rates of detection. Information not within conscious awareness is less likely to receive this level of processing and so fewer anomalies will be detected. A related argument, in respect of the modulation of conscious language

processing due to stylistic features, is advanced by literary theorists such as Miall & Kuiken (1994). They suggest that stylistic features such as foregrounding (which is used to refer to a range of devices such as alliteration, rhyme, inversion, ellipsis, metaphor or irony), has the effect of both capturing attention, but more importantly guiding the emotional reaction to, and interpretation of, literature. The interplay between focus, depth of processing, and conscious awareness, therefore, also requires further investigation.

Conclusions and way forward

In summary, the experimental findings reported in this thesis provide additional evidence for shallow processing in language processing, as evidenced by non-detection of semantically anomalous words presented within short stories. The on-line processing of detected and missed anomalies was investigated with eye-tracking and ERP paradigms. The results demonstrate that detection of anomalies results in clear effects (with eye-tracking, increased total time, number of fixations etc on the anomalous word; with ERPs an increased late positivity within a 500-900ms latency), and in cases where anomalies went unreported, there was some evidence to suggest that these may be unconsciously detected (with eye-tracking more fixations were made on the anomalous word and readers slowed down in the subsequent region; with ERPs a more negative-going waveform within the N400 time window was observed). However, these effects were weak and require replication.

The evidence presented in this thesis has also ruled out the possibility that the failure to detect an anomaly is due to a lack of encoding, if so then there would have been clear differences in the eye movement data, for example in a higher rate of skipping or lower overall reading time on the anomalous word. Further, in the ERP data there was no

evidence that participants were in a qualitatively different state (e.g. due to attentional differences as evidenced by P1 or N1) to explain when readers did or did not detect anomalies. This then raises the intriguing question of what processing does actually occur when the missed anomalous word is read.

There is substantial evidence that both semantic and syntactic processing occurs incrementally (e.g. Altmann & Kamide 1999; Altmann & Steedman, 1988; Frazier, 1979; Frazier & Rayner, 1982; Kamide, Altmann & Haywood, 2003; Marslen-Wilson, 1973; Pickering & Traxler, 1998; Sedivy, Tannenhaus, Chambers, & Carlson, 1999; Sturt & Lombardo, 2005; Traxler & Pickering, 1996), and because the evidence suggests that the anomalous word was fixated and not skipped, we can presume that some processing of the anomalous word does occur.

On the one hand non-detection may be due to the activation or retrieval of only a limited set of semantic features which define the anomalous word (assuming a semantic feature model of semantic representation). This explanation for the failure to detect anomalies has been discussed previously (see Erickson & Mattson, 1981; Kamas, Reder & Ayers 1996; Reder & Cleermans 1990; Reder & Kusbit 1991; Van Oosendorp & De Mul, 1990; Van Oostendorp & Kok 1991). This argument suggests that shallow processing is due to limited lexical retrieval.

On the other hand non-detection may also be affected by a strong biasing context. A strong context is likely to constrain the predictability of a critical word within a given scenario (Kutas & Hillyard 1984 refer to this as cloze probability). The predictability of a word within a scenario may modulate depth of processing and the likelihood of subsequent detection in at least two possible ways. A strong context may have the effect of either making an anomalous word more detectable because of the violation, or

it may make it less detectable because a shallow processing strategy has been adopted. Conversely, a weak context is likely to either lead to more detection if words are processed to a greater depth in order to establish their relevancy to the scenario, or it may lead to fewer anomaly detections if there are several possible scenario-relevant words that the reader will accept (for example, in the scenario of a plane being hijacked, the predictability of the term hijackers being used may be as likely as other agents such as terrorists, extremists, Islamisists, agitators, suicide bombers, as well as the anomalous term *hostages*). Further work is needed to explore the relationship between anomaly detection and the predictability of an anomalous word within a given scenario. One possible way of doing so is by investigating the detectability of an anomalous word as a function of the likelihood of the words use within a sentence completion task. This argument suggests that shallow processing may in fact be modulated by the reader's expectations.

The studies reported as part of this thesis have contributed to our understanding of the time course of anomaly detection, the influence of task demands on processing, and the role of conscious awareness in anomaly detection. However, there are many more questions that have been raised and these will hopefully be the focus of future research. For example, whether or not shallow semantic processing is due to inefficient retrieval or failure to integrate semantic information has not been established. Future investigation utilising new statistical techniques such as Time Frequency Analysis may help to resolve this issue. The modulation of experimental task demands on language comprehension, and how this is reflected in eye tracking and ERP paradigms, as well as the relationship between focus and consciousness awareness, all require further investigation.

Appendices

Appendix 1: Materials from original pilot study (chapter 2).

Appendix 2: Materials used in Experiment 1: Cognitive load and anomaly detection.

Appendix 3: Materials used in Experiment 2: Preliminary eye-tracking study.

Appendix 4: Materials used in Experiments 3, and 5. Including, post tracking anomaly knowledge test used in experiments 3, 4, and 5.

Appendix 5: Materials used in Experiment 4.

Appendix 6: Materials used in Experiment 6. Including, anomaly knowledge test.

Appendix 1: Materials from original pilot study (chapter 2).

Semantic anomalies are presented with their rates of non-detection for both auditory and text-based versions (expressed in percentages). Each item was presented to 15 participants. Anomalous word is underlined.

1) Pan Am flight 004 from Chicago was forced at gunpoint to land at New York's John F. Kennedy Airport. The emergency services responded quickly and all were in attendance around the international terminal building. Time was running out for the airport police. They knew that people would be killed. Under these circumstances, should the authorities meet the demands of the hostages or stand up to international terrorists?

(AUDIO 80%) (VISUAL 73%)

2) The calculated abduction and murder of a 7 year old boy shocked the nation. Evidence from a psychologist proved vital in locating the suspects. Under prosecution questioning the step-father confessed to the murder. Do you feel that the media was justified in calling for a full enquiry after the judge sentenced the victim to only 10 years?

(AUDIO 93%) (VISUAL 73%)

3) 1996, saw an increase in helplines available for the public. The creation of the 0345 prefix meant that calls would be charged at the local rate only. It was hoped that this would encourage callers to use the correct enquiry services. Towards the end of the year the government set up helplines of its own. Do you think that the use of these phonelines, such as one for members of the public to report anyone who was illegally claiming additional taxes, are an invasion of civil liberties?

(AUDIO 87%) (VISUAL 80%)

4) Scottish Police figures, for 1996, revealed a sharp increase in violent crime. Of particular concern is the increase in young people offending in Scotland. It has become apparent that there is a strong correlation between adolescent crime and alcohol consumption. Glasgow City Council has pioneered several initiatives, although it has been argued that some these might actually erode the rights of citizens. In an attempt to reduce such crimes, is it right for local councils to introduce such laws, for example, banning people from drinking in private, or should other initiatives be explored?

(AUDIO 27%) (VISUAL 20%)

5) December 1996 was a harrowing time for British lorry drivers reliant on traveling to France. The French roadways were at a standstill due to blockages organized by fellow truck drivers, involved in a pay dispute. Neither side appeared to be willing to accept the others proposals. In view of the ensuing violence, was the rejection of the pay-offer by the French Government justified, or should there have been more attempts to compromise?

(AUDIO 93%) (VISUAL 87%)

6) Last year saw a lavish open-air staging of Puccini's La Boheme, with Lucianno Pavarotti. However, bad weather brought the production to a standstill. The downpour

was so strong, the orchestra were unable to play in time because they could not even see the baton. Given that some people had paid over £100 for their seat was it justifiable for the baton to be flung down by the exasperated composer and the production stopped?

(AUDIO 93%) (VISUAL 87%)

7) Many companies are now using telephone marketing to sell their products, instead of the traditional door-to-door sales technique. The effectiveness of each method has been extensively researched. The face to face approach does appear to result in slightly more sales, but by using the telephone more people can be contacted quicker and cheaper. However, both have been accused of invading peoples privacy. Which do you consider is more intrusive, a telephone call, or someone knocking on your front door bell, to sell you goods?

(AUDIO 67%) (VISUAL 60%)

8) At the start of 1996 there was great hope in Northern Ireland. The cease fire was in its sixth month and many thought that the trouble had come to an end. However, the reality was that the sectarian divide was as strong as ever. What could a British government do to reduce the level of hatred between the Catholics and Irish, or do you think there is no solution to the problem?

(AUDIO 80%) (VISUAL 53%)

9) Recent studies conducted by psychologists at Glasgow University investigated the influence of t.v. violence on violent behaviour. They asked subjects to keep detailed diaries of their daily routines, and asked them to assess how much of their own behaviours were influenced by what they view. They found a surprising large number of people were aware that they copied the behaviour of characters on t.v.. Do you think it is worrying when people are aware that they themselves cannot separate fact from reality, or does this do very little harm?

(AUDIO 87%) (VISUAL 73%)

10) There are plans to locate a sunken galleon off the south-west coast of England. The boat is thought to have been carrying many valuable artifacts. Many other attempts have been made to retrieve the treasure, but have failed due to the depth of the sunken wreckage. However, new sophisticated diving equipment has been developed allowing divers to remain submerged for longer, but is unfortunately rather cumbersome. What dangers would you expect the divers to be facing, apart from having to carry bulky replacement water tanks, in their bid to retrieve the treasure?

(AUDIO 87%) (VISUAL 73%)

11) Recent psychological research suggests that Scottish school children score above average on tests of self-esteem and self-image, compared nationally. One interesting conclusion, is the influence of modern representations of Scottish heroes. Braveheart was quoted as inspiring over 60% of children aged between 11 and 17. How influential do you think media representations are, such as in Braveheart when Wallace's small army, courageously stood up to fight the mightier English foe, even in the face of victory, in promoting a positive national identity?

(AUDIO 87%) (VISUAL 73%)

12) The future of the NHS has been a major electoral issue. There is increasing concern from nursing unions, that their members are under-paid. UNISON has threatened strike action if a new government does not improve the present situation. However, critics argue that strike action could dangerously affect the people in their care. After considering these arguments, would you support a national strike until there is a reasonable pay settlement for all patients in NHS hospitals?

(AUDIO 67%) (VISUAL 47%)

13) In the last five years there have been an increasing number of accidents on Scottish roads. The police revealed that the majority of the fatalities involve young drivers. The government and the DVLC are considering various plans to combat this, including increasing the age when people can drive. If it would reduce fatalities on the roads, would you support plans to increase the age from 16 to 18 when people can sit their driving licenses, or are there more appropriate measures?

(AUDIO 73%) (VISUAL 47%)

14) The buoyant share prices of jewelry companies indicates that wearing bracelets, rings and necklaces containing fine jewels, is still a popular form of displaying wealth. However, whilst this is good news for retailers, it seems that the men and women employed to produce them are still badly paid. Given that their are dangers in producing these items, should consumers be forced to pay more, so that better wages can be paid to workers mining pearls in countries such as South Africa?

(AUDIO 80%) (VISUAL 60%)

15) The sectarian divide and the resultant violence in Glasgow is increasingly causing alarm. Police and Glasgow council are especially sensitive to the Rangers/Celtic clashes that they believe fuel much of this antagonism. Proposals have been considered that might limit the aggressive behaviour at football matches. If such proposals were implemented, for example, banning a full stadium with thousands of players singing sectarian anthems every time their team won, would it ruin the atmosphere of the games?

(AUDIO 73%) (VISUAL 73%)

16) Heat insulation is an important consideration for many of Glasgow's city residents living in old tenemant flats. Heating can be very expensive, especially with added VAT on fuel bills. However, many of the older and poorer members of society cannot afford to insulate their home becomes it is quite a major expense. Given the improvements to property, and the reduction in wasteful use of fuel, should Glasgow council be willing to assist some people in the cost of installing double-glazed walls, or would this be a waste of public funds?

(AUDIO 67%) (VISUAL 33%)

17) The final event of the Olympic Games in Atlanta resulted in disappointment for Britain when one runner dropped the baton, and lost valuable time. In previous events the runners had all achieved medal positions but this race would require more stamina and

endurance. What skills do you think are most important in race events such as the 100m marathon, skill, speed, team work, or mental readiness?

(AUDIO 53%) (VISUAL 13%)

18) A recent TV show was based on the idea of comparing sports stars from different sports to try to determine who was the "greatest". The panel judged each sports person according to a range of criteria, including athleticism, style and impact. Do you think it makes sense to compare across different sports in this way or is it impossible to judge, say, whether a striker's tally of winning goals is more impressive than a batsman's total of wickets in test matches?

(AUDIO 40%) (VISUAL 27%)

note: 9 of the 15 subjects reported they knew very little about cricket and were rated as detecting item, that is, due to lack of knowledge they cannot be considered as being susceptible to the illusion. Of the six remaining subjects there was 100%non-det. in audio and 67 % in text.

19) The drivers in the 1996 RAC Rally Championship had to cope with terrible conditions over the Scottish stages. The weather conditions were so bad that some contestants lost their way, and dangerously sped down roads not intended for the course. Only the four wheel drive cars were able to make it up the steep hills and a lack of grip meant that they were nearly flying on the downward tracks. Were the organisers irresponsible in allowing contestants to race, or were the drivers' aviation problems due solely to the bad weather conditions?

(AUDIO 60%) (VISUAL 40%)

20) In 1996, Chris Evans was announced as Britain's best loved entertainer. Following the popularity of "Don't forget your toothbrush" Chris was snapped up by the BBC to become the highest paid radio one DJ. However, this contract came to an unexpected end in January 1997, due to his erratic behaviour. Fans claimed this was due to Chris' excrutiating work schedule at the time, and have campaigned to see him return. In view of his past behaviour, do you think Chris would be as popular if we were able to watch him on the radio again?

(AUDIO 27%) (VISUAL 13%)

21) In a report published last year, it was claimed that the level of general knowledge of British students is extremely poor. This is present even amongst university students. In the recent Scottish Universities Mastermind quiz, one Paisley University student finished last due to a poor performance in the general knowledge round. In your opinion is it fair to tar all students with the same brush, just because one student couldn't answer the question, 'Where's Amsterdam?, because her knowledge of history was so poor?

(AUDIO 40%) (VISUAL 33%)

22) Educationalists have been calling for more advanced screening of children at primary school. It appears that children are failing in education, especially young girls. However, critics argue that young children find testing very stressful, and developmental differences

can be so varied that children should not be categorised so early. Should testing be employed, or should it be left to the responsibility of classroom teachers, to diagnose reading problems such as anorexia at a primary school level?

(AUDIO 33%) (VISUAL 13%)

23) The number of accidents involving ambulances have been on the increase. Many of the accidents have happened whilst transporting patients in need of emergency care. The public were outraged when a speeding ambulance killed a young girl who ran across a road. The driver reacted quickly, but could not divert the disaster. Should the driver be prosecuted, even though he made every effort to stop and slammed his foot on the accelerator as soon as possible?

(AUDIO 13%) (VISUAL 7%)

24) The Virgin hot air balloon took to the sky in 1997 to attempt a record braking round the world voyage for the second time. However, the heroic trip ended in near disaster when they lost altitude and nearly hit a mountain range. Richard Branson announced that he owed his life to one of the engineers on board. What must have been the most dangerous aspect facing the engineer when he climbed to the top of the aircraft to repair the damaged wing which was preventing them from gaining height?

(AUDIO 87%) (VISUAL 67%)

25) Crime is a major social worry for both citizens and the state alike. However, the measures the police have to resort to to capture criminals is causing some worry. The success in capturing major criminals by bugging private conversations has led to many convictions. However, there are calls to extend police powers, allowing the routine use of this technique. Would it be an invasion of civil liberties, if police chiefs could decide to routinely tap into local lines and bug suspects televisions, without more stringent legal controls?

(AUDIO 80%) (VISUAL 53%)

26) A recent report in the British Medical Journal suggests that back problems may stem from a combination of poor quality beds and poor sleeping position. Soft and old beds tend to be the worst offenders as the spine is not straightened. Do you think that to save the NHS money it would be a good idea to have a special allowance for people to replace the springs in their pillows every 10 years?

(AUDIO 13%) (VISUAL 0%)

27) There has been much debate in Christian communities about the issue of admitting homosexual ministers into the clergy. This follows the heated debates that preceded women being ordained into the ministry. Many progressive religious leaders have argued that the church faces falling congregational numbers and must modernise to appeal to wider sections of the community. Do you think that the current instability presents a serious challenge to the future of the Church of Scotland's monarchy, and that they must modernise?

(AUDIO 67%) (VISUAL 47%)

28) There have been many documentaries devoted to the issue of sectarianism within Christian communities. However, there are more similarities than differences between these communities. Is it your opinion that these similarities actually can unite faiths, for example basic and fundamental religious beliefs such as Jesus being the son of God, their symbols depicting Jesus on the cross at the time of the resurrection, and many moral teachings in the bible?

(AUDIO 100%) (VISUAL 87%)

29) Most of Britain's high-street banks have dramatically improved security in their branches. However, many crimes are committed by employees. The BCCI and Bearings banks were both ruined by illegal financial dealings. Do you believe that white collar crime is a greater threat to banks than the possibility of being held-up by armed police, or are both equally important?

(AUDIO 60%) (VISUAL 40%)

30) Many classical music artists have had surprising chart success with their own music. Next month sees soprano Cecilia Bartillo attempting to join their ranks. However, this trend has been criticised for reducing these works to '3-minute' pop songs. Do you share this criticism, or do you think it is a good thing that opera singers sing famous symphonies for a popular audience?

(AUDIO 80%) (VISUAL 73%)

31) Since the dismantling of the old Soviet Union, Russia has been struggling to establish a new identity as an important player in the international community. However, it has been beset with problems during this transition, including widespread criticism of the heavy-handed response to calls for independence in Chechnya. How damaging do you think the war with Chechnya was to Russia, in terms of world standing, when pictures of poorly equipped but valiant Chechnyan soldiers firing their guns in the air to celebrate their enemies' victory, was broadcast around the world?

(AUDIO 87%) (VISUAL 80%)

32) The Spice Girls have taken the world by storm. The quality of the their music appeals to the large teen market, and they have reached No.1 in many countries throughout the world. What do you think is the reason for this phenomenal success; their provocative looks, the appeal of infectious pop, that many people can't help but hum their lyrics, or is it merely successful marketing?

(AUDIO 100%) (VISUAL 80%)

33) Scottish Power carried out a nationwide survey and discovered that British electricians are being called out for the most trivial things. Rather than becoming better able to cope with appliances and gadgets the British people seem to be getting worse. Is it right for customers to be charged upwards of £50, if electricians are called out for trivial problems, such as failing to plug in the switch, or should manuals be better written?

(AUDIO 87%) (VISUAL 73%)

34) The Consumer Advisory Board, acting on information from hairdressers and dermatologists, are warning the public to stay away from dandruff shampoos. Many dandruff shampoos tested were found to have long-term damaging side-effects. Even though these products are clearly labeled, there are demands that they should be banned. Do you believe that these shampoos, which when over used can cause damage by stripping the skull, should be banned?

(AUDIO 87%) (VISUAL 73%)

35) There are many deaths every year among hillwalkers exposed to Scotland's unpredictable weather. Hillwalking centres are willing to give advice and training. Such training proved beneficial to one troop of boyscouts earlier this year. When they realised that they were lost, they used their skills to calculate where they were in relation to the sun and found their way home. They had not taken proper provisions because they were not intending to walk far. How important , do you think, that even people taking short walks in hills should be properly equipped with maps, sundials, warm clothes and water?

(AUDIO 93%) (VISUAL 80%)

36) Las Vegas is still the gambling capital of the world. Every year millions of tourists go to gamble there. Many tourists enjoy the glamour of the city. Others appreciate the opportunity to see world-famous performers. Las Vegas has staged everybody from Tony Bennet to Diana Ross and even the Kirov ballet. What, in your opinion is the main appeal of Las Vegas, gambling, brightly lit street of lights flashing with helium, or the entertainment.

(AUDIO 87%) (VISUAL 67%)

37) Levi jeans are being sold at discount prices at selected Tesco stores. The denims which normally retail at fifty pounds will be sold by the supermarket chain for thirty pounds. Levi's are reported to be very perturbed especially as they have invested heavily in promoting a high-class brand image. Do you think it should be legal for price limits to be set by consumers on their own products?

(AUDIO 47%) (VISUAL 40%)

38) Over the last few years climatic changes has resulted in more frequent natural disasters. Insurance companies, because of these, have had to pay out a great deal of money and prompted some to review their policies. Claims for events that were once thought of as 'freaks of nature', may in the future, no longer be covered by their policies. Is it ethical for insurance companies to exclude from their policies accidents such as being struck by thunder and falling on ice?

(AUDIO 47%) (VISUAL 33%)

39) A new programme on Channel 4 is reporting the sometimes bizarre customs and rituals of the British nation. The second programme investigated the British attitude to drinking tea. The British have a world-wide reputation as a nation of tea drinkers, and apparently have strict rules for how long the tea should be brewed for, and how it should be served. In your opinion, once the tea has been brewed, when should it be poured from the kettle, before or after the milk?

(AUDIO 87%) (VISUAL 67%)

40) Earlier this year saw a brave but dangerous attempt by Sharon Portmann, to be the first ever female to reach the summit of Everest alone. The freezing temperatures and high altitude defeated her solo bid. Although she survived it took two days to find her. The rescue operation cost over £500,000. Given the cost incurred , do you think that the record-breaking attempt by this team, is nothing more than vain glory, and a waste of money?

(AUDIO 93%) (VISUAL 80%)

41) One of the most popular events of the Olympics was the gymnastics. Almost everyone admired the amazing strength and control these young athletes displayed. The stadium could barely hold the number of people who wanted to watch the girls championships. The Chinese girl was champion overall, but only after a slip was caught on camera resulting in the disqualification of the Russian girl. Do you think it was fair to disqualify the Russian spectator, or should judges have been given a chance to review their scores?

(AUDIO 47%) (VISUAL 33%)

42) A U.S. serial killer recently caused uproar in the States when he requested to attend the funeral of his final victim. The murderer handed himself over to the authorities shortly after committing the brutal crime. Through his lawyer, the murderer expressed remorse for his actions, and has stated that he wants to attend the funeral service so that he will always be reminded of his crime. As yet the family have not responded. Do you think it would be fitting for a murderer to be allowed to offer his condolences to the deceased at the funeral service?

(AUDIO 93%) (VISUAL 93%)

43) Michael Jackson's baby attracted a lot of unwanted media attention . Things became particularly fierce at the child's baptism. Only a small party were invited to witness the church service, however, the press turned up in force to cover the story. Jackson and his wife released a statement after the ceremony, stating that they were angry at the press intrusion. Do you believe that there should be stricter regulations on media reporting, that would protect celebrities from unwanted press intrusion, such as in the case of the Jackson baby's adoption, earlier this year?

(AUDIO 67%) (VISUAL 53%)

44) At the Olympic games in Atlanta, the equestrian team managed to win a silver medal. Dressage was the teams strong point, and their performance was spectacular. Their success demonstrated the superb training of the horses, which required very little control. Do you think that in an event such as Dressage, medals should be presented to the trainers rather than the riders, given that the horses are so obedient that they can be halted without even having to pull on the stirrups, or is it really down to the riders?

(AUDIO 87%) (VISUAL 80%)

45) Over the last year the RNIB have been working to encourage more theatre companies to provide services for blind customers. They have had some success in Glasgow, with both the Theatre Royal, and the Kings Theatre agreeing to provide support services on all their productions. Whilst this is certainly beneficial to blind customers, it will obviously entail an increase in ticket prices. Do you think it is right that the sighted audiences should pay extra for the provision of sign interpreters for disabled theatre goers?

(AUDIO 87%) (VISUAL 73%)

46) The medical profession are advising expectant mothers not to try to deliver at home. In the last few years women have shown a greater interest in more traditional, natural birth techniques. However, a recent scare involved a young girl who narrowly escaped death when experiencing difficulties during her labour at home. Do you believe that all expectant mothers, because of the potential dangers for both parent and child, be required to enter hospital as soon as their ovulations begin, or should there be improved home care services?

(AUDIO 67%) (VISUAL 47%)

Appendix 2: Materials used in Experiment 1: Cognitive Load and Anomaly Detection.

Each item is presented with an initial context paragraph, followed by a critical question containing an anomalous word. Both high and low load versions are presented. Anomalous words are underlined (underlining was not presented to participants).

1) Many classical music artists have had surprising chart success with their own music. Next month sees soprano Cecilia Bartillo attempting to join their ranks. However, this trend has been criticised for reducing these classical works to '3-minute' pop songs.

LOW: In your opinion, is Bartillo, currently singing highlights of famous symphonies, damaging to the integrity of great classical works?

HIGH: In your opinion, is Bartillo , currently considered to be in her prime, singing highlights of famous symphonies damaging to the integrity of great classical works?

2) The consumer advisory board, acting on information from hairdressers and dermatologists, are warning the public to stay away from dandruff shampoos. Many dandruff shampoos tested were found to have long-term damaging side-effects. Even though these products are clearly labelled, these labels are often small and left unread.

LOW: In your opinion, how worrying is it that these shampoos, which through over-use can strip the skull, can be easily bought and cause serious damage?

HIGH: In your opinion, how worrying is it that these shampoos, which can be easily bought, can strip the skull and cause serious damage?

3) Whilst the worlds' stock markets have been rocked by financial disasters, companies trading in valuable jewels and metals have largely withstood the crises. Champions of developing countries, such as Oxfam, have argued that many such valuables are obtained at the expense of poorer nations, where workers are often paid very poor wages for their dangerous jobs.

LOW: In your opinion, should higher wages, which would subsidise poverty-stricken workers mining pearls in countries such as South Africa, be paid in the interests of humanitarian need, even if it would mean higher prices?

HIGH: In your opinion, should higher wages, in the interests of humanitarian need, be paid to poverty-stricken workers mining pearls in countries such as South Africa, even if it would mean higher prices?

4) There are plans to locate a sunken galleon off the south-west coast of England. The boat is thought to have been carrying many valuable artefacts. Many other attempts have been made to retrieve the treasure, but have failed due to the depth of the sunken wreckage. However, new sophisticated diving equipment has been developed allowing divers to remain submerged for longer, but is unfortunately rather cumbersome.

LOW: In your opinion, what need is there for the divers to carry, for example bulky replacement water tanks, complex machinery for hundreds of fathoms below the sea, in their bid to retrieve the treasure?

HIGH: In your opinion, what need is there for the divers to carry, for hundreds of fathoms below the sea, bulky replacement water tanks, in their bid to retrieve the treasure?

5) The future of the NHS has been a major electoral issue. There is increasing concern from nursing unions that their members are under-paid. UNISON has threatened strike action if the government does not improve the present situation. However, critics argue that strike action could dangerously affect the people in their care.

LOW: Would you support a national strike, that is demanding a reasonable pay settlement for all patients, even though it may be quite lengthy and disruptive to NHS hospitals?

HIGH: Would you support a national strike, possibly quite lengthy and disruptive, that demanded a reasonable pay settlement for all patients in NHS hospitals?

6) The calculated abduction and murder of a 7 year old boy shocked the nation. Evidence from a psychologist proved vital in locating the suspects. The world's media listened to the prosecution's questioning as the step-father confessed to the murder.

LOW: In your opinion, was the international media attention, resulting from the 10 year sentence given to the victim, a threat to there being a fair verdict given in this case, and to impartial court procedures generally?

HIGH: In your opinion, was the international media attention, resulting from the 10 year sentence decided by the judge, a threat to there being a fair verdict given to the victim, and to impartial court procedures generally?

7) December 1996 was a harrowing time for British lorry drivers reliant on travelling to France. The French roadways were at a standstill due to blockages organised by fellow truck drivers, involved in a pay dispute. Neither side appeared to be willing to accept the others proposals.

LOW: In view of the ensuing violence, should the pay offer, after being rejected by the French Government, have been reconsidered by both parties, instead of stalling the negotiations?

HIGH: In view of the ensuing violence, should the pay offer, after failing to reach an acceptable solution, have been rejected by the French Government, instead of stalling the negotiations?

8)Last year saw a lavish open-air staging of Puccini's La Boheme, with Lucianno Pavarotti. However, bad weather brought the production to a standstill. The downpour was so strong, the orchestra were unable to play in time because they could not even see the baton.

LOW: In your opinion, were the rain-soaked audience, who witnessed the baton being flung down by the frustrated composer, fairly treated when they waited for over 30 mins for the concert to begin, and were subjected to such petulant temper tantrums?

HIGH: In your opinion, were the rain-soaked audience, who waited over 30 mins for the concert to begin, fairly treated when they witnessed the baton being flung down by the frustrated composer, and subjected to such petulant temper tantrums?

9) Scottish Power carried out a nation-wide survey and discovered that British electricians are being called out for the most trivial things. Rather than becoming better able to cope with appliances and gadgets the British people seem to be getting worse.

LOW: In your opinion, if electricians are called out for trivial problems, for example for failing to plug in the switch, is it fair to penalise customers by charging £50 for the service?

HIGH: In your opinion, if electricians are called out for trivial problems, many of which are explained in the manual, is it fair to penalise customers if they have failed to plug in the switch by charging £50 for the service?

10) A U.S. serial killer recently caused uproar in the States when he requested to attend the funeral of his final victim. The murderer handed himself over to the authorities shortly after committing the brutal crime. Through his lawyer, the murderer expressed remorse for his actions, and has stated that he wants to attend the funeral service so that he will always be reminded of his crime. As yet the family have not responded.

LOW: In your opinion, should the family permit the murderer, respectfully offering his condolences to the deceased, to attend the funeral, or would this denigrate a religious service?

HIGH: In your opinion, should the family permit the murderer, respectfully attending the funeral, to offer his condolences to the deceased, or would this denigrate a religious service?

11) Pan Am flight 004 from Chicago was forced at gunpoint to land at New York's John F. Kennedy Airport. The emergency services responded quickly and all were in attendance around the international terminal building. Time was running out for the airport police. They knew that people would be killed.

LOW: Under these circumstances, what difficulties would the officials at John F. Kennedy Airport, who must negotiate the demands of the hostages, be facing when they must ensure passenger safety and possible further threats to airport security?

HIGH: Under these circumstances, what difficulties would the officials at John F. Kennedy Airport, who must ensure the safety of their passengers, be facing when they must negotiate the demands of the hostages and possible further threats to airport security?

12) Earlier this year saw a historic moment in Northern Ireland. Expectations were high as the public voted to support the Good Friday agreement. However, there are fears that some Unionists will oppose the plans and attempt to ruin the most promising project for peace the province has ever known.

LOW: In your opinion, would you expect the Good Friday agreement, which aims to restore peace between the Catholics and Irish, be successful only if dissenting Unionists factions accept the proposal, or is the history of conflict too powerful an influence?

HIGH: In your opinion, would you expect the Good Friday agreement, which is rejected by many Unionist factions, be successful in restoring peace between the Catholics and Irish, or is the history of conflict too powerful an influence?

13) Due to demographic trends detailing increased violence in our homes, studies conducted by psychologists at Glasgow University investigated the influence of t.v. violence on violent behaviour. They asked students to keep detailed diaries of their daily routines, the programmes that they watched, and assess how much of their own behaviour was influenced by what they saw. A surprisingly large number of people reported that they were aware, in hindsight, of the influence t.v. played on their own behaviour.

LOW: In your opinion, is the increase in violent behaviour due to aggressive t.v. programmes, often accused of not convincingly separating fact from reality, due to viewer's susceptibility to cosmetic portrayals of violence, or are there other causes?

HIGH: In your opinion, is the increase in violent behaviour due to aggressive t.v. programmes, often accused of being cosmetic in their portrayal of violence, due to viewer's inability to separate fact from reality, or are there other causes?

14) A recent report from the World Health Organisation shows that there are now fewer reported deaths from Aids-related illness'. This is mostly due to the development of new drugs that slow-down the spread of the disease. However, and of more worry, is that other trends report a significant increase with patients newly diagnosed with contracting the HIV virus.

LOW: In your opinion, does this trend reflect the importance of education, often supported by health prevention initiatives, in combating the spread of diseases?

HIGH: In your opinion, does this trend reflect the importance of education, often required to combat the spread of diseases, through health prevention initiatives?

15) A new programme on Channel 4 is reporting the sometimes bizarre customs and rituals of the British nation. The second programme investigated the British attitude to drinking tea. The British have a world-wide reputation as a nation of tea drinkers, and apparently have strict rules for how long the tea should be brewed for, and how it should be served.

LOW: In your experience, to make the ideal cup of tea, when would you add the milk, once brewed for 5 mins in the kettle, before or after the tea is poured into the teacups?

HIGH: In your experience, to make the ideal cup of tea, when would you add the milk, once brewed for 5 minutes, before or after the tea is poured from the kettle into the teacups?

16) Earlier this year saw a brave but dangerous attempt by Sharon Portmann, to be the first ever female to reach the summit of Everest alone. The freezing temperatures and high altitude defeated her solo bid. Although she survived it took two days to find her. The rescue operation cost over £500,000.

LOW: Given the cost incurred, do you think that this record-breaking attempt, costing so much time and effort of this team, is nothing more than vain glory, and a waste of money?

HIGH: Given the cost incurred, do you think that this record-breaking attempt, costing so much time and effort, is nothing more than the vain glory of this team, and a waste of money?

Appendix 3: Materials used in Experiment 2: Preliminary eye-tracking study.

Materials were re-written so that the anomalous word was part of the story and appeared in the second sentence. Anomalous words are underlined. Non-anomalous versions were achieved by manipulating a prior context word. This manipulation is printed in italics (as, *anomalous / non-anomalous*).

1. Opera has entered the mainstream. *Divas / conductors* have recorded many symphonies and have helped to popularise the music. The record industry is making a lot of money.
2. Oxfam champions the poor in underdeveloped countries. *Miners / divers* are often underpaid to extract pearls in harsh conditions. There is usually little medical care as well.
3. There have been many disruptions in the NHS recently. The *strike / relocation* action by the patients has been delayed due to a new offer. The governement hope to resolve the issue soon.
4. A recent trial for the abduction of a 7 year old boy shocked the nation. A 10 year *sentence / care order* was finally given to the victim but this was immediately appealed. The appeal is expected to fail.
5. Exploring undersea wreckage is a dangerous activity. The *divers / ships* need to carry replacement water tanks when they are at sea. Few trips result in new discoveries.
6. A recent opera staged in Hyde Park was brought to a standstill due to rain. *Puccini's La Boheme / Lloyd Weber's Evita* was halted by the frustrated composer when the rain got too heavy. The audience were very disappointed.
7. Criminals often regret their crimes after they have been captured. The murderer sent his *condolences / apologies*, with permission from the authoroties, to the deceased after he expressed remorse. He received a life sentence for his crime.
8. A jumbo jet was forced at gunpoint to land. *Negotiation / communication* by the authorities with the hostages was brief. The siege lasted for two days.
9. Non-denominational schools have banned the telling of religious stories. The parable of Jesus *on the cross / leaving the tomb* at the time of the resurrection is one such story that has been banned. Church leaders are furious.
10. Violent crimes are on the increase. *Ransacking / defending* the bank were armed police carrying automatic shotguns. They escaped with half a million pounds.
11. How to make the perfect cuppa: Pour the *tea / water* slowly from the kettle into a china cup. Add milk and sugar.
12. A successful stable requires well-trained horses. To stop the horse the rider *pulls / stands*, gently but firmly, on the stirrups and speaks to the horse. The rider should feel in control at all times.
13. Ante-natal care is often very good in NHS hospitals. *Expectant / hopeful* mothers are admitted to hospital as soon as their ovulations begin and receive the highest care. Many patients are happy with the care they receive.

14. Football clubs have made a stance against religious bigotry. Sectarian chanting by t*housands of / several* boisterous and ill-mannered players has brought the game into disrepute. Clubs are operating a zero tolerance policy.
15. Distressing news reports from war torn countries is causing concern. Images of soldiers *celebrating / weeping over* their enemies' victory has upset many. Viewers argue news companies should respect the 9pm watershed.
16. Sharon Portmann attempted to scale Everest in harsh conditions. Her *solo / brave* bid to reach the summit in record time was abandoned when her team refused to carry on. The attempt was criticised for poor planning.
17. A pay dispute between lorry drivers and their employers reached a crucial stage in negotiation. The *government / union* rejected the pay offer as insufficient. Eventually a compromise was accepted, however.
18. People do extraordinary things when they are in an emergency. Recently, a damaged *hot air balloon / aeroplane* was repaired by a passenger who climbed out to repair a broken wing whilst at an altitude of 10,000 feet. Fellow passengers called the man a hero.
19. Dangerous chemicals are too easily available. *Shampoos / detergents* which can strip the skull, should be banned. The most dangerous ones should only be used under special licence.
20. The Good Friday agreement offered new hope to Northern Ireland. The *sectarian violence / national identity* between the Catholics and the Irish has divided the country for many years. There is still popular support for the peace process.
21. The Inland Revenue are cracking down on criminals. They will prosecute anybody who tries to *claim / avoid* additional taxes and the penalties will be severe. Many people still attempt it though.
22. Gas and electricity companies have been criticised for their door-to-door sales techniques. Salesmen *knocking on / ringing* the front door bell of homeowners has led to many angry complaints. Some companies have even been fined.
23. Houses are cheaper to heat if there is good insulation. *Double-glazing / insulation* of windows and walls is popular amongst homeowners. This can be a big saving in cold winters.
24. Primary school teachers can help children in many ways. *Eating / reading* problems such as dyslexia should be diagnosed as soon as possible. Teachers should get involved quickly.
25. A recent race was beset with accidents. The *cyclists / pilots* faced severe aviation problems due to stormy weather. Fortunately there were no fatalities.
26. Many school children have poor general knowledge *"Where is Amsterdam?" / "Who was Churchill?"*, was failed by 40% of 15 year olds in one exam, demonstrating such a poor knowledge of history that many parents were appalled. Teachers have been criticised for this failure.

Appendix 4: Materials used in Experiments 3, and 5.

Non-anomalous versions were achieved by manipulating a prior context word. This manipulation is printed in italics (as, *anomalous / non-anomalous*). Anomalous words are underlined.

1. A recent trial for the abduction of a young boy shocked many. A 5 year *prison sentence / care order* was finally given to the victim but was later appealed. The appeal is expected to fail.
2. A pay dispute between lorry drivers and their employer reached a crisis in negotiation. The *Government / union* rejected outright the conciliatory pay offer and halted the talks. Eventually a compromise was accepted, however.
3. The high standards and success of ante-natal care in NHS hospitals has been praised. The *pregnant / hopeful* mothers enter hospital when they ovulate and are promptly seen. Many go away very happy indeed.
4. The Inland Revenue are cracking down on criminals. They will prosecute anybody who tries to *claim / avoid* new and sometimes hidden additional taxes and fine them heavily. Many people still attempt it though.
5. There is concern over the number of household products that contains highly dangerous chemicals. Some *shampoos / detergents* can dangerously strip bare the skull and should be banned. Better monitoring procedures are really needed.
6. Exploring undersea wreckage is dangerous but well worth it if sunken treasure is found. The *divers / ships* need to carry bulky replacement water tanks on their trips. Few trips result in new discoveries.
7. A recent opera staged in Hyde Park was brought to a standstill due to rain. *Puccini's La Boheme / Lloyd Webber's Evita* was halted by the frustrated composer when the clouds burst. The audience were understandably very disappointed.
8. Recently some non-denominational schools have banned the telling of religious stories. The parable of Jesus *on the cross / leaving the tomb* at the time of the resurrection has been banned first. Many church leaders are very angry.
9. The ever-growing popularity of classical music can be seen reflected in recent chart success. Respected *singers / conductors* have recorded several highly successful symphonies and have become stars. Several records have even gone platinum.
10. There was a daring and violent bank raid in Glasgow this month. The bank was *ransacked / defended* by a group of armed police who carried loaded guns. The bank lost a million pounds.
11. A North American jumbo jet was forced at gunpoint to land in Canada. The authorities *negotiated / communicated* with the scared and desperate hostages and calmed them down. The siege lasted for two days.
12. Consumer groups are increasingly concerned about highly aggressive sales techniques. Reports that door-to-door salesmen who *knock / ring* aggressively at homeowner's front door bells can be very persistent. This has led to many complaints.
13. Houses are cheaper to heat if there they are insulated. Money is well spent on *double glazing / good insulation* of the building's windows and walls and helps future sales. It makes environmental sense as well.

14. Many people are choosing not to fly because of recent aviation disasters. In one a *hot air balloon / busy charter plane* crashed due to a damaged wing, and the wreck exploded. Luckily, some did manage to escape.
15. Distressing news reports from Iraq has caused concern with many parents. Television images of soldiers *celebrating / weeping* openly due to their enemies' victory has upset many viewers. The news should censor the violence.
16. This is how to make the perfect cup of tea using a t-bag: Pour the *tea / water* carefully and slowly from the kettle into a china cup. Some serve it with a scone.
17. General working conditions for migrants is often very poor. There have been recent cases of *miners / divers* whose job is to extract pearls in often dangerous places. Safety is minimal and accidents high.
18. Sharon Portman, the champion Scottish climber, attempted to scale Everest in very harsh conditions. Her *solo / brave* bid was abandoned when her team refused to carry on. The expedition was very poorly planned.
19. Schoolteachers are being asked to help spot children's problems. Teachers may notice a problem with *eating / reading* which may be diagnosed as dyslexia sooner than their parents. They can then help the child.
20. In a recent international high profile race the leading contestants had a nasty accident. The *cyclists / pilots* contended with dangerous and severe aviation problems due to gales. Fortunately there were no serious injuries.
21. General knowledge is apparently poor for many schoolchildren. In one exam pupils failed the question, *"Where is Amsterdam?" / "Who was Churchill?",* demonstrating such bad knowledge of history that parents were appalled. Teachers were blamed for the results.
22. A successful stable requires well-trained horses and competent riders. To stop the horse the rider *pulls / stands* firmly but gently on the stirrups and clearly says, "halt". The rider is always in control.

Additional items included in experiments 5 and 6:

23. Criminals often regret their crimes, usually after they have been captured. The murderer sent his *condolences / apologies*, which were heartfelt, to the deceased along with some flowers. He received a double life sentence.
24. New changes to the National Health Service have resulted in many organisational problems. A recent *strike / relocation* lasted nearly a week by patients and greatly disrupted services. This is harming hospitals a lot.
25. The Good Friday agreement offered new hope to Northern Ireland. The central issue of the *sectarian violence between / national identity of* the Catholics and the Irish has divided the country. Many still support the peace process.
26. Scotland's football clubs have made a highly publicised stand against religious bigotry. Sectarian chanting by *thousands of / several very* loud, boisterous and unruly players has caused much offence. Clubs are trying to change attitudes.

Example participant knowledge check questionnaire used in experiments 3, 4, and 5.

Please tick the correct answer to the following statements:

1. Symphonies are recorded by:
- ❑ Divas
- ❑ Conductors
- ❑ Pop stars

2. Pearls are extracted by:

- ❑ Miners
- ❑ Divers
- ❑ Fish

3. The many disputes in the NHS has seen strike action taken by:
- ❑ Patients
- ❑ Nurses
- ❑ Teachers

4. In a court of law the judge sentences the:
- ❑ Victim
- ❑ Criminal
- ❑ Lawyer

5. Deep sea divers carry
- ❑ Water tanks
- ❑ Air tanks
- ❑ Petrol tanks

6. The composer Lloyd Webber is:

- ❑ Alive
- ❑ Dead
- ❑ Not sure

7. At a funeral you would send your condolences to the
- ❑ Deceased
- ❑ Bereaved
- ❑ Both
- ❑ Not sure

8. If an aeroplane was hijacked the authorities would negotiate with:
- ❑ The hostages
- ❑ The hijackers
- ❑ Not sure

9. In the bible, the resurrection was when:
- ❑ Jesus was re-born
- ❑ Jesus was nailed to the cross
- ❑ Jesus went to the desert alone

10. In a robbery a bank would be broken into by:
- ❑ Thieves
- ❑ Police
- ❑ The public

11. Tea is brewed in the:
- ❑ Kettle
- ❑ Tea pot
- ❑ Not sure

12. The stirrups worn by a horse are:

- ❑ Pulled
- ❑ Pushed
- ❑ Stood in

13. An ovulation is when
- ❑ An egg is released from a females' ovaries
- ❑ The contractions a woman has before birth
- ❑ Not sure

14. At a football match there might be thousands of:

- ❑ Spectators
- ❑ Players
- ❑ Dogs

15. In a battle it would be usual to celebrate your enemies
- ❑ Victory
- ❑ Defeat
- ❑ Not sure

16. The number of people involved in a solo attempt to climb Everest would be:

- ❑ One
- ❑ Two
- ❑ More

17. When a pay offer has been made, it may be rejected by
- ❑ The employees
- ❑ The government
- ❑ Not sure

18. A wing would be commonly found on a:
- ❑ Hot air balloon
- ❑ An aeroplane

- ❑ A go-kart

19. Bone can be dissolved using
- ❑ Shampoo
- ❑ Acid
- ❑ Detergent

20. The conflict in Northern Ireland is between the Catholics and:
- ❑ The Protestants
- ❑ The Irish
- ❑ Not sure

21. It is illegal to
- ❑ Pay more tax
- ❑ Avoid paying tax
- ❑ Not sure

22. Front door bells are:

- ❑ Knocked
- ❑ Rung
- ❑ Hammered

23. You would double-glaze
- ❑ Windows
- ❑ Walls
- ❑ Gardens

24. Dyslexia is:

- ❑ An eating problem
- ❑ A reading problem
- ❑ A financial problem

25. Aviation is to do with
- ❑ Flying
- ❑ Cycling
- ❑ Driving

26. The question “Who was Churchill?” would probably be found in:

- ❑ A geography exam
- ❑ A history exam
- ❑ A maths exam

27. The composer Puccini is:

- ❑ Alive
- ❑ Dead
- ❑ Not sure

28. In the Bible, the Crucifixion was when:

- ❑ Jesus was re-born
- ❑ Jesus was nailed to the cross
- ❑ Jesus went to the desert alone

Appendix 5: Materials used in Experiment 4.

Each item is presented in high load and low sentential load versions. The anomalous nature was affected via prior context manipulation (presented in italics, as *anomalous / non-anomalous*). The anomalous word is underlined.

1. LOW: A recent trial for the beating of a delinquent young boy shocked many, even the judge described it as appalling. A 5 year *prison sentence / care order* given to the victim was later appealed. The case has upset many people.

HIGH: A recent trial for the beating of a delinquent young boy shocked many. A 5 year *prison sentence / care order*, which was awarded by the judge and given to the victim, was later appealed. The case has upset many people.

2. LOW: A pay dispute between lorry drivers and their employer reached a crisis in negotiation, even the mediators seemed very dejected. The *Government / Union* rejected the initial payoffer and halted the talks. Eventually a compromise was accepted, however.

HIGH: A pay dispute between lorry drivers and their employer reached a crisis in negotiation. The *Government / Union*, who were negotiating firmly with the mediators, rejected the initial payoffer and halted the talks. Eventually a compromise was accepted, however.

3. LOW: Exploring undersea wreckage, according to the Diving Association of Scotland, is dangerous but well worth it if sunken treasure is found. The *divers / ships* carry bulky replacement water tanks on their trips. Few trips result in new discoveries.

HIGH: Exploring undersea wreckage is dangerous but well worth it if sunken treasure is found. The *divers / ships*, who are regulated by the Diving Association, carry bulky replacement water tanks on their trips. Few trips result in new discoveries.

4. LOW: A recent opera staged in Hyde Park, and starring the famous singer Kiri Te Kanawa, was brought to a standstill due to rain. *Puccini's La Boheme / Lloyd Webber's Evita* was halted by the composer when the clouds burst. The audience were understandably very disappointed.

HIGH: A recent opera, staged in Hyde Park, was brought to a standstill due to rain. *Puccini's La Boheme / Lloyd Webber's Evita*, which was starring the singer Kiri Te Kanawa, was halted by the composer when the clouds burst. The audience were understandably very disappointed.

5. LOW: Recently some non-denominational schools have banned the telling of religious stories that have been popular at Sunday School. The parable of *Jesus on the cross / leaving the tomb* during the resurrection has been banned first. Many church leaders are very angry.

HIGH: Recently some non-denominational schools have banned the telling of religious stories. The parable of Jesus *Jesus on the cross / leaving the tomb* ,which is a popular Sunday school story, during the resurrection has been banned first. Many church leaders are very angry.

6. LOW: The ever-growing popularity of classical music has a fan base around the world that can be seen reflected in recent chart success. Respected *singers / conductors* have recorded several symphonies that have topped the charts. Several records have even gone platinum.

HIGH: The ever-growing popularity of classical music can be seen reflected in recent chart success. Respected *singers / conductors*, who have many fans around the world, have recorded several symphonies that have topped the charts. Several records have even gone platinum.

7. LOW: There was a daring and violent bank raid in Glasgow this month in front of twenty terrified cashiers. The bank was *ransacked / defended* by highly armed police who carried loaded shotguns. The bank lost a million pounds.

HIGH: There was a daring and violent bank raid in Glasgow this month. The bank was *ransacked / defended* in front of twenty terrified cashiers by highly armed police who carried loaded shotguns. The bank lost a million pounds.

8. LOW: A North American jumbo jet was forced at gunpoint to land in Canada, where trained psychologists were called to help. The authorities' *negotiations / communications* with the scared hostages helped to calm them down. The siege lasted for two days.

HIGH: A North American jumbo jet was forced at gunpoint to land in Canada. The authorities' *negotiations / communications*, which used the expertise of trained psychologists, with the scared hostages helped to calm them down. The siege lasted for two days.

9. LOW: In a recent international high profile race the leading contestants had a nasty accident whilst they were competing furiously against each other. The *cyclists / pilots* contended with severe aviation problems due to gales. The organisers stopped the race early.

HIGH: In a recent international high profile race the leading contestants had a nasty accident. The *cyclists / pilots*, who were competing furiously against each other, contended with severe aviation problems due to gales. The organisers stopped the race early.

10. LOW: The high standards of antenatal and fertility care in NHS hospitals have been praised for its well trained specialists. The *pregnant / hopeful* women enter hospital when they ovulate and are carefully monitored. Many go away very happy indeed.

HIGH: The high standards of antenatal and fertility care in NHS hospitals have been praised. The *pregnant / hopeful* women enter hospital and are treated by specialists when they ovulate and are carefully monitored. Many go away very happy indeed.

11. LOW: The Inland Revenue are cracking down on criminals thanks to the Chancellor of the Exchequer. They will prosecute anybody who tries to *claim / avoid* newly introduced business taxes, and fine them heavily. Many people still attempt it though.

HIGH: The Inland Revenue are now finally cracking down on criminals. They will prosecute anybody who tries to *claim / avoid* the Chancellor of the Exchequer's newly introduced business taxes, and fine them heavily. Many people still attempt it though.

12. LOW: There is concern over the number of household products that contains highly dangerous chemicals which are freely available in shops. Some *shampoos / detergents* can strip the skull if used too much. Better monitoring procedures are really needed.

HIGH: There is concern over the number of household products that contains highly dangerous chemicals. Some *shampoos / detergents*, which are sold in shops, can strip bare the skull if used too much. Better monitoring procedures are really needed.

13. LOW: The Good Friday agreement, which the government has largely supported throughout, offered new hope to Northern Ireland. The central issue of the *sectarian violence / national identity* between Catholics and Irish, has divided the country. Many still support the peace process.

HIGH: The Good Friday agreement offered new hope to Northern Ireland. The central issue of the *sectarian violence / national identity*, which the government has tried to improve, between Catholics and Irish, has divided the country. Many still support the peace process.

14. LOW: Consumer groups are increasingly concerned about highly aggressive sales techniques that can intimidate and anger many innocent people. There are reports that salesmen who *knock / ring* repeatedly on doorbells are often too insistent. This has led to many complaints.

HIGH: Consumer groups are increasingly concerned about highly aggressive sales techniques. There are reports that salesmen who *knock / ring* repeatedly, which in itself can intimidate many people, on their doorbells are often too insistent. This has led to many complaints.

15. LOW: Scotland's football clubs are fighting religious bigotry on and off the pitch, as reported in the local press. Sectarian chanting by *hundreds of / several* boisterous and unruly players, has caused much offence. Clubs are trying to change attitudes.

HIGH: Scotland's football clubs are fighting religious bigotry on and off the pitch. Sectarian chanting by *hundreds of / several*, according to the press, boisterous, unruly and criminally irresponsible players, has caused much offence. Clubs are trying to change attitudes.

16. LOW: New changes to the National Health Service have resulted in many organisational problems for the increasingly beleaguered hospital management teams. A recent *strike / relocation* lasted a week by / as patients and services were disrupted. Hospitals are struggling to deliver services.

HIGH: New changes to the National Health Service have resulted in many organisational problems. A recent *strike / relocation*, which really annoyed the hospital management team, lasted a week by / as patients and services were disrupted. Hospitals are struggling to deliver services.

17. LOW: Houses are cheaper to heat if they are insulated throughout and a reputable firm has been used. Money is well spent on *double-glazing / good insulation* of windows and walls which helps conserve heat. It makes environmental sense as well.

HIGH: Houses are cheaper to heat if they are insulated. Money is well spent on *double-glazing / good insulation* from a reputable firm, who will fix all the windows and walls helping to conserve heat. It makes environmental sense as well.

18. LOW: Many people are choosing not to fly because of recent aviation disasters. In one serious incident in the Lake District, a *hot air balloon / busy charter plane* crash-landed because of a damaged wing, and later it exploded. Luckily, everybody managed to escape.

HIGH: Many people are choosing not to fly because of recent aviation disasters. In one serious incident a *hot air balloon / busy charter plane* crash-landed in the Lake District because of a damaged wing, and later it exploded. Luckily, everybody managed to escape.

19. LOW: Distressing news reports transmitted live from the busy battlefields of Iraq has caused concern with many parents. Television images of soldiers *celebrating / weeping* after their enemies' victory has upset some viewers. Some children have even had nightmares.

HIGH: Distressing news reports from Iraq has caused concern with many parents. Television images of soldiers *celebrating / weeping*, often transmitted live from the battlefield, after their enemies' victory has upset some viewers. Some children have even had nightmares.

20. LOW: When you are ready for a drink, this is how to make the perfect cup of tea using a t-bag: Pour the *tea / water* carefully from the kettle into a china cup. Some serve it with a scone.

HIGH: This is how to make the perfect cup of tea using a t-bag: Pour the *tea / water*, when you are ready for your drink, carefully from the kettle into a china cup. Some serve it with a scone.

21. LOW: In developing countries some people often work in terrible conditions and companies fail to properly train their staff. In one country recently there were *miners / divers* who were extracting pearls without any safety equipment. The business closed after several fatalities.

HIGH: In developing countries some people often work in terrible conditions. In one country recently there were *miners / divers* who had not been trained by the company, and were extracting pearls without any safety equipment. The business closed after several fatalities.

22. LOW: Sharon Portman, the champion Scottish climber who had been sponsored by Sainsburys, attempted to scale Everest in very harsh conditions. Her *solo / brave* bid ended when her team refused to carry on. The expedition was very poorly planned.

HIGH: Sharon Portman, the champion Scottish climber, attempted to scale Everest in very harsh conditions. Her *solo / brave* bid, which had been sponsored by Sainsburys, ended when her team refused to carry on. The expedition was very poorly planned.

23. LOW: Children's behavioural problems are often spotted first by their schoolteachers, and according to experts these are increasing. Teachers may notice problems with *eating / reading*, which could be dyslexia, sooner than their parents. Parents may be too close to notice.

HIGH: Children's behavioural problems are often spotted first by their schoolteachers. Teachers may notice problems with *eating / reading*, which experts say is on the increase and could be dyslexia, sooner than their parents. Parents may be too close to notice.

24. LOW: General knowledge is apparently poor for many schoolchildren. In one exam, which judges had rated as easy, pupils failed the question, *"Where is Amsterdam?" / "Who was Churchill?"*, demonstrating such a bad knowledge of history that parents were appalled. Teachers were blamed for the results.

HIGH: General knowledge is apparently poor for many schoolchildren. In one exam, pupils failed the question, *"Where is Amsterdam?" / "Who was Churchill?"* which judges had rated as easy, demonstrating such a bad knowledge of history that parents were appalled. Teachers were blamed for the results.

25. LOW: A successful stable requires well-trained horses, which clearly understand clear messages that are used by competent riders. To stop the horse, the rider *pulls / stands* firmly on the stirrups, and loudly says "halt". The rider is always in control.

HIGH: A successful stable requires well-trained horses and competent riders. To stop the horse, the rider *pulls / stands* firmly, which is a clear message to the horse, on the stirrups and loudly says "halt". The rider is always in control.

26. LOW: Some criminals express remorse once they have been captured, often using their lawyers to make announcements. One American serial killer sent his *condolences / apologies* to the recently deceased after he was caught. He was given a life sentence.

HIGH: Some criminals express remorse once they have been captured. One American serial killer sent his *condolences / apologies*, and a large wreath via his lawyer, to the recently deceased after he was caught. He was given a life sentence.

Appendix 6: Materials used in Experiment 6.

Context manipulation is in italics (as in, *anomalous / non-anomalous*).

Anomalous word is underlined.

1. A recent trial for the beating of a delinquent young boy by his father shocked many; even the judge described it as appalling. In the end a 5-year *prison sentence / care order* was finally given to the victim for the terrible crime.
2. A pay dispute between lorry drivers and their employer reached a crisis in negotiation, even the professional mediators seemed dejected. After five days of discussion the *Government / union* rejected outright the final conciliatory pay-offer and halted the talks.
3. According to the Diving Association of Scotland, exploring undersea wreckage is dangerous, but well worth it if new discoveries are made. On these long trips the *divers / ships* need to carry tanks filled with water where ever they go.
4. A recent opera staged in Hyde Park, and starring the famous singer Kiri Te Kanawa, was brought to a standstill due to rain. The popular open-air production of *Puccini's La Boheme / Lloyd Webber's Evita* was halted by the frustrated u when the heavens opened.
5. Recently some multi-faith schools have banned the telling of religious stories that do not reach across different faiths. The story of Jesus *on the cross / leaving the tomb* at the time of the resurrection has been banned first.
6. The ever-growing popularity of classical music has a fan base around the world that can be seen reflected in mainstream chart success. Many world famous, and very respected *singers / conductors*, have recorded several highly successful symphonies that have sold millions.
7. There was a daring and violent bank raid in Glasgow this month in front of twenty terrified cashiers. The Clydesdale bank in Govan was *ransacked / defended* by a squad of armed police, who carried loaded shotguns.
8. A North American jumbo jet was forced at gunpoint to land in Canada, experts were quickly on hand to help. First of all the authorities' initial *negotiations / communications* with the scared and desperate hostages, helped calm the situation.
9. In a recent international high profile race the leading contestants had a nasty accident whilst they were competing furiously against each other. All of the very competitive *cyclists / pilots* contended with very dangerous and severe aviation problems due to gales.
10. The high standards of antenatal and fertility care in NHS hospitals has been praised for its well trained specialists. At the very beginning the *pregnant / hopeful* women enter hospital as soon as they ovulate and are expertly nursed.
11. The Inland Revenue are cracking down on criminals thanks to the Chancellor of the Exchequer. They will prosecute anybody who tries to *claim / evade* his new value added taxes, and fine them heavily.
12. There is concern over the number of household products that contain highly dangerous chemicals which are freely available in shops. There are some very popular new *shampoos / detergents* that can strip bare the skull if used too much.

13. The Good Friday agreement, which the government has largely supported throughout, offered new hope to Northern Ireland. The central issue of the *sectarian violence / national identity* between/of the Catholics and the Irish, has divided the country.
14. Consumer groups are increasingly concerned about highly aggressive sales techniques that can intimidate and anger many innocent people. There are some salesmen who will *knock / ring* aggressively at homeowner's front door bells and are too insistent.
15. Scotland's football clubs are fighting religious bigotry, both on and off the pitch, as reported in the local press. When there is sectarian chanting by *hundreds / several of/very* loud, boisterous and unruly players, it can cause offence.
16. New changes to the National Health Service have resulted in many organisational problems for the increasingly beleaguered hospital management teams. In one hospital there was a *strike /relocation*, that lasted a week, by / of patients due to dirty wards.
17. Houses are cheaper to heat if they are insulated throughout and a reputable firm has been used. Money is well spent on *double glazing / good insulation* of the building's windows and walls which helps conserve heat.
18. Many people are choosing not to fly because of recent aviation disasters. In one accident a *hot air balloon / busy charter plane* crash-landed because of a damaged wing in the Lake District.
19. Distressing news reports transmitted live from the busy battlefields of Iraq has caused concern with many parents. News reports on television showing soldiers *celebrating / weeping* openly due to their enemies' victory has upset many viewers.
20. When you are ready for a drink, this is how to make the perfect cup of tea. When using a teabag, pour the *tea / water* carefully and slowly from the kettle into a china cup.
21. In developing countries some people often work in terrible conditions and companies fail to properly train their staff. Recently, in one country there were *miners / divers* whose job was to find pearls without any safety equipment.
22. Sharon Portman, the champion Scottish climber who had been sponsored by Sainsburys, attempted to scale Everest in very harsh conditions. Unfortunately tragedy struck, and then her *solo / brave* bid was abandoned when her team refused to carry on.
23. Children's behavioural problems are often spotted first by their schoolteachers. Observant teachers may notice problems with *eating / reading*, which may be diagnosed as dyslexia sooner than their parents.
24. General knowledge is apparently poor for many schoolchildren. A frequently failed question, "*Where is Amsterdam?" / "Who was Churchill*, demonstrates such bad knowledge of history, that teachers were appalled.
25. A successful stable requires well-trained horses, which responds promptly to the rider. To stop a horse, the rider *pulls / stands* firmly, but gently, on the stirrups and loudly says, halt.
26. Some criminals express remorse once they have been captured, often wanting to send messages to their victims. One American serial killer sent his *condolences / apologies*, which sounded heartfelt, to the deceased after he was caught.

27. The Captain of the athletics team was asked if his team would win the competition again. The tournament, which is staged every *two / single* year(s), would be theirs every year, was his arrogant opinion.
28. Psychologists' believe that soap operas can take over peoples' lives when viewers identify with popular characters. Some have become so confused between *factual / fictional* events on the telly and reality that they need help.
29. Statistics show that it is young car drivers that cause the majority of road accidents. Increasing the age when drivers can *sit / apply* for their very first driving licence might help reduce accidents.
30. The hugely anticipated Olympic events were physically and emotionally draining on these world-class professional athletes. It had been a hard, gruelling *100-metre / 26 mile* race, but very soon the marathon would finally be over.
31. Many people die, some because they are addicted smokers, and others because they inhale their smoke. Recently, the Scottish executive decided to *ban / permit* smoking in all enclosed private places to reduce smoking-related illnesses.
32. Jonathan Ross is one Britain's most popular and loved presenters and has legions of adoring fans. One obsessive fan, Mary Ogden, loves *watching / listening* (to) him so much on the radio that she writes daily.
33. He could see the other car speeding towards the junction and he realised that they were going to collide unless one of them got across first. The driver decided he had to *stop / speed up* so he quickly pressed the accelerator and prayed for luck.
34. Sleeping in old and poor quality bedding can lead to serious postural problems. Experts recommend that you check the *springs / feathers* you have in your own pillows at least every year.
35. Pete heard the new song by Kate Bush on the radio and liked it a lot. He really could not stop himself *humming / singing* those quite silly and annoying lyrics for the whole day.
36. When Dorothy returned from the supermarket she was dismayed to find that her fridge was broken. She checked that she had *plugged in / turned on* correctly the appliance's small white switch at the back end.
37. It is hoped that the democratic elections in Iraq will eventually lead to peace and stability. The new government's aim is to *restore / lessen* throughout the troubled country the chaos seen in recent times.
38. There were problems for one contestant when the music system failed at an ice skating competition. The young Canadian ice skater was *disqualified / jeered* by the panel of international spectators because she wouldn't dance.
39. Scotland has chronic levels of heart disease and obesity, and Scotland's politicians want to change this. The Scottish Executive is trying to *prevent / encourage* people from / to adopting a healthy lifestyle and halt this trend.
40. Cathy found herself caught in the storm, which was the most violent she had every seen. She was lucky not to be *struck / deafened* by the wild and terrifying thunder which really scared her.
41. The job interview is the favourite, albeit difficult, method used by firms for recruiting new members. When it comes to hiring staff *customers / employers* must assess the qualifications of applicants and their general suitability.
42. The young lovers met secretly beside the old yew tree, far away from their families gaze. The warm, golden light of the *evening / morning* sun kissed the sky as dawn spread across the land.

43. The police and government, fearful of violent terrorists, now have increased powers to help fight criminals. The police can now place a *tap on / bug in* any potentially dangerous suspect's television without going to court.
44. It was the height of the holiday season in the ever-popular Northern coastal resort of Blackpool. In every direction there were bright *lights / balloons* that were completely filled with helium advertising fun and festivities.
45. The long-standing war between the Hutus and Tutsis had been violent, and many people had died. The odds were against the Hutus *losing / winning*, but even as they faced defeat, they raised their weapons.
46. He was accused of killing the seven women, and the court case had lasted three months. After a trial that established his *innocence / guilt*, the judge finally gave him life, much to everybody's surprise.
47. In an age of political uncertainty the United Nations vows that it will protect defenceless countries. An innocent country, with a *friendly / aggressive* neighbour who wants to do them harm, will always receive support.
48. A survey by a travel company asked their customers what they wanted in a holiday destination. Many British tourists hit Spain to get *away/closer from/to* the high summer sun for two relaxing weeks.
49. On November the 5th all over Britain, bonfires are lit, Guy Fawkes' burn, and fireworks explode. Many children and adults enjoy watching the *Catherine Wheels / rockets* in the dark autumnal sky because they are pretty.
50. Billy joined the local charity and asked his supervisor what they could do for the children. The volunteer worker went to *take from / give to* the deprived children their Christmas presents and wish them luck.
51. Many schools are struggling to teach Art properly because of the spiralling cost of art supplies. In one class the pupils kept *sharpening / refilling* their new and expensive calligraphy pens that they were using.
52. Rosie's grandmother went to the fashionable boutique to get some ideas for her granddaughter's birthday present. She learnt that scarves are very *popular / unpopular* amongst girls who spend little money on them each month.
53. Clive was angry after Amanda called him a fool, but didn't want to lose his temper. During the fight, she had *looked at him / spoke to him* with such a contemptuous voice that he almost cried.
54. Meredyth and Derrick had been arguing when she suddenly accused him of having an affair. The shocking accusation just took the *wind / words* right out of his gaping mouth and he fell silent.
55. The pressure was on as the contestants prepared for the next quick fire round of questions. Judith had quickly slumped in to *a deep coma / state of profound* and emotionally intense competitiveness after the last round.
56. The boyscout had worked hard and won an award for all the knots he could tie. He was happy to wear his *shirt / badges* right on his crisply ironed sleeve and did so proudly.
57. The tackle was dirty and the referee blew his whistle as the player writhed in agony. The medic quickly assessed the man's *shoulder / leg* as serious and the player limped painfully off the pitch.
58. It has taken Kate Bush twelve years to release a new album and expectations were high. The majority of music critics couldn't *fault / praise* the new double album too highly and it sold well.

59. Many people take board games very seriously and in school teams the competition can be fierce. Andrew never recklessly moved his own *pawns / pieces* and was the champion of checkers in his school team.

60. Margaret had a bad fall and everybody was very concerned. The doctor suspected she'd a *broken leg / had a seizure* and requested they scan her brain as soon as possible.

61. June's sales team had worked hard and they expected a good bonus for all their effort. Finally they could see there was *light / gold* at the end of the rainbow if they kept going.

62. Janet was a truly exceptional singer and nobody was surprised when she won the talent show. She was determined to enjoy her *night / day* in the hot and shining sun, no matter how brief.

63. It was the premiere night and the excited audience began to rush to their seats as the auditorium lights began to dim. The actors quickly took to *the/their stage / seats* and were ready for the film to begin on time.

64. The rugby team have won all their matches and were beginning to act a little cocky. They were seen as the big *birds / fish* to be removed from their perch by the other teams.

65. Some cosmetic companies have been criticised because their advertising campaigns for anti-ageing products target people's insecurities. One slogan stated that every *second / minute* of every day we age a minute, was especially singled out.

66. The snow had been falling all morning and now it had even begun to freeze over. Bill went to work even though *driving / walking* was very slow and difficult underfoot because of the ice.

67. A man was attacked in Kelvingrove Park and was stabbed and had his wallet stolen. As yet there's been no *statement from / identification of* the elderly man who was murdered which is very unhelpful.

68. Sometimes we all have to take risks and tackle difficult situations in life to get anywhere. The saying, if you play with *fire / bees* you are going to get stung, warns you of this.

69. Marian's grandmother said she was still behaving as if she'd just discovered sex. Sex is even older than *Jonah / Noah*, who was famous for building the ark, she told her sternly.

70. Zoos allow us to study animals, where we can learn to understand and help them better. Some zoos even provide useful *informational resources on / breeding grounds for* exotic species that are extinct for visitors to enjoy.

71. A book about Princess Diana's rows with her family is out which describes her long feuds. One argument lasted for six months *after / before* she visited Paris and was killed in a car accident.

72. Every now and then an amazing sporting celebrity appears on the scene and captures everybody's imagination. The amazing debut of *Tiger Woods/ Steve Davis* was the most exciting thing in snooker because he raised standards.

73. The psychology demonstrator hung up several posters informing students that they couldn't eat in the labs. For the past two years these *mandatory / particular* rules had not always been insisted upon in the labs.

74. Rugby can be a dangerous sport and violent clashes on the pitch can result in injuries. One injured player had to *finally walk / be carried* off on an old fashioned stretcher to see a doctor.

75. Whilst on the bus, on her way home from work, Judith began to list her evening chores. She went home to wash and *cook / feed* her three wild and unruly children before doing the hoovering.

76. Everybody was disappointed with the Glasgow University team's performance in the end of year relay race. The team were losing, and kept *kicking / shooting* themselves right in the proverbial foot by dropping the baton.

77. The school coach had trained the children well, but was really proud of one pupil particularly. Peter was so amazingly fast, he *swam / ran* just like he was a greyhound, easily beating the others.

78. Theresa had just started her final year of University studies and she intended to do well. After working hard there was a *carrot / light* at the end of the tunnel because of her holiday.

79. The naughty schoolboys had been fighting on the muddy playing field and both had black eyes. The headmaster scolded them and they *hung /wrung* their tired and very muddy heads in deeply felt shame.

80. Delia and Don really enjoyed eating soft-boiled eggs, especially when they cut their toast into soldiers. She dipped her toast into the *white / yellow* of the very lovely runny yolk and ate it up.

81. Once more, James had possession of the ball and had a clear shot of the goal. His second great goal was a *blueprint / copy*, it was said, of the first and was pretty cool.

82. The television crew interviewed the old vaudeville star that had made his reputation in the 1930's. Most of the other stars who *can / could* remember him performing are now dead because they're so old.

83. David had worked very hard on his party costume and hoped to win the top prize. The fancy dress theme was *Robin Hood / Three Muskateers*, so he dressed up as D'Artagnian, and won the prize.

84. One high street store made a list recently of their biggest selling and most popular items. Plasma screen televisions are popular and *consumers / shops* have set a very high price for these desirable goods.

85. On Kate Bush's album, Aerial, there is a song about the Catholic saint Joan of Arc. Many think that Joan of Arc was *immortal / fictitious*, but she did in fact exist in 15th century France.

86. A boxing match is meant to last for four rounds, each one lasting for two minutes. Whilst this brutal sport sees many *deaths / injuries* every year, none are actually serious and the sport's popular.

87. The restaurant manager regretfully told Mary that they were very busy and she'd have to wait. He had a table that was currently *empty / full* but was likely to be vacated in about 20 minutes.

88. A jockey without a whip at the Grand National is useless. Metaphorically, he is just like a *carpenter / plumber* without his trusty and useful spanner vainly attempting the job.

89. George Bizet's most popular opera, Carmen, is the story of a wild and passionate Spanish gypsy. He didn't appreciate the success *until after / properly before* his surprising and rather sudden death later on that year.

90. Investing money on the stock market can be a high-risk strategy for earning high profits. Money can quickly go *up and down / round and round* like happy children playing on swings in the school playground.

91. It was an icy cold day outside and Jack decided to put on his warmest clothes. He put on his patent leather *boots / jacket* and then his warm woolly socks so he'd stay warm.

92. Few people ever walk down those secret corridors of power where cigar-chomping businessmen decide our fate. International monetary decisions are made *behind doors / within rooms*, that are filled with blue smoke, where the powerful meet.

93. Bob and Clare had saved up money all year for a fantastic holiday after they graduated. They travelled all the way to *Siberia / Ecuador* where the weather was usually hot for the whole year.

94. The men were proud in knowing how they should properly behave in an aristocratic country house. The very old and extremely tired *Lord / workman* listened to, and obeyed the butler, quickly finishing the job.

95. Emily warned Sarah to expect a lot of mess when she walked in to the room. However, when she saw a fine *white / black* dust everywhere due to the coal delivery, she was angry.

96. Barbara and Adam were on a weekend city break and had spent the morning shopping in the busy market, but wanted to go somewhere quiet for lunch. One of the quietest spots in *London / Paris* is next to the river Seine where they could sit.

97. The two boys enjoyed playing games together and tonight it was a close match. Charles surveyed his pieces in the *chess / board* game and promptly moved the monk which infuriated his opponent.

98. The Romans worshipped many different Gods and Goddesses and they each represented something different. One of the most important was *Neptune / Mars*, who was the God of war and he was powerful.

99. Rebecca and James were very much in love, but James had gone on a long holiday to Brazil. Poor Rebecca mournfully counted the many *years / hours* until the end of the month when they'd meet again.

100. The Giant Panda bear exists in the wild in only a few isolated places in China. These wild and fragile habitats contain *palm / bamboo* trees which pandas need for food if they're to survive.

101. When volcanoes erupt the amount of destruction they cause can be terrifying. In one disaster, many people in *Rome / Pompeii* died because the eruption of Vesuvius took them by surprise.

102. It was the biggest ship of its day and no one was prepared for the disaster. On her maiden voyage in the *Indian / Atlantic* Ocean an accident sunk the Titanic in a few hours.

103. There was a recent boating disaster and all hands were lost at sea. The boat was sailing in the *Mediterranean sea / Arctic ocean* when it hit an iceberg and very quickly sunk.

104. Robin Hood is one of England's greatest myths and has inspired countless stories. Dressed in green, his merry men *stole from / gave to* the many unobservant local poor their gold and valuables.

105. The thieves wanted to steal one of the world's most famous paintings from the Parisian museum. Hanging majestically in the Louvre is *Rembrandts / Da Vinci* most famous paintings, the exquisite Mona Lisa, and they wanted it.

106. Children love fairy tales because they're great stories. In one story there were three *pigs / bears* and a little girl called Goldilocks, who'd eaten their porridge.

107. It was the children's dream holiday; the best funfair in America, with the biggest rides, and all their favourite cartoon characters too. Emma enjoyed most her meeting *Bugs Bunny / Mickey Mouse* on her second day at Disneyland because he hugged her.

108. The fairy princess had danced all evening with the handsome prince, but the ball would soon be over. It was so very late when *Snow White / Cinderella* ran away and left her slipper on the stone steps.

109. It had been a long journey, walking through the forest and carrying a basket of food. Opening the bedroom door *red riding hood was / the seven dwarves were* surprised to see a wolf in the large bed.

110. Australia is famous for its flora and fauna. One quintessentially Australian animal is the *emu / koala* that eats the leaves of eucalyptus plants in vast quantities.

111. The sociologist was interested in the home life of Britons in the 21st century. He carried out a survey of *industrial / nuclear* and, when he could, extended families, in five British cities.

112. The war in Iraq had been a failure, and many people were very angry. The anti-war protestors waited for the *president / prime minister* and when he finally appeared Blair was booed and jeered.

113. Many British holidaymakers want more adventure and are willing to travel further to get it. One of the most popular *islands / countries* for a long summer holiday is Brazil because of the rainforest.

114. All the children enjoyed the holiday season because they could dress up and have fun. In every window were *carved pumpkin heads / colourful little decorations* that glowed brightly to celebrate Christmas and everybody had fun.

115. When Dorothy found herself in Oz she realised she had to find a way home. Her travelling companions included a kind *scarecrow / lion* who did not have any courage, on her brave quest.

116. In a recent BBC programme, Robin Muir explored the relationship between photography and the perception of famous people. One of the most photographed was *John F Kennedy, who was America's / Rainier Grimaldi, who was Monaco's* most famous prince in the 20th century.

117. David's African safari was a great success and they had just spotted a solitary animal in the distance. Using his binoculars he saw *five / two* humps on the back of the camel as it drank water.

118. King Arthur was a proud host and liked to look after his guests himself. Along with the king, *seated at / waiting on* the round table, were all the peasants of the royal court.

119. They awoke to find the snow outside had settled thickly, but they were more excited about their presents. The couple opened their *anniversary / Christmas* presents, which had been delivered overnight by Santa and they laughed happily.

120. Specialist equipment allows us to see nature more clearly. Jeremy had to quickly focus his *microscope / telescope* so as to see the bird landing in its nest.

121. Sarah's GP prescribed her a fast acting drug. She sipped cold water with her *sedative / stimulant* and soon she felt like dancing right around the room. .

122. Comic strips are a popular feature of newspapers. A very popular character is the *cat / dog* from Peanuts, who was called Snoopy created by Charles Shulz.

123. In a recent poll Americans voted for the most influential figures of the 20th century. The very first man *on the moon / of traditional jazz* was the now legendary and respected Louis Armstrong**,** and he came 7th.

124. In 1939 the Second World War began and Hitler seemed unstoppable. The turning point came when the *Germans / Japanese* attacked the unsuspecting people of Pearl Harbour and America declared war.

125. The relationship between American and Russia used to be very frosty. Events changed when visionary leaders of *capitalist / communist* countries, for example men like Gorbachev, opened up their borders.

126. The weather had taken everybody by surprise and John and Jane were happy to get inside. They rushed to switch on the *central heating / air conditioning* so the room would cool to a comfortable temperature.

127. Dorothy and Sam were having a dinner party, but their guests were arriving and they weren't ready. Sam quickly mashed up some fresh *artichokes / avocadoes* to make his favourite dip, guacamole, which he served first.

128. The sea captain stood on the deck urgently surveying the horizon. Once he had focussed his telescope Captain *Nemo / Ahab* could clearly make out Moby Dick in the far distance.

129. Mary was studying Shakespeare for her English literature exam and was memorising quotes for the exam. *"To be or not to be, that is the question" / "Double, double toil and trouble; fire burn and cauldron bubble"* from his Macbeth was one she learnt.

130. Historical dramas are very popular on television, especially when they're about the royal family. A popular story is of the *8 / 6* women who ended up marrying Henry 8th and is being filmed now.

131. Many countries have their traditional forms of entertainment. There were specialist female entertainers in *China / Japan*, who were traditionally known as geisha, and were highly regarded.

132. There seemed to be no end to the villainy of Lex Luther after he had kidnapped Miss America. Clark Kent had to *change in to Superman / travel along the highway* so he went to the tollbooth on the busy road.

133. The primary school teacher carefully prepared her lesson plan. They would learn all twenty six *numbers / letters* that make up the whole alphabet in the class today.

134. The Scottish Tourist Board is trying to attract more tourists to Scotland who want short city breaks. One popular destination is also the *largest / capital* city of Scotland, which is Glasgow on the river Clyde.

135. Peter enjoyed watching other people when he was out for a drink, especially when they were behaving interestingly. He was attracted to a young *brunette / skinhead* who had very brown wavy hair and acted very sulkily.

Example participant knowledge check questionnaire used in experiment 6.

PLEASE CIRCLE THE CORRECT ANSWER(S):

1. Tiger Woods is famous for which sport?

Tennis Snooker Golf

2. Fairground lights are powered by:

Helium Water Neon

3. An innocent man should be:

Given life Set free Remanded in custody

4The number of players you would expect to be on a pitch would be:

Hundreds Thousands Several

5. Who would you expect to go on strike in a hospital?

The nurses The patients The visitors

6. A blueprint is:

A copy A plan Don't know

7. If you weren't doing well, you might be said to have "shot yourself in the …."

Foot Leg Knee

8. Can a murdered man make a statement to the police?

Yes No Don't know

9. How long does a second of every day last?

A minute A second An hour

10. At the time of the resurrection, what was Jesus doing?

Leaving the tomb Nailed to the cross Feeding the 5,000

11. Aviation problems are usually met by:

Cyclists Pilots runners

12. Driving a car over ice would be difficult

Underfoot Under the car tyres Don't know

13. Where would you watch a Catherine wheel?

Nailed to a post In the sky In the house

14. An annual tournament would be

Every year Every two years Every three years

15. When the end is in sight, you might say, "you can see the light at the end of –"

The rainbow The tunnel The stick

16. When offering an incentive, you might say, “dangling a carrot at the end of -”

The rainbow | The tunnel | The stick

17. A prison sentence is given to the -

Victim | Criminal | Judge

18. Which of these can strike you?

Thunder | Lightning | Clouds

19. Could you correctly describe a 100-metre race as a:

Sprint | Race | Marathon

20. What colour is the yolk of an egg?

White | Yellow | Red

21. When you hum a song, do you hum the,

Tune | Lyrics | Words

22. If somebody is immortal, does it mean that

They live forever | They are dead | They are fictional

23. If somebody had a broken leg, would the doctors x-ray the,

Brain | Leg | Arm

24. If you ask somebody to reminisce about the past, they usually need to be,

Alive | Dead | Either

25. Schoolchildren might study where capital cities are in which subject?

History | Geography | Maths

26. Governments tend to encourage their citizens to,

Live healthily | Live unhealthily | Adopt alternative lifestyles

27. People who have arguments with their families must be,

Alive | Dead | Either

28. A shock may “take the wind out of …..”

The sails | The mouth | The cat

29. Would you knock on a -

Door knocker | Door bell | Door mat

30. Would you celebrate your enemies -

Victory | Defeat | Either

31. Are the Americans in Iraq trying to restore,

Peace | Chaos | Don’t know

32. Describing somebody’s tone, would this refer to,

How they speak | How they look | How they act

33. If the odds are against you losing, then you are likely to,

Win Lose Don't know

34. Which would you double-glaze,

Windows Walls Floors

35. Which tools would a carpenter carry,

A chisel A spanner A blow torch

36. New staff are hired by,

Customers Bosses Applicants

37. Would you enjoy the sun at

Night time During the day Both

38. Most package holidays to the Mediterranean would take you,

To the sun Away from the sun Both

39. Over-use of shampoos might strip

The skin The skull The fingers

40. If you wanted to praise something, you wouldn't be able to fault it too …

Highly At all Not sure

41. Do you go round and round on

Swings Roundabouts Slides

42. Which is an eating disorder?

Autism Dyslexia Anorexia

43. Are the troubles in Ireland between the Catholics and

Protestants Irish English

44. Where is smoking banned in Scotland,

Open space Private spaces Public spaces

45. Is the composer, Puccini,

Alive Dead Don't know

46. What are pillows stuffed with?

Spring Feathers Cement

47. When a plane has been hijacked, who would the authorities negotiate with?

The hostages The hijackers The psychologists

48. When making dinner, would a parent wash and cook the,

Vegetables Children Plates

49. To appreciate something, do you need to be,

Alive Dead Doesn't matter

50. Negotiations for a new pay deal would normally be between the employees and

The employers The government The customers

51. How do you connect a fridge to a power supply?

With a fridge	With a switch	With a fuse

52. Where do you find fish?

On perches	In pools	In trees

53. If you touched a naked flame you would be,

Burnt	Stung	Bruised

54. Do pregnant women still ovulate?

Yes	No	Don't know

55. If a man is limping, he's likely to have hurt his

Leg	Shoulder	Either

56. If a stretcher is used to transport a man, he would be

Being carried	Walking	Floating

57. Which would you watch somebody on -

The radio	The TV	A book

58. Which story is D'Artagnian from?

The three musketeers	Robin Hood	Cinderella

59. Are mandatory rules,

Enforced	Ignored	Don't know

60. Symphonies are recorded by,

Conductors	Singers	Rappers

61. If a table is soon to be vacated, then it is presently

Occupied	Empty	Don't know

62. Who was Jonah?

He lived in a whale	He built the Ark	Don't know

63. Would a solo walker also have a team?

Yes	No	Don't know

64. What kind of tanks do underwater divers need to carry?

Water	Air	Soup

65. When is the dawn?

Morning	Evening	Afternoon

66. If you are shamefaced, would you wring your,

Heads	Hands	Necks

67. When somebody is killed, is it a

Serious injury	Not a serious injury	Sometimes

68. Pawns are found in which game?

Chess	Checkers	Scrabble

69. Where do you pour tea from?

A kettle	A teapot	A teacup

70. Which is illegal,

To claim taxes	To avoid taxes	Not sure

71. A contestant would be disqualified by,

The judges	The spectators	The coaches

72. To stop a car, you would step on the,

Brake	Accelerator	Clutch

73. Which would a friendly neighbour NOT do to you?

Hurt you	Help you	Leave you alone

74. Which of these would you get through mining?

Pearls	Diamonds	Ice cubes

75. If something is extinct, it means that it is,

Still alive	No longer exists	Don't know

76. In surveillance, which household objects could you place a tap on?

T.V.'s	Telephones	Hairdryers

77. If you were in a coma, would you be,

Asleep	Competitive	Unsure

78. Which flying object has a wing?

A plane	A balloon	Both

79. Film actors star on

The stage	The screen	The chipshop

80. Shops are generally broken in to by,

Robbers	Police	Rabbits

81. The opposite of factual is

Reality	Fictional	biographical

82. The prices for consumer goods is determined by,

The customer	The seller	Not sure

83. Who do charities give to?

The poor	The rich	Both

84. Who would you send your condolences to?

The bereaved	The deceased	Both

85. Which do you sharpen?

Pens	Pencils	Rulers

86. Which animals move by swimming?

Fish	Greyhounds	Parrots

87. What do horse riders pull on?

Stirrups	Reins	Saddles

88. When learning to drive, what do you sit for?

Driving licence	Driving test	Motor insurance

References

Albrecht, J., E., & Clifton, C., Jr. (1998). Accessing singular antecedents in conjoined phrases. *Memory and Cognition*, 26, 599-610.

Altmann, G.T.M., &Kamide, Y. (1999). Incremental interpretation at verbs: Restricting the domain of subsequent reference. *Cognition*, 73, 247-264.

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

Ariel, M. (1990). *Accessing noun-phrase antecedents*. London: Routledge.

Baker, L., &Wagner, J., L. (1987). Evaluating information for truthfulness: The effects of logical subordination. *Memory and Cognition*, 15, 247-255.

Barton, S., &Sanford, A.J. (1993). A case study of anomaly detection: Shallow semantic processing and cohesion establishment. *Memory and Cognition*, 21, 477-487.487.

Bastiaansen, M.C.M., Van der Linden, M., ter Keurs, M., Dijkstra, T., &Hagoort, P. (2005). Theta responses are involved in lexical-semantic retrieval during language processing. *Journal of Cognitive Neuroscience*, 17(3):1-12.

Bastiaansen, M.C.M. (2005). *Do you see what I mean? Theta power increases are related to the retrieval of lexical semantic information*. 5th Endo-Neuro-Psycho meeting.

Bastiaansen, M.C.M., &Hagoort, P. (2006). Oscillatory neuronal dynamics during language comprehension. In: C. Neuper, W. Klimesch (Eds.), Event-related dynamics of brain oscillations. *Progress in Brain Research series*, Vol. 159. Elsevier, Amsterdam: 179-196

Besson, M., Kutas, M., & Van Petten. C. (1992). An event-related potential ERP analysis of semantic congruity and repetition effects in sentences. *Journal of Cognitive Neuroscience*; 4: 132–149.

Bever, T.G. (1970). The cognitive basis for linguistic structures. In R. Hayes (Ed.), *Cognition and language development* (pp. 277-360). New York: Wiley and Sons, Inc

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

Brédart, S., &Docquier, M. (1989). The Moses illusion: A follow-up on the focalisation effect. *Cahiers de Psychologie Cognitive/ European Bulletin of Cognitive Psychology*, 9, 357-362.

Brédart, S., &Modolo, K. (1988). Moses strikes again: Focalisation effect on a semantic illusion. *Acta Psychologica*: *International Journal of Psychonomics*, 67, 135-144.

Büttner, A.C. (2007). Questions versus statements: Challenging an assumption about semantic illusions. *The Quarterly Journal of Experimental Psychology*. 60, (6), 779-789.

Carpenter, P.A., Miyake, A., & Just, M.A. (1995). Language comprehension: sentence and discourse processing. *Annual Review of Psychology*, 46, 91-121.

Christianson, K., Hollingworth, A., Halliwell, J.F., & 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(2), 205-238.

Coulson, S., King, J., & Kutas, M. (1998): Expect the unexpected: Event-related brain responses to morphosyntactic violations. *Language and Cognitive Processes* 13:21-58.

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

Daneman, M., Lennertz, T., & Hannon, B. (2007). Shallow semantic processing of text: Evidence from eye movements. *Language and Cognitive Processes*, 22, 85-105.

Daneman, M., Reingold, E.M., & Davidson, M. (1995). Time course of phonological activation during reading: Evidence from eye fixations. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 21, 884-898.

Dien, J., & Santuzzi, A.M. (2005). Application of repeated measures ANOVA to high-density ERP datasets: a review and tutorial. In: T.C. Handy, Editor, *Event-Related Potentials: A Methods Handbook*, MIT Press, Cambridge, MA, pp. 57–82.

Eimer, M., & Mazza, V. (2005). Electrophysiological correlates of change detection. *Psychophysiology*, 42, 328-342.

Erickson, T.D., & Mattson, M.E. (1981) From Words to Meaning: A Semantic Illusion. *Journal of Verbal Learning and Behaviour*, 20, 540-551.

Fernandez-Duque, D., Grossi, G., Thornton, I.M., & Neville, H.J. (2003) Representation of change: separate electrophysiological markers of attention, awareness, and implicit processing. *Journal of Cognitive Neuroscience*. 15: 491-507.

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

Ferreira, F., & Henderson, J.M. (1999). *Good enough representations in visual cognition and language.* Paper presented at Architectures and Mechanisms of Language Processing Conference, Edinburgh, Scotland.

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

Filik, R., Sanford, A.J., & Sturt, P. (2005). Anaphoric reference to structured entities. Cited in, Poesio, M., Sturt, P., Artstein, R., and Filik, R. (2006) Underspecification and anaphora: Theoretical issues and preliminary evidence. *Discourse Processes* 42(2) pp 157-175.

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

Fodor, J. & Pylyshyn, Z. (1988). Connectionism and cognitive architecture: A critical analysis. *Cognition*, 28: 3-71.

Frazier, L. (1979). *On comprehending sentences: Syntactic parsing strategies*. Bloomington, In: Indiana University Linguistics Club.

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

Friederici, A.D., Pfeifer, E., & Hahne, A. (1993). Event-related potentials during natural speech processing: Effects of semantic, morphological and syntactic violations. *Cognitive Brain Research*, 1, 183–192.

Fries, P. (2005). A mechanism for cognitive dynamics: neuronal communication through neuronal coherence. *Trends in Cognitive Science* 9, (10), pp. 474–480.

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

Garrod, S.C., Freudenthal, D., & Boyle, E. (1994). The role of different types of anaphor in the on-line resolution of sentences in a discourse. *Journal of Memory and Language*, 33, p39-68.

Garrod, S.C. & Sanford, A.J. (1999). Incrementality in discourse understanding in: H.Van Oostendorp & S.R.Goldman (Eds.), *The construction of mental representations during reading*. Mahwah, NJ: Lawrence Erlbaum Associates .

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

Gibson, E. (2000). The dependency locality theory: A distance-based theory of linguistic complexity. In Miyashita, Y., Marantz, A., and O'Neil, W. (Eds.), Image*, language, brain* (pp. 95-126), Cambridge, MA: MIT Press.

Glenberg, A.M., Wilkinson, A.A., & Epstein, W. (1982). The Illusion of Knowing: Failure in the Assessment of Comprehension. *Memory and Cognition*, 10, 597-602.

Gray, C.M., König, P., Engel, A.K., & Singer, W. (1989). Oscillatory responses in cat visual cortex exhibit inter-columnar synchronization which reflects global stimulus properties. *Nature*, 338, 334-337.

Grice, H.P. (1975). Logic and Conversation, in P. Cole and J.L. Morgan, editors, *Synatx and Semantics*, Academic Press

Gross, D., Allen, J., & Traum, D. (1993). *The TRAINS 91 dialogues* (TRAINS Tech Note No. 92-1). Computer Science Department, University of Rochester, New York.

Gunter, T.C., Stowe, L., & Mulder, G. (1997). When syntax meets semantics. *Psychophysiology*, 34, 660–676.

Hagoort, P., Brown C, & Groothusen, J. (1993). The syntactic positive shift (SPS) as an ERP measure of syntactic processing. *Language and Cognitive Processes*, 8: 439–83.

Hagoort, P., Hald, L., Bastiaansen, M.C.M., & Petersson, K.M. (2004). Integration of Word Meaning and World Knowledge in Language Comprehension. *Science*, 304 (5669), 438-440.

Hahne, A., & Friederici, A.D. (1999). Electrophysiological evidence for two steps in syntactic analysis. Early automatic and late controlled processes. *Journal of Cognitive Neuroscience*. 11, pp. 194–205.

Hald, L., Bastiaansen, M., & Hagoort, P. (2006). EEG theta and gamma responses to semantic violations in online sentence processing: an oscillatory correlate of the N400 effect. *Brain and Language*, 96 (1), 90-105.

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

Hannon, B., & Daneman, M. (2001a). Susceptibility to semantic illusions: An individual-differences perspective. *Memory and Cognition*, 29, 449-461.

Hannon, B., & Daneman, M. (2001b).A new tool for measuring and understanding individual differences in the component processes of reading comprehension. *Journal of Educational Psychology*, 93, 103-128.

Hauk, O., & Pulvermüller, F. (2004). Effects of word length and frequency on the human event-related potential. *Clinical Neurophysiology*, *115*, 1090-1103.

Heinze, H., Muente, T., & Kutas, M. (1998). Context effects in a category verification task as assessed by event-related brain potential (ERP) measures. *Biological Psychology*, 47, 121–135.

Henderson, J.M., & Hollingworth, A. (1999). High-level scene perception. *Annual Review of Psychology*, 50, 243-271.

Hobbs, J.R. (1985). *Granularity*. Paper presented at the International Joint Conference on Artificial Intelligence, Los Angeles.

Hoeks, J.C.J., Stowe, L.A., & Doedens, L. H. (2004). Seeing words in context: the interaction of lexical and sentence level information during reading. *Cognitive Brain Research*, 19(1), 59-73.

Holcomb, P.J. (1988). Automatic and attentional processes: An eventrelated brain potential analysis of semantic priming. *Brain and Language*, 35, 66-85.

Holcomb, P.J., & Neville, H.J. (1990). Semantic priming in visual and auditory lexical decision: A between modality comparison. *Language and Cognitive Processes.* 5, 281-312.

Hollingworth, A., Schrock, G., & Henderson, J.M. (2001). Change detection in the flicker paradigm: The role of fixation posi- tion within the scene. *Memory and Cognition*.;29:296–304.

Hollingworth, A., Williams, C.C., & Henderson, J.M. (2001). To see and remember: Visually specific information is retained in mem- ory from previously attended objects in natural scenes. *Psychonomic Bulletin and Review*.;8:761–768.

Irwin, D.E. (1996). Integrating information across saccadic eye movements. *Current Directions in Psychological Science*, 5, 94-100.

Jaarsveld, H., Van Dijkstra, A., & Hermans, D. (1997). The detection of semantic illusions: Task-specific effects for similarity and position of distorted terms. *Psychological Research*, 59 (4), 219-230.

Just, M.A., & Carpenter, P.A. (1980). A theory of reading: from eye fixations to comprehension. *Psychological Review* 87:329–354.

Kaan, E., Harris, A., Gibson, E., & Holcomb P. (2000). The P600 as an index of syntactic integration difficulty, *Language and Cognitive Processes* 15, pp. 159–201.

Kahneman, D., Slovic, P., & Tversky, A. (1982). *Judgement under uncertainty: Heuristics and biases*. Cambridge, England: Cambridge University Press.

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

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

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

Kintsch, W., & T.A., van Dijk. (1978). Toward a model of text comprehension and production. *Psychological Review* 8, 363-94.

Koh, S., Sanford, A.J., Clifton, C., & Dawydiak, E. (in press) Good Enough Representation in Plural and singular pronominal reference: Modulating the conjunction cost. In, Gundel J., and Hedberg, N., (Eds.) *Interdisciplinary perspectives on referencing processing*. Oxford: OUP.

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

Kolk, H.H.J., Chwilla, D.J., Van Herten, M., & Oor, P.J.W. (2003). Structure and limited capacity in verbal working memory: a study with event-related potentials. *Brain and Language*, 85 (1), 1-36.

Kuperberg, G.R., Kreher, D.A., Sitnikova, T., Caplan, D., & Holcomb, P. (2007). The role of animacy and thematic relationships in processing active English sentences: Evidence from event-related potentials. *Brain and Language*; 100: 223-238.

Kuperberg, G.R., Sitnikova, T., Caplan D., & Holcomb, P.J. (2003). Electrophysiological distinctions in processing conceptual relationships within simple sentences. *Cognitive Brain Research*;217:117-29.

Kuperberg, G.R. (2007). Neural mechanisms of language comprehension: Challenges to syntax. *Brain Research (Special Issue);* 1146:23-49. Epub 2006 Dec 2.

Kutas, M., & Federmeier, K.D. (2000). Electrophysiology reveals semantic memory use in language comprehension. *Trends in Cognitive Science* 4:463-470.

Kutas, M. & Hillyard, S.A. (1980). Reading Senseless Sentences: Brain Potentials Reflect Semantic Incongruity. *Science*, 207: 203-205.

Kutas, M., & Hillyard, S.A. (1983). Event-related brain potentials to grammatical errors and semantic anomalies. *Memory and Cognition*; 11: 539–50.

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

Kutas, M., & Van Petten, C.K. (1994). Psycholinguistics electrified: event-related brain potential investigations. In: Gernsbacher MA, editor. *Handbook of psycholinguistics*. San Diego: Academic Press, p. 83–143.

Kutas, M., Lindamood, T.E., & Hillyard, S.A. (1984). Word expectancy and event-related brain potentials during sentence processing. In S. Kornblum and J. Requin (Eds.), *Preparatory States and Processes* (pp. 217-237). Hillsdale, NJ: Lawrence Erlbaum Associates.

Lucas, M. (2000). Semantic priming without association: a meta-analytic review. *Psychonomic Bulletin and Review*, Vol.7, no.4, 618-630.

Marslen-Wilson, W.D. (1973) Linguistic structure and speech shadowing at very short latencies. *Nature*, *244*, 522-523.

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

McClelland, J.L., St.John, M., & Taraban, R. (1989). Sentence comprehension: A parallel distributed processing approach. *Language and Cognitive Processes*, 4 (3/4), SI 287-335.

Miall, D., & Kuiken, D. (1994). Foregrounding, defamiliarization, and affect response to literary stories. *Poetics*, 22, 389-407.

Moxey, L.M., Sanford, A.J., Sturt, P., & Morrow, L. (2004). Constraints on the formation of plural reference objects: The influence of role, conjunction, and type of description. *Journal of Memory and Language*, 51:346-364.

Munte, T.F., Heinze, H.J., Matzke, M., Wieringa, B.M., & Johannes S. (1998). Brain potentials and syntactic violations revisited: no evidence for specificity of the syntactic positive shift, *Neuropsychologia* 36, pp. 217–226.

Neville, H., Nicol, J.L., Barss, A., Forster, K.I., & Garrett, M.F. (1991). Syntactically based sentences processing classes: Evidence from event-related brain potentials. *Journal of Cognitive Neuroscience*, 3, 151–165.

Ni, W., Fodor, J.D., 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(7), 1098-1111.

Nieuwland, M.S. & Van Berkum, J.J.A. (under review). The interplay between semantic and referential aspects of anaphoric noun phrase resolution: Evidence from ERPs.

Osterhout, L., & Mobley, L.A. (1995). Event-related brain potentials elicited by failure to agree. *Journal of Memory and Language*, 34,739–773.

Pickering, M.J., McElree, B., Frisson, S., Chen, L., & Traxler, M.J. (2006). Underspecification and aspectual coercion. *Discourse Processes*, 42, 131-155.

Pickering, M.J., & Traxler, M.J. (1998). Plausibility and recovery from garden paths: An eye-tracking study. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 24*, 940-961.

Pickering, M.J., Traxler, M.J, & McElree, B. (2005). The difficulty of coercion: A response to de Almeida. *Brain and Language*, 93, 1-9.

Pierrehumbert, J.B. (1980). *The Phonetics and Phonology of English Intonation*, PhD dissertation. MIT. Published by Garland Press, New York, 1990.

Piñango, M.M., Zurif, E., & Jackendoff, R. (1999). Real-time processing implications of enriched composition at the syntax-semantics interface. Journal of *Psycholinguistic Research*, 28, 395-414.

Poesio, M., Sturt, P., Artstein, R., & Filik, R. (2006). Underspecification and anaphora: Theoretical issues and preliminary evidence. *Discourse Processes* 42(2) pp 157-175.

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.

Reder, L.M., & Cleeremans, A. (1990). The role of partial matches in comprehension: The Moses illusion revisited. In A. Graesser and G. Bower, (Eds.), *The psychology of learning and motivation*, Vol. 25, New York: Academic Press, pp. 233-258.

Roskies, A.L. (1999). The binding problem. *Neuron*, 24, 7–8.

Rugg, M.D., & Doyle, M.C. (1994). Event-related potentials and stimulus repetition in indirect and direct tests of memory. In: H. Heinze, T. Munte & G.R. Mangun, Editors, *Cognitive electrophysiology*, Birkhauser, Boston, pp. 124–148.

Sanford, A. J., Filik, R., Emmott, C. & Morrow, L. (in press). They're digging up the road again: The processing cost of Institutional "They". *Quarterly Journal of Experimental Psychology*.

Sanford, A.J., & Garrod, S.C. (1981). *Understanding Written Language*: Explorations of Comprehension Beyond the Sentence. Chichester:John Wiley and Sons.

Sanford, A J & Garrod, S C (1989).What, when and how? Questions of immediacy in anaphoric reference resolution. *Language and Cognitive Processes*, 4, p235-262.

Sanford, A. J., Garrod, S.C., Lucas, A., & Henderson, R.J. (1983). Pronouns without explicit antecedents? *Journal of Semantics*, 2, p303-318.

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

Sanford, A.J.S., Sanford A.J., Molle, J., & Emmott, C. (2006). Shallow processing and attention capture in written and spoken discourse *Discourse Processes* 42 pp 109-130.

Sanford, A.J.S., Sanford, A.J., Filik, R., & Molle J. (2005). Depth of lexical-semantic processing and sentential load. *Journal of Memory and Language* 53(3) pp 378-396.

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

Sanford, A.J. (2002). Context, attention, and depth of processing during interpretation. *Mind and Language*, 17, 188-206.

Sanford, A.J., & Young, K. (2002) Unpublished data.

Sanford, A.J. & Garrod, S.C. (1998). The role of scenario mapping in text comprehension, *Discourse Processes*, 26, 159 - 190.

Sedivy, J.C., Tanenhaus, M.K., Chambers, C.G., & Carlson, G.N. (1999). Achieving incremental processing through contextual representation: Evidence from the processing of adjectives. *Cognition, 71,* 109-147.

Sergent C., Baillet S., & Dehaene S. (2005). Timing of the brain events underlying access to consciousness during the attentional blink. *Nature Neuroscience* Oct 8(10), 1391-1400, 2005.

Sereno, S. C., Posner, M. I. & Rayner, K. (1998). Establishing a time-line of word recognition: Evidence from eye movements and event-related potentials. *Neuroreport, 9*, 2195-2200.

Sereno, S. C. & Rayner, K. (2003). Measuring word recognition in reading:Eye movements and event related potentials. *Trends in Cognitive Sciences, 7*, 328-333.

Simons, D. (2000). Current Approaches to Change Blindness. *Visual Cognition*, 7, 1-15.

Simons, D.J., & Levin, D.T. (1997). Change blindness. *Trends in Cognitive Sciences*, 1(7), 261-267.

Singer, W. (1999). Neuronal synchrony : A versatile code for the definition of relations? *Neuron*, 24, 49--65.

Singer, W., & Gray, C.M. (1995). Visual feature integration and the temporal correlation hypothesis. *Annual Review of Neuroscience*, 18, 555-586.

Slovic, P. & Fischhoff, B. (1977). On the psychology of experimental surprises. *Journal of Experimental Psychology: Human Perception and Performance*, 3, pp 544-551.

Sturt, P. & Lombardo, V. (2005). Processing coordinate structures: Incrementality and connectedness. *Cognitive Science*, 29:291-305.

Sturt, P., Sanford, A.J., Stewart, A.J., & Dawydiak, E. (2004). Linguistic focus and good-enough representations: an application of the change-detection paradigm *Psychonomic Bulletin and Review* 11 pp 882-888.

Thornton, I.M. & Fernandez-Duque, D. (2000). An implicit measure of undetected change. *Spatial Vision* 14(1), 21-44

Thornton, I.M., & Fernandez-Duque, D. (2002). Converging evidence for the detection of change without awareness. *The Brain's Eyes: Neurobiological and Clinical Aspects of Oculomotor Research*, 99-118. (Eds.) Hyönä, J.; Munoz, D. P.; Heide, W.; et al. Elsevier Science B.V., Sara Burgerstraat 25, 1000 AE Amsterdam

Todorova, M., Straub, K., Badecker, W., & Frank, R. (2000). *Aspectual coercion and the online computation of sentential aspect*. Proceedings of the Twenty-second Annual Conference of the Cognitive Science Society, (pp.3-8).

Traxler, M.J., & Pickering, M.J. (1996). Plausibility and the processing of unbounded dependencies: An eye-tracking study. *Journal of Memory and Language*, 35, 454-475.

Traxler, M.J., McElree, B., Williams, R.S., & Pickering, M.J. (2005). Context effects in coercion: Evidence from eye-movements. *Journal of Memory and Language*, 53, 1-25.

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: 657–71

Van Berkum, J.J.A., Zwitserlood, P., Bastiaansen, M.C.M., Brown, C.M. & Hagoort, P. (2004). So who's "he" anyway? Differential ERP and ERSP effects of referential success, ambiguity and failure during spoken language comprehension. *Journal of Cognitive Neuroscience*, 16, 70. Suppl

Van Herten, M., Kolk, H.H.J., & Chwilla, D.J. (2005). An ERP study of P600 effects elicited by semantic anomalies. *Cognitive Brain Research*, 22 (2), 241-255.

Van Oostendorp, H. & De Mul, S. (1990). Moses Beats Adam: A Semantic Relatedness Effect on a Semantic Illusion. *Acta Psychologica*, 74, 35-46.

Van Oostendorp, H. & Kok, I. (1990). Failing to Notice Errors in Sentences. *Language and Cognitive Processes*, 5, 105-113.

Van Petten, C, Kutas, M, Kluender, R, Mitchiner, M, & McIsaac, H. (1991). Fractionating the word repetition effect with event-related potentials. *Journal of Cognitive Neuroscience*; 3: 131–50.

Van Petten, C., & Kutas, M. (1991). Influences of semantic and syntactic context on open- and closed-class words. *Memory and Cognition.*; 19: 95–112

Varela, F., Lachaux, J.P., Rodriguez, V., & Martinerie, J. (2001). The brainweb: phase synchronization and large-scale integration, *Nature Review: Neuroscience*. 2 229–239.

Vissers, C., Chwilla, D.J., & Kolk, H.H.J. (2007). The interplay of heuristics and parsing routines in sentence comprehension: Evidence from ERPs and reaction times. *Biological Psychology*, 75, (8-18).

Vogel, E.K., Luck, S.J., & Shapiro, K.L. (1998). Electrophysiological evidence for a postperceptual locus of suppression during the attentional blink. *Journal of Experimental Psychology: Human Perception and Performance*, 24, 1656-1674.

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

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

Weiss, S., & Müller, H.M. (2003). The contribution of EEG coherence to the investigation of language. *Brain and Language, 85*, 325–343.

Williams, P., & Simons, D.J. (2000). Detecting changes in novel 3D objects: Effects of change magnitude, spatiotemporal continuity, and stimulus familiarity. *Visual Cognition*, 7, 297-322.