On Repairing Sentences: An Experimental and Computational Analysis of Recovery from Unexpected Syntactic Disambiguation in Sentence Parsing

Submitted by Matthew James Green to the University of Exeter
as a thesis for the degree of
Doctor of Philosophy in Psychology
in July 2013

I certify that all material in this thesis which is not my own work has been identified and that no material has previously been submitted and approved for the award of a degree by this or any other University.

Signature:

Dedication

I would like to dedicate this thesis to my grandmother Eve McRae.

Acknowledgements

I would like to acknowledge those people who helped me with my PhD. Thank you Don Mitchell, my supervisor, for your guidance, your patience, and for having confidence in me when I lacked it myself, and for putting retirement to one side for the final push. Thanks to my family and friends who were there for me unfailingly, when it seemed like I would never finish: my mum Elspeth Green; Lily McGuire; Dan and Linny Morgan; Jerry Evans. Thanks to Marisa Boston for reinvigorating my efforts with your comments, and for helping from afar to make software work. Thanks to the NLG group at Aberdeen for useful comments and discussion.

Abstract

This thesis contends that the human parser has a repair mechanism. It is further contended that the human parser uses this mechanism to alter previously-built structure in the case of unexpected disambiguation of temporary syntactic ambiguity. This position stands in opposition to the claim that unexpected disambiguation of temporary syntactic ambiguity is accomplished by the usual first pass parsing routines, a claim that arises from the relatively extraordinary capabilities of computational parsers, capabilities which have recently been extended by hypothesis to be available to the human sentence processing mechanism. The thesis argues that, while these capabilities have been demonstrated in computational parsers, the human parser is best explained in the terms of a repair based framework, and that this argument is demonstrated by examining eye movement behaviour in reading. In support of the thesis, evidence is provided from a set of eyetracking studies of reading. It is argued that these studies show that eye movement behaviours at disambiguation include purposeful visual search for linguistically relevant material, and that the form and structure of these searches vary reliably according to the nature of the repairs that the sentences necessitate.

Contents

List of Tables

List of Figures

Abbreviations

ANOVA Analysis of variance

AP Adjectival phrase

BIC Bayesian information criterion

DSURP Dependency surprisal

DTIME Dependency retrieval time

ER Entropy reduction

FFD First fixation duration

FPRT First pass reading time

GDP Grammatical dependency principle

LMER Linear mixed effects regression

NP Noun phrase

NP/S Noun phrase / sentential complement ambiguity

NP/Z Noun phrase / no complement ambiguity

PCFG Probabilistic context-free grammar

PP Prepositional phrase

PREG Proportion of first-pass regressions out

RPD Regression path duration

S Sentence

SC Sentential complement

TDP Thematic processing domain

TSR Time spent regressing

TSURP Phrase structure surprisal

VP Verb phrase

Introduction

There follows an overview of the rest of the document.

The thesis is arranged into chapters as follows.

Chapter 1 covers reading in its visual, linguistic, and computational aspects. Eye movement metrics are explained here. Complement ambiguity is presented here. Chapter 2 explains how probabilistic approaches to reading are implemented. Chapter 3 presents a review of theories of sentence processing, focussing on disambiguation. Chapter 4 sets out the issues that the rest of the thesis focusses on in detail. The thesis uses mixed effects regression models for hypothesis testing. The relevant advantageous properties of these models are compared with those of traditional models for hypothesis testing (5). Chapters presenting simulation and eyetracking data from experiments are chapters 7, 8, 9, 10, 11, and 12. The first three experiments (in chapters 7, 8, and 9) are concerned with how the spatial arrangement of text affects measures of parsing. The next three experiments (in chapters 10, 11, and 12) are concerned with how regressive eye movements differ according to what kind of repair the partial representation of the sentence needs in order to accommodate new material. The final chapter (13) draws together the evidence across the experiments to evaluate the theories explored in the background material in the light of the evidence, and offers some guidance for future work.

Chapter 1

The reading process

This chapter focusses on the processes that are involved in reading. The chapter has three parts dealing respectively with reading as a visual process, a linguistic process, and a computational process.

In this thesis I will use the word *reading* to indicate skilled fluent reading that is undertaken with the intention of understanding the text. A *reader* then is someone who is engaged in gathering information from the text and does not resort to skimming or similar behaviour. The reading process involves coordination of the linguistic system and the visual system.

1.1 Reading as a visual process

The oculomotor muscles move the eyes in a series of jumps through the words of a sentence. The landing points between jumps are referred to as fixations, and the jumps themselves are called saccades. This contrasts with other oculomotor routines such as those involved in smooth pursuit where the eyes glide smoothly. The retina has a small area of high density of photoreceptors called the fovea, surrounded by a lower density in the periphery. By making saccades the oculomotor system is able to position the fovea to receive light reflected from the word that is currently being read.

When a word is fixated foveally it is represented with high resolution in the visual cortex. Information passes from visual cortex to the linguistic system that is charged with identifying the word from its physical representation in text and retrieving semantically associated properties. While the reading process is going smoothly the eyes tend to keep moving through the text. However there are times when the reading process does not go smoothly. Under these circumstances the eyes pause for longer in fixations, and sometimes move back into parts of the sentence that have already been fixated, before moving on to fixate new material. In this way there can be several passes over the text, and it is useful to distinguish first pass routines from those associated with second and subsequent passes.

Basic properties of eye movements in reading In order to make sensible inferences from eye movements to word integration events, it is important to consider fundamental information about how the eyes move in reading more generally, and how those movements vary. Eye movements in fluent reading are characterised by a series of fixations and saccades: an observation made by Emile Javal in the late nineteenth century if not before. While Javal relied on observing with the naked eye the eye movements made by another individual, subsequent technologies have allowed more detailed measurements that help delimit the scope of these fixations and saccades.

Fixations are typically 200 ms to 250 ms long ranging between 150 ms to 500 ms so that the eyes move about 4 or 5 times a second in reading Rayner, Pollatsek, Ashby, and Clifton Jr (2012). Some of the variance in fixation durations is due to linguistic properties of the text. There are constant small movements of the eye muscles that are made so that neurons in the retina can continue to fire (this avoids synaptic fatigue). These tiny movements are called tremor or nystagmus. They are such small movements that they do not get considered when it comes to reading. The eyes also drift slightly while reading, and small micro saccades are made to correct for drift. Again these are not typically considered in reading research. Typically, experimenters might pool micro saccades by adding their durations to nearby ‘proper’ fixations.

Saccades typically last for about 20-35 ms. The typical range of a saccade in fluent reading is about 7 to 9 characters. Saccades cannot be changed once they start, like other ballistic movements. The phenomenon of saccadic suppression (Matin, 1974) indicates that no information is taken in during a saccade. Campbell and Wurtz (1978) found that at best a smeared image of what is presented during a saccade can be extracted, but they had to prevent visual stimulation before and after the saccade to get this result. In normal circumstances, Wolverton and Zola (1983) demonstrated that replacing text with a mask during each saccade was not perceived by participants, and nor did it exert an influence on any measure of processing.

The amount of text that can be taken in at a fixation is sometimes called the perceptual span. Early research in this area using a tachistoscope (Marcel, 1974) indicated that it is about 3 to 4 words. Taylor (1965) used a simple number of words per fixation measure and offered 1.11 words as an estimate. This low estimate fails to take into account the suggestion that perceptual spans might overlap, under which circumstances the estimate is revealed to be conservative. Moving window techniques (McConkie & Rayner, 1975), in which a sentence is displayed for self-paced reading in such a way that the current word (the one in the window) is presented normally while the other words are replaced, typically with a series of dashes, or ‘x’ characters, indicated that allowing 31 letter spaces for the moving window, reaching 15 characters either side of the fixation, resulted in no loss of reading speed. This finding has appeared stable in subsequent work (DenBuurman, Boersma, & Gerrisen, 1981; Miellet, O’Donnell, & Sereno, 2009; Rayner & Bertera, 1979; Rayner, Castelhano, & Yang, 2009; Rayner, Inhoff, Morrison, Slowiaczek, & Bertera, 1981). Since the fovea extends 3 characters either side of a fixation, it is clear that much of the perceptual span is spread over the parafovea, and that – depending on the word lengths involved – the span covers some part of new words to the right of the word being fixated. In other words, the fovea is not the limit of what can be perceived in a fixation. Pollatsek, Raney, Lagasse, and Rayner (1993) showed that although the perceptual span may extend horizontally beyond the currently fixated word, it does not extend vertically below the currently fixated line. This changes in visual search where the span may extend beyond the currently fixated line.

In the event that a short word appears in the right parafovea, and that it is sufficiently frequent, like ‘the’, one might imagine that this word need not be fixated itself, and there is evidence that word skipping is a normal part of fluent reading. This skipping of words represents one departure from straightforwardly word-by-word eye movements. Frequency of skipping increases as words get more frequent, and increases for shorter words. Skipping frequency ranges between about 10% for little-known long words up to about 70% for some very frequent and very short words like articles (Brysbaert, Drieghe, & Vitu, 2003; Rayner & McConkie, 1976; Rayner, Slattery, Drieghe, & Liversedge, 2011).

Another departure from word-by-word eye movements in fluent reading occurs when words are revisited having once been read. This happens when regressive eye movements are launched from one word and visit words that were read earlier. There are broadly 3 types. One is a simple regression to the immediately previous word. These are common and quick. Another is the conscious deliberate and slow re-inspection of earlier material. Yet another is the very rapid regression that the reader is unaware of that is launched from a given word and takes in some earlier words before moving on so fast that the reader is still unaware of having made the movements. Frequency of regressing is estimated between 10% to 15% (Buswell, 1922; Rayner et al., 2012) with the highest estimate 15.3% coming from a large corpus of adults reading a novel (Vitu & McConkie, 2000). Even higher estimates of 30% have been made for subsets of readers (Radach & McConkie, 1998).

Immediacy and eye-mind assumptions Data from eye tracking is widely used in research on parsing (a review of 100 papers may be found in Clifton Jr, Staub, and Rayner (2007)). The logic of using eye movements to infer parser activity rests on two assumptions that were made explicit by Just and Carpenter (1980). These are: the *immediacy* assumption; and the *eye-mind* assumption. The immediacy assumption says that *a reader tries to interpret each content word of a text as it is encountered, even at the expense of making guesses that sometimes turn out to be wrong.* (Just & Carpenter, 1980, p. 330). The eye-mind assumption says that *the eye remains fixated on a word as long as the word is being processed ... [S]o the time it takes to process a newly fixated word is directly indicated by the gaze duration* (Just & Carpenter, 1980, p. 331). The balance of the evidence supports the view that eye movements during reading admit treatment under the immediacy and eye-mind assumptions – although see Vitu (1991) for an extreme view where low-level characteristics entirely determine eye movements during reading.

Definitions of eye-movement measures Rayner et al. (2012, p. 93) set out some eye tracking measures in common use. The designation *eye tracking temporal measures* is used in the thesis to cover the following measures: First Fixation Duration (FFD); First Pass Reading Time (FPRT); Regression Path Duration (RPD); Time Spent Regressing (TSR); Probability of Regression (PREG). These measures are defined here, along with their standard interpretations, taken from Rayner et al. (2012) .

First fixation duration (FFD) is the mean duration of the first fixation on a word regardless of other possible fixations on the word. It has traditionally been treated as a measure of early processing. First fixation duration is interpreted to index lexical access.

First pass reading time (FPRT) also known as gaze duration, is the sum of the durations of all fixations on the word that occur before leaving the word in any direction. This still captures the early processing (FFD is a subset of FPRT) but FPRT also includes any refixations that there might be on the word before a regression is launched from it. First pass reading time is often interpreted to index lexical integration into the phrase marker.

Regression path duration (RPD) includes FPRT but adds to it the durations of fixations on preceding words that the eyes regress to before leaving the word to the right to take in new material, as well as any refixations on the launch word that occur before new material is taken in. In this way RPD is sensitive to integration difficulties that yield regressive eye movements but is also confounded by early processing. Regression path duration is often interpreted to index incremental syntactic integration of the new word into the sentence's representation including any semantic problems that arise from this.

Time spent regressing (TSR) is a custom measure that I use in the thesis and is calculated by subtracting FPRT from RPD to give a measure that reflects the duration of regressive eye movement events including subsequent refixations on the launch word before new material is taken in. TSR is the only measure that is sensitive to variance in time spent on the regression path without also being confounded by early processing of the low level properties of the disambiguating word, since those properties are removed by subtracting FPRT. In the thesis this time spent regressing is valued, and analysed, as a measure of how long it took the parser to seek out a new solution, over and above how long it took the parser to realise initially that there was a problem with the phrase marker. It is contended in this thesis that time spent regressing is time spent seeking solutions.

Total reading time (TRT) is the sum of all fixations on the word regardless whether they were from the first or subsequent passes over the text. This measure is often used when comparing computational parser actions to reading times.

Proportion of regressions (PREG) A standard binomially distributed measure Proportion of regressions (PREG) was also computed. This measure is insensitive to temporal factors but gives an indication of whether any first-pass regressions were launched from the disambiguation. Proportion of regressions is often taken to index difficulty integrating the current word into the current existing partial representation of the sentence. Probability of (first-pass) regression is a measure that indicates whether a given trial resulted in a regression from the critical word on the first pass, often the disambiguating word. When aggregated over items nested in conditions, for example, it represents how likely the average participant was to make a regression in that condition. Because it is a binary valued measure before aggregation, it is well modelled by the binomial link function in the Linear Mixed Effects Regression (LMER) framework, but rather poorly modelled in the ANOVA framework because in the aggregate it is a proportion. Jaeger (2008) presents compelling arguments to the effect that analysing proportions with ANOVA is wrong and unnecessary. The measure is used by Rayner, Ashby, Pollatsek, and Reichle (2004) to show that the parser is influenced by lexical anomaly, and by Frazier and Rayner (1982) to show how parsing is influenced by unexpected disambiguation of syntactic ambiguity. In this thesis probability of regression is taken to index the likelihood that solutions need seeking. The term PREG is used in the thesis as a shorthand for probability of regression.

Two metrics that focus on the spatio-temporal distribution of regression paths saccades are given in chapter 4.3

1.2 Reading as a linguistic process

A sentence can be described as a sequence of states, where a state is a word and the sequence is the order of words in a sentence. The goal of sentence processing is to achieve a representation of the sequence that carries the intended meaning of the sentence. In order to achieve a representation of a sentence, people must have a way to combine the individual words that make up the sentence. This representation of the intended meaning of the sentence is built from the sentence's constituent parts. A grammar is assumed to be available that specifies rules for the combination of constituents.

Phrase structural grammar One proposal for how people combine words into sentences appeals to the notion of *constituency* to provide units of meaning at a level intermediate between the word and the sentence. When a group of words behaves as an intermediate-level unit in a sentence, the unit is referred to as a constituent of the sentence.

One type of constituent is the noun phrase. The sentence *The tall man loves Phoebe* contains two examples of a noun phrase: one is the bare noun *Phoebe* and the other is *the tall man*, where a person is individuated with reference to some property that he exhibits, and the sequence of words means something that is not conveyed by any of its elements taken on its own.

The observation that constituents of the same type often appear in similar syntactic contexts is taken to support the idea of constituency. For example the words preceding a verb often constitute a noun phrase.

Evidence that constituents behave as units can be seen in the observation that prepositional phrases can be moved without affecting the grammaticality of a sentence. For example *In January the weather is cold.* can be re-phrased as *The weather is cold in January.* without affecting the meaning of the sentence.

Evidence that constituents have properties that the individual words that make them up do not have can be seen from the ungrammaticality of moving one of the words that comprises the prepositional phrase, relative to the grammaticality of moving the entire phrase. For example, these sentences are both grammatical: *At 5 o'clock they have a meal.* and *They have a meal at 5 o'clock.*; but the following is not grammatical: *At 5 they have o'clock a meal.* Radford (1988) gives more examples of groups of words behaving as a single constituent.

The *context-free grammar* (also known as *phrase structure grammar*) provides a way to describe how constituents are related to each other in a structure. This grammar formalism assumes constituency as a primitive. Essentially the claim is that parses can be represented by hierarchical trees that connect words into nodes, where a node is a phrasal unit or constituent. The idea can be traced back to Wundt (1900) but was formalised by Chomsky (1956).

In phrase structure grammar each constituent has a head. While noun phrases have nouns at their heads, verb phrases have verbs at their head. Nouns have *feature structure* that determines how the elements of the noun phrase are governed by their head nouns. Verbs have *argument structure* that determines how the elements of the verb phrase relate to the head verb. Verb phrases consist of a verb and some other constructions (e.g., NP; PP; NP+PP; NP+SC, described in Table 1.1). These other constructions are referred to as *complements* of the verb.

Within a verb phrase the relations between the verb and each constituent can be described. Different verbs admit different complements. The possible set of

Table 1.1: Some common grammatical constituents

Label	Description	Example
NP	Noun phrase	*The tall man* bought a book.
VP	Verb phrase	Luke *loves Phoebe*.
PP	Prepositional phrase	The weather is cold *in January*.
SC	Sentential complement	The man believed *the government was lying*.

Table 1.2: Some common thematic roles[a]

Role	Definition
agent	The volitional causer of an event
experiencer	The experiencer of an event
force	The non-volitional causer of the event
theme	The participant most directly affected by the event
result	the end product of an event
content	The proposition or content of a propositional event
instrument	An instrument used in the event
beneficiary	The beneficiary of an event
source	The origin of the object of a transfer event
goal	The destination of an object of a transfer event

[a] from Jurafsky and Martin (2009, p. 655)

complements that a particular verb admits is referred to as that verb's *subcategorisation frame*. The verb *find* subcategorises for a NP: *find the ball*; the verb *want* subcategorises for either an N: *want the food*; or for a non-finite VP: *want the food to be nice*.

Subcategorisation frames allow description of the syntactic relations that a verb may participate in, whereas descriptions of the semantic relations that a verb may participate in are termed *thematic roles* (Fillmore, 1968; Gruber, 1965). For example the *agent* of a verb is the volitional causer of an event expressed in the verb; where the *force* of a verb is the non-volitional cause of the event (see Table 1.2). The set of thematic role arguments that a verb can take is sometimes called its *thematic grid*; *θ-grid*; or *case frame*.

Dependency grammar *Dependency grammar* is an example of a grammar formalism that does not assume constituency: there are no phrasal nodes in a dependency grammar. Instead dependency grammar describes sentences as words; and binary semantic or syntactic relations between pairs of words. Work on dependency grammars stems from the work of Tesnière (1959) who described sentences as sequences of "word-to-word connections". Typed dependency parsers label these relations, e.g., *subject*, *direct object*, but the simplest dependency parsers merely establish that there is a relation between pairs of words, one element of the pair being the *head* and the other element being the *dependent*.

Ambiguity In a sentence with no structural ambiguity from start to finish, as each new word is encountered, there is only one way for it to be integrated into the partial structure that exists at the time the new word is processed. This lack of ambiguity is unusual. Recall the idea of a sentence as a sequence of states. The sentence *John loves Mary.* has four transitions across states: (1) the appearance of *John*; (2) the appearance of *loves*; (3) the appearance of *Mary*; and (4) the appearance of a punctuation marker indicating the completion of the sentence, i.e., the full stop. At each of these four transitions the grammar indicates only one possible integration with the partial structure that has been built at the last transition.

The more common pattern as a parser effects state transitions is for the grammar to indicate several possibilities for integration at some state. To make this more concrete it is necessary to introduce an example of a sentence and a grammar where for some transition it is not clear what the correct integration is as the parser transitions from one state to the next. Example 1.1 provides such a sentence.

(1.1) The horse raced past the barn.

It is also necessary to provide a more complex grammar, given in Table 1.3. This grammar introduces determiners which were not present in the simple example; and two ways of composing a verb phrase where the simple grammar only offered one way; as well as offering two rules that are satisfied by the same input word *raced*.

Consider the state transitions in this sentence. There are 7 transitions but the focus is on transition 3: (1) start → *The*; (2) *The* → *horse*; (3) *horse* → *raced*; (4) *raced* → *past* (5) *past* → *the* (6) *the* → *barn*; (7) *barn* → *full stop* . The initial assumption that what follows is a sentence permits the first high-level partition

Table 1.3: A grammar that contains ambiguity

rule 1:	S	→	NP + VP
rule 2:	NP	→	DET + N
rule 3:	VP	→	V [active] + NP
rule 4:	VP	→	V
rule 5:	DET	→	the, The
rule 6:	N	→	barn
rule 7:	N	→	horse
rule 8:	V [active]	→	raced
rule 9:	V [passive]	→	raced
rule 10:	NP	→	P + NP
rule 11:	P	→	past
rule 12:	PP	→	P + NP

into NP and VP with reference to rule 1. The first state transition from the start to *The* involves attaching *The* as a determiner with reference to rule 5, and as part of a noun phrase by reference to rule 2. The next transition from *The* to *horse* is straightforwardly handled by rule 7 identifying *horse* as a noun, and rule 2 integrating that noun into the existing partially formed NP.

At this point there is no ambiguity. Ambiguity is introduced by transition (3) from *horse* to *raced*. Ambiguity arises because there are two rules that could integrate *raced*: rules 8 and 9 both have *raced* as the right hand side of the rule, but each rule maps the word to a different syntactic unit: active verb (rule 8) and passive participle (rule 9).

The general form of the question which rule applies to *raced* is the subject of some controversy. The general form of the question can be considered to be this: when a given constituent can be integrated into the existing partial phrase marker in more than one way, how should the transition be handled by the parser? The next two sections each offer a different way of approaching the integration of a word like *raced* that satisfies more than one rule from the grammar into the existing partial phrase marker. The two approaches differ according to whether the parser requires a single integration to be made, or whether the parser permits multiple integrations to co-exist.

Complement ambiguity Sentential complement ambiguities exploit the properties of 'complement' verbs like *believe* that can be followed either by a complement clause or by a direct object, or by no complement. When such verbs are followed by complements and an overt complementiser like *that* is used, no temporary syntactic ambiguity is present: however, when the complementiser is omitted, which may be done without violating the grammar, temporary syntactic ambiguity arises with respect to the first few words of the complement. These words may be taken as a direct object instead, and then when the complement verb appears, disambiguation ensues as the words that were taken to be part of a direct object of the verb are revealed necessarily to be part of a complement. Another possibility afforded by the multiple subcategorisation frame of words like *believe* is that the words immediately following could properly be the start of a main clause where the clause containing *believe* is properly a subordinate clause. Such cases are sometimes referred to as reduced complements. In these cases only the presence of a main verb resolves the temporary syntactic ambiguity, and when it appears, some major restructuring is involved. Complement ambiguities of both kinds have been used to investigate the parsing of ambiguous clauses (Clifton Jr, 1993; Ferreira & Henderson, 1991b; Holmes, Kennedy, & Murray, 1987; Pickering & Traxler, 1998; Rayner & Frazier, 1987; Sturt, Pickering, & Crocker, 1999; Trueswell, Tanenhaus, & Kello, 1993).

Sentences 1.2 and 1.3 provide an example of the sentential complement ambiguity that is used in this thesis.

(1.2) John knows the truth. (*direct object analysis*)

(1.3) John knows the truth hurts. (*sentential complement analysis*)

This ambiguity type has been used to provide evidence for claims about how reanalysis is implemented (Ferreira & Henderson, 1991b, 1998; Frazier & Rayner, 1982; Rayner, 1998; Warner & Glass, 1987), questions about whether ultimately incorrect analyses are retained or deleted (Christianson, Hollingworth, Halliwell, & Ferreira, 2001; Staub, 2007b), and the nature of the use of verb-subcategorisation preferences in parsing (Adams, Clifton Jr, & Mitchell, 1998; Mitchell, 1987; Staub, 2007a; Traxler, 2005; van Gompel, Pickering, & Traxler, 2001).

Below three cases of the ambiguity are considered.

Case 1: Consider the following example sentences. Slashes mark analysis region boundaries.

(1.4) The maid disclosed / the safe's location within the house / to the / officer.
(No complementiser, noun phrase complement)

(1.5) The maid disclosed that / the safe's location within the house / had been / changed. *(Overt complementiser, sentential complement)*

(1.6) The maid disclosed / the safe's location within the house / had been / changed. *(No complementiser, reduced sentential complement)*

Holmes et al. (1987) used sentences like Examples 1.4 – 1.6 in a self-paced reading paradigm, presenting cumulatively one word at a time so that the final display was of the whole sentence. They found that the reduced complement sentences like 1.6 were not harder to read than the overt complement sentences like 1.5, which finding led them to propose that the supposed Minimal Attachment effect was really an effect of the greater clause complexity in 1.6 and 1.5 versus 1.4. On this interpretation, the two sentential complement conditions do not differ in difficulty, and are both harder than the direct object condition – the direct object condition has only one set of clausal relations, but the complementiser conditions each have two sets (the second is introduced by the verb *had been changed*) and therefore the difficulty should be attributed to differential complexity of clausal relations, and not to any Minimal Attachment-related garden-pathing caused by complementiser omission.

Case 2: Consider the following sentences.

(1.7) The contestant imagined that the small tropical islands would be completely deserted. *(Overt complementiser, sentential complement)*

(1.8) The contestant imagined the small tropical islands would be completely deserted. *(No complementiser, sentential complement)*

(1.9) The contestant imagined the small tropical islands to be completely deserted. *(Infinitival clausal complement)*

(1.10) The contestant imagined the small tropical islands in the middle of the Pacific. *(Direct object noun phrase complement)*

Rayner and Frazier (1987) used sentences like Examples 1.7 – 1.10 to show, using a measure they called *mean first-pass reading time per character*, that participants experienced more difficulty with unmarked complements than marked complements, consistent with the predictions of Minimal Attachment. They showed using regression probability that the unmarked complement condition induced more regression probability than the marked complement condition, also consistent with the predictions of Minimal Attachment. Rayner and Frazier (1987) interpreted the difference between their findings and the findings of Holmes et al. (1987) in several ways. Firstly they argued that the effects of Minimal Attachment are expected to exert themselves in initial syntactic processing, and that therefore self-paced reading might obscure these effects by adding time due to different, later, processing to the measure for the disambiguation. Secondly, they considered specifically the cumulative self-paced reading paradigm used by Holmes et al. (1987). This method allows readers to look back into the sentence, but, in the absence of concurrent eye movement recording, it is not possible to tell for a given region how much of the time was spent looking at the region versus looking back into earlier material, since both patterns contribute to the same button-press latency. Third they noted that the cumulative nature of the paradigm has been shown to induce readers to decouple button-pressing from linguistic processing. Just, Carpenter, and Woolley (1982) showed that subjects press rapidly to get the whole sentence up on screen first, and then read it. Fourth they noted that the imposition of a secondary task in self-paced reading slows reading down versus eye-tracking, and that because the syntactic effects predicted by Minimal Attachment exert their influence early in the reading process, slowing down the process with a secondary task will obscure the influence of the effects of interest.

Case 3: From cases 1 and 2 it is possible to conclude that complementiser omission causes readers difficulty by forcing them to reanalyse, and that it is better to examine this difficulty using eye-tracking than by using self-paced reading. In order to examine the difficulty caused by reanalysis at the disambiguation of these complement ambiguities, it is useful to examine a case where the difficulty of reanalysis is modulated by some other factor.

In cases 1 and 2, each verb was obliged to take a complement of some kind. Some verbs, like *remember* have several legitimate options with respect to complements: they may take no complement (e.g., 1.11 below); they may take a noun phrase direct object as a complement (e.g., 1.12 below); or they may take a sentential complement (e.g., 1.13 below).

(1.11) The man remembered. *(Verb takes no complement)*

(1.12) The man remembered the anniversary. *(Verb takes a noun phrase complement)*

(1.13) The man remembered the anniversary was a disaster. *(Verb takes a sentential complement)*

Other verbs license only a subset of these possibilities. For example, verbs like *salute* take only a noun phrase complement or no complement at all, which makes temporarily ambiguous sentences like 1.14 possible; and verbs like *notice* take sentential complements or noun phrase complements, but must take some complement, which makes temporarily ambiguous sentences like 1.15 possible.

Following Sturt et al. (1999) I will use the term *noun phrase / no complement (NP/Z)* ambiguity to indicate a verb that may take a noun phrase as a complement or, also legitimately, may take no complement at all, where Z stands for 'zero'. I will use the term *noun phrase / sentential complement (NP/S)* ambiguity to indicate a verb that must take some complement, either a noun phrase complement NP or, also legitimately, a sentential complement S.

(1.14) After the cadet saluted the captain walked to the gates of the enclosure. *NP/Z, ambiguous*

(1.15) The cadet noticed the captain walked to the gates of the enclosure. *NP/S ambiguous*

Ignoring for a moment that 1.14 starts with *After*, the other words in the sentences can be held constant except for the first verb. In this way we can manipulate the type of complement ambiguity independently from the rest of the words in the sentences.

Pritchett (1988, 1992) recorded his non-empirical intuition that NP/Z structures like 1.14 caused him more reanalysis difficulty than NP/S structures like 1.15. Empirical support for this claim was obtained by Sturt et al. (1999). Theoretical attention was directed at the distinction by Fodor and Inoue (1998), who offered an explanation of the differential difficulty in their Diagnosis model of reanalysis. In 1.15 the ultimately correct interpretation of the sentence requires the sentential complement relation. Therefore if the parser assigns the more frequent direct object relation, some revision of that assignment will have to take place in order to arrive at the correct interpretation of the sentence.

Table 1.4: A simple grammar

rule 1:	S	→	NP + VP
rule 2:	NP	→	N
rule 3:	VP	→	VP + NP
rule 4:	VP	→	V

Table 1.5: A simple lexicon

N	→	{John, Mary}
V	→	loves

1.3 Reading as a computational process

Parsing as search A parser can be seen as implementing a search through the space of possible parses in order to find the appropriate parse for the sentence it is parsing. The goal of parsing is then to identify all the trees whose root is in the top level S node, and which contain exactly the words of the sentence. There are two kinds of constraint on the search: constraints imposed by the requirement to include all of the string of words in the sentence; and constraints imposed by the grammar – the parse must have one root and that root must be the start symbol S. For purposes of illustration of the basic machinery of parsing, an example of sentence processing in the simplest case follows, with reference to a sentence given in Example 1.16; a grammar given in Table 1.4, a lexicon given in Table 1.5, and a phrase marker in Figure 1.1. *Top-down* parsers start from the root node S at the top of these inverted trees, and work their way down the inverted tree to the words at the leaves of the tree. *Bottom-up* parsers start with the words and work their way up the tree until they reach the root node S. Some parsing algorithms combine these strategies, such as the *left-corner* parser.

(1.16) John loves Mary.

Top-down parsing The top-down parser begins with the assumption that the sentence has a root node S. The next move is to find the tops of all trees that begin with S: these are yielded by the grammar, in rules that have S as their left hand side. In the simple grammar in Table 1.4 there is only one such rule, rule

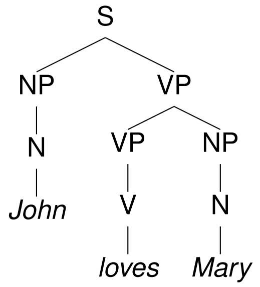


Figure 1.1: A simple phrase marker for *John loves Mary*.

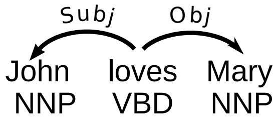


Figure 1.2: Dependency parse of *John loves Mary*.

1. Thus the next level of the search space has one partial tree. The constituents NP and VP of this rule are expanded next. The constituents are looked up in the grammar rules. NP has one expansion, from rule 2. The right hand side of this rule provides an expectation for a N node. Since no rule in the grammar expands N, it is sought for in the lexicon. The lexicon informs the parser that N expands to either *John* or *Mary*. *John* is found in the input string at position 1, so the parser has worked its way all the way down one branch of the tree at this point.

The next step is to go back to the expectation for a VP that was set up by grammar rule 1. VP is found on the left hand side of rules 3 and 4. Two expansions of the tree are made, one to account for the right hand side of rule 3 (VP → VP + NP) and one to account for the right hand side of rule 4 (V). Taking rule 3 first, VP expands to V, for which the lexicon contains *loves*, which matches a word in the input, and NP expands to N which is satisfied by *Mary*. Pursuing rule 4 yields V and *loves*, but cannot account for *Mary*, so the partial structure representing rule 4 is removed, leaving only the successful parse indicated in Figure 1.1.

Bottom-up parsing The bottom-up parser starts by looking up each input word in the lexicon and building three partial trees that contain the part of speech for each word that was yielded by the lexicon. Each of these trees is then expanded. Rule 2 is applied to each of the nouns, enabling the parser to work up to NP from each of the nouns. The verb enables the parser to build up to VP by applying rule 4. The parser expands one level to the next by looking for places in the partial trees where the right hand side of some grammar rule might fit. In the example, NP *John* and VP *loves Mary* constitute the right hand side of a rule that has the goal node S on its left hand side and the parser succeeds.

Both top down and bottom up parsing algorithms have advantages and disadvantages. The top-down strategy is efficient in the sense that it does not consider any partial trees that do not lead to S, because it starts out from S. In contrast, a strict bottom-up strategy proposes potentially very many ways to combine terminals that can never lead to S. The top-down strategy does involve some redundancy though. This is because it can explore many trees that start with S but do not go on to match the words of the input. The bottom-up strategy prevents this kind of inefficiency by only proposing partial trees that do match the input.

There is a parsing strategy that combines top-down and bottom-up parsing in a way which prevents the worst inefficiencies of both these strategies when applied strictly: this combined strategy is known as *left-corner parsing*. The left corner of a rule is the left-most part of the right hand side of the rule. In the simple grammar in Table 1.4 the left-corner of rule 1 is NP, and the left-corner of rule 3 is VP. The left-corner strategy alternates steps of top down and bottom-up parsing.

Left corner parsing Assuming that the parser has identified that John matches a grammar rule with NP at its left hand side, the left-corner parser then seeks NP at the left-corner of a grammar rule. In the simple grammar this is rule 1. In order to be able to make use of this rule, the parser must be able to find VP as the next item in the input string. This sets up a top-down expectation for VP. A top-down strategy can then be used to match VP to words in the input string. If they are not found, the parser can reject the rule at this stage, but if words can be found to match VP, the parser can make use of the rule identified by the left-corner strategy. The left-corner parser begins by stating the assumption that what follows is a sentence, which gives the initial highest level decomposition into noun phrase and verb phrase, in the same way as a pure top-down strategy. Next it switches to bottom-up mode, taking in the input word John, and identifying it as a N with reference to the lexicon. This is a bottom-up step made with reference to the lexicon. Then a grammar rule is sought that has N as its left corner: rule 2, NP $\rightarrow$ N. This is a bottom-up step made with reference to the grammar. Next it seeks a rule that has NP as its left corner: rule 1, S $\rightarrow$ NP $+$ VP. This is a bottom-up rule made with reference to the grammar. Because this step has the goal node S as its left hand side the parser has succeeded for that branch of the tree. Also an expectation has been created top-down for a VP because of the presence of VP in the rule that identified an S. The next step is to take in input: loves. Its category V is recognised bottom-up with reference to the lexicon. The left corner rule supplies the expansion to VP by rule 4 in the grammar, applied bottom-up. This represents a match with the top-down expectation for a VP and yields an accurate tree for the whole sentence.

Dependency parsing So far the focus has mainly been on phrase-structural grammar. An alternative grammar formalism is *dependency grammar* (Tesnière, 1959). In a dependency grammar there are no phrasal nodes in contrast with a phrase structure grammar. However dependency grammar parses still represent the structure of the sentence. Figure 1.2 shows the dependency parse for *John loves Mary*. It is possible to convert between phrase-structural representations and dependency representations of a sentence's structure. Conversion from a phrase structure parse to a dependency tree is done by making the head of each non-head child of a node depend on the head of the head child. This conversion is illustrated below in: Figure 1.3, which represents a phrase structural parse of the sentence in Example 1.17; and Figure 1.4, which shows the resulting dependency parse after conversion. Although a dependency parse can be written out vertically, as in Fig 1.4, it is more common to see dependency parses written out horizontally, as in Fig 1.5 with curved arrows indicating the head and dependent of each dependency relation. Notice that there are no phrasal nodes.

(1.17) John liked the dog in the pen.

The equivalent of a phrase structure grammar in dependency parsing is the *transition system*, described below in section 2.3.

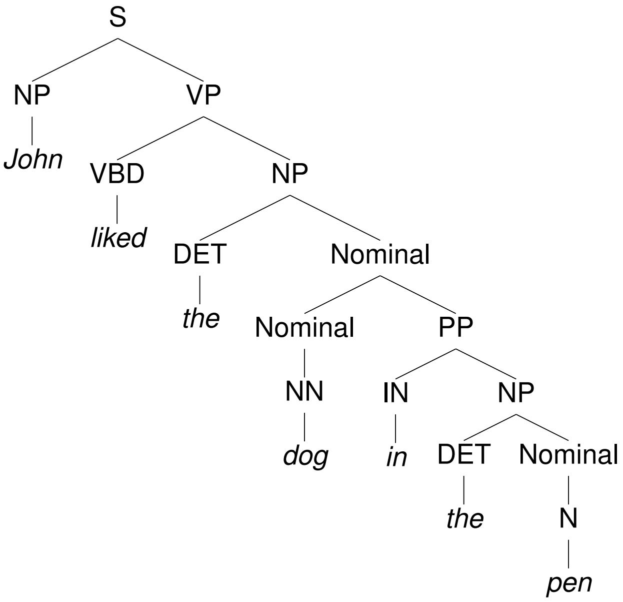


Figure 1.3: Phrase structural representation of *John liked the dog in the pen.*

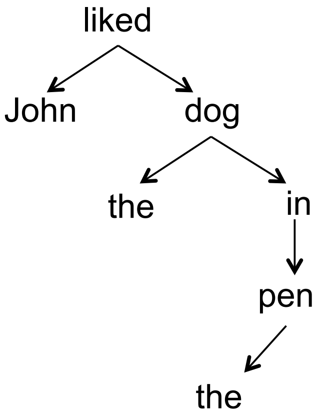


Figure 1.4: Dependency representation of *John liked the dog in the pen. Arrows point from heads to dependents.*

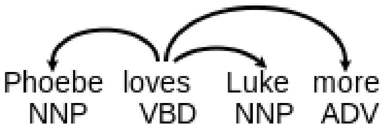


Figure 1.5: Dependency parse of the sentence *Phoebe loves Luke more*, showing the dependency relations. Arrows point from heads to dependents.

Chapter 2

Reading as a probabilistic process

This chapter focusses on ambiguity of sentence structure, and how parsers proceed when faced with ambiguous structure. The approaches can be divided into single path and parallel types.

2.1 PCFGs

A probabilistic context-free grammar (PCFG; T. L. Booth (1969); Manning and Schütze (1999)) is a collection of context-free grammatical rewrite rules of the form $X \rightarrow \alpha$ which specify how a constituent may decompose into more constituents or a terminal. For each constituent that decomposes, the probabilities of its decompositions must sum to one. Table 2.1 illustrates a small probabilistic context-free grammar (PCFG) with weights on the rules, and which generates phrase markers for main clause and direct object analyses of the partial sentence *When the dog scratched the vet and his new assistant.*

(2.1) When the dog scratched the vet and his new assistant removed the muzzle.

2.2 A top-down probabilistic phrase-structure parser

A top-down probabilistic parser TDPARSE is provided by Roark (2013) and described in Roark (2001) and Roark (2004). The parser is set up to generate mea-

Table 2.1: A small probabilistic context-free grammar (PCFG) that generates example 2.1[a]

		Rule	Prob			Rule	Prob
S	→	SBAR S	0.3	N	→	dog	0.2
S	→	NP VP	0.7	N	→	vet	0.2
SBAR	→	COMPL S	0.5	N	→	assistant	0.2
SBAR	→	COMPL S COMMA	0.5	N	→	muzzle	0.2
COMPL	→	When	1.0	N	→	owner	0.2
NP	→	DET N	0.33	Adj	→	new	1
NP	→	DET Adj N	0.33	VP	→	V NP	0.5
NP	→	DET Conj NP	0.33	VP	→	V	0.5
Conj	→	and	1.0	V	→	scratched	0.25
Det	→	the	0.8	V	→	removed	0.25
Det	→	its	0.1	V	→	arrived	0.5
Det	→	his	0.1	COMMA	→	comma	1

[a] When the dog scratched the vet and his new assistant removed the muzzle

sures of psycholinguistic interest (Roark, Bachrach, Cardenas, & Pallier, 2009). As presented in Roark (2004), the probabilities in the PCFG used in the parser are smoothed so that the parser is guaranteed not to fail due to garden-pathing, despite following a beam search strategy. Hence there is always a non-zero prefix probability as defined in Equation 2.7. The parser was trained on the Wall Street Journal part of the Penn Treebank (Charniak, 2000).

The parser follows a top-down leftmost derivation strategy. The parser maintains a set of possible connected derivations, weighted via the PCFG. It uses a beam search, whereby the highest scoring derivations are worked on first, and derivations that fall outside of the beam are discarded. The model conditions the probability of each production on features extracted from the partial tree, including non-local node labels such as parents, grandparents and siblings from the left-context. The final step in parsing, following the last word in the string, is to *complete* all non-terminals in the yield of the tree. The parser is a k-best parallel parser whose beam width varies as described below.

The parser takes as input a string of n words w_0^n; a PCFG G; and a queue of candidate analyses. A candidate analysis $C = (D, S, P_D, F, w_1^n)$ where D is a derivation, S is a stack, P_D is a derivation probability, and w_1^n is the string of remaining words in the sentence. The first word in the string remaining to be parsed, w_1^n, is called the *lookahead word*. The derivation D is a sequence of rules used from G. The stack S contains a sequence of nonterminal symbols and an end-of-stack marker $\$$. The symbol $\langle/s\rangle$ denotes the end-of-sentence terminal. The probability P_D is the product of the probabilities of all rules used in D. F is the product of P_D and a look-ahead probability $PAD(S, w_i)$ which measures the likelihood of the stack S rewriting with w_i as its left corner.

The parse begins with a single candidate analysis on the queue. Next the top-ranked candidate is popped from the queue. If $S = \$$ and $w_i = \langle/s\rangle$ then the analysis is complete. If not, all C' such that C derives C', denoted $C \Rightarrow C'$ are pushed onto the queue.

This is implemented as beam search. For each word position i there is a separate queue H_i with look-ahead word w_i. When there are 'enough' analyses on H_{i+1} all candidate analyses on H_1 are discarded. All parses pushed onto H_{i+1} are complete. The parse on H_1 with the highest probability is returned for evaluation. In the event that no complete parse is found, a partial parse is returned and evaluated.

The beam threshold at word w_i is a function of the probability of the top-ranked candidate analysis on H_{i+1} and the number of candidates on H_{i+1}. Essentially, the beam width should be wide when there are few analyses on H_{i+1}, but relatively narrow when there are many analyses on H_{i+1}. If p is the probability of the top-ranked parse on H_{i+1} then an analysis is discarded if its probability falls below $pf(\gamma, |H_{i+1}|)$, where γ is an initial parameter calleed the *base beam factor* set to 10^{-11} and $f(\gamma, |H_{i+1}|)$ is $\gamma|H_{i+1}|)^3$. So, if 100 analyses have already been pushed onto H_{i+1} then a candidate analysis must have a probability greater than $10^{-5}p$ to avoid being pruned. After $1,000$ candidates, the beam has narrowed to $10^{-2}p$. There is also a maximum number of allowed analyses on H_1 in case the parse does not put an analysis for H_{i+1} and this maximum was $10,000$.

2.3 Nivre's non-projective transition system parser

Nivre (Nivre, 2004b, 2006) defines each parser state as consisting of a tuple:

$$\text{state} = (\Sigma, B, A) \qquad (2.2)$$

Table 2.2: Transitions for the Nivre non-projective transition system[a]

transition	definition			condition
Left–arc	$([\sigma\|i,j],B,T,A)$	$\Rightarrow$	$([\sigma\|j],B,T,A\cup\{(j,i)\})$	$i\neq 0$
Right–arc	$([\sigma\|i,j],B,T,A)$	$\Rightarrow$	$([\sigma\|i],B,T,A\cup\{(j,i)\})$	
Shift$_\beta$	$(\sigma,[i\|\beta],T,A)$	$\Rightarrow$	$([\sigma\|i],\beta,T,A)$	$\beta\neq 0$
Shift$_r$	$(\sigma,B,[i\|\tau],A)$	$\Rightarrow$	$([\sigma\|i],B,\tau,A)$	$\tau\neq 0$
Swap	$([\sigma\|i,j],\beta,T,A)$	$\Rightarrow$	$([\sigma\|j],[i\|\beta],T,A)$	$0<i<j$

[a] Table taken from Boston (2012, p.50)

Σ is a stack consisting of already-parsed words that still require heads or dependents. B is a buffer of upcoming words, and A holds the dependency analysis information. This transition system handles non-projective analyses by allowing already-parsed words to be pushed back onto the buffer B so that the sentence can be reordered and attachments can be made locally. One problem with this transition system is that buffered, already-parsed and unparsed words are held in one data structure, B, which could lead to problems for incrementality. For example, the parser should be blind to any words that haven't been parsed yet, but already-parsed words should be available.

To meet this requirement a data structure is added to Nivre's tuple, T, as in equation (2.3). Following Nivre (2004a), T holds all words that have not yet been parsed. B only holds words that have been taken off the stack for reordering.

$$\text{state} = (\Sigma, B, T, A) \tag{2.3}$$

To transition from one state to the next, Nivre defines four actions: Left–arc; Right–arc; Shift; and Swap. Aside from the distinction between Shift$_\beta$ and Shift$_r$, the definitions in Table 2.2 are directly from Nivre (2009, p. 353). Following his conventions, Σ's top appears on the right and B and T's tops appear on the left.

Dependencies can be formed only between the two top elements in the stack, $\sigma 1$ and $\sigma 2$. For left-arc transitions $\sigma 1$ (or j) becomes the head of $\sigma 2$, i, and i is removed from the stack. For right-arc transitions, the opposite occurs: $\sigma 1$ (or j) becomes the dependent of $\sigma 2$, i, and j is removed from the stack. In the implementation, shift is one action: until β is empty, shift pops elements off β and onto σ. Once β is empty shift pops elements off τ. In this way the extra data structure does not make any changes to the way the parser itself works (Boston, 2012).

Finally, the Swap action is what gives this parser its ability to handle non-projective structures. Essentially, Swap reorders elements so that two discontinuous elements can be side-by-side in the stack as $\sigma 1$ and $\sigma 2$. Swap pops $\sigma 2$, in this case i, off of σ and pushes it onto β. This allows a new word, originally $\sigma 3$, to be available for dependencies with $\sigma 1$. Additionally, this reorders the sentence: if $\sigma 2$ is pushed back onto the stack, the two words are inverted.

The parser know which actions to take to generate the correct analysis by reference to an *oracle*. This section shows how the *oracle* is created in steps: (1) convert treebank to state-action bank: The treebank was the full Wall Street Journal corpus of the Penn Treebank; the tool PENNCONVERTER (Johansson & Nugues, 2007) was used to convert from CFG to DG form. For each sentence, the sequence of parser states necessary to build the correct dependency analysis is listed, as well as the correct actions to take to get to the next parser state. This treebank of state-action pairs is then used to train probabilistic features. (2) generate feature banks from state-action banks: The state-and-action banks contain all the information in each parser state: all the words in the stack, all the words in the buffer, all the dependency analyses created. If the exact parser state is not found in the treebank, the parser can still use its experience with the treebank to decide on a parser action This is done by extracting *features* and only using some of them at a given state transition. (3) machine learn the probabilistic features from the feature banks. The feature banks consist of transitions and feature instances. But, the parser requires probabilities for taking each action. Machine learning provides a method for learning the patterns in the feature banks, which result in weights for each transition and feature instance. The machine learning implementation used here is LIBLINEAR1.5 (Lin, Weng, & Keerthi, 2008). The output of LIBLINEAR1.5 is a list of feature instances, along with weights for each parser action. These weights are then normalised: each weight is divided by the sum of the four action weights, resulting in probabilities for each parser action. A set of probabilistic features with weights constitutes the *oracle*.

The particular implementation of a non-projective transition system that I use in the thesis is HUMDEP3.0 (Boston, 2013). It is a k-best parallel parser where beam width k is a parameter that the user specifies. All the results in the thesis were generated using the default beam width $k = 3$. HUMDEP uses four data structures as follow. σ is a stack of already-parsed non-reduced words; τ is an ordered input list of unparsed words; h is a function from dependent words to head words; d is a function from dependent words to arc types. Permissible transition types for HUMDEP are given here:

Left-arc a left arc is drawn from the current word being parsed i to the first word j of σ

Right-arc a right arc is drawn from the first word j on σ to the current word being parsed i making i the head of j; j is pushed onto σ

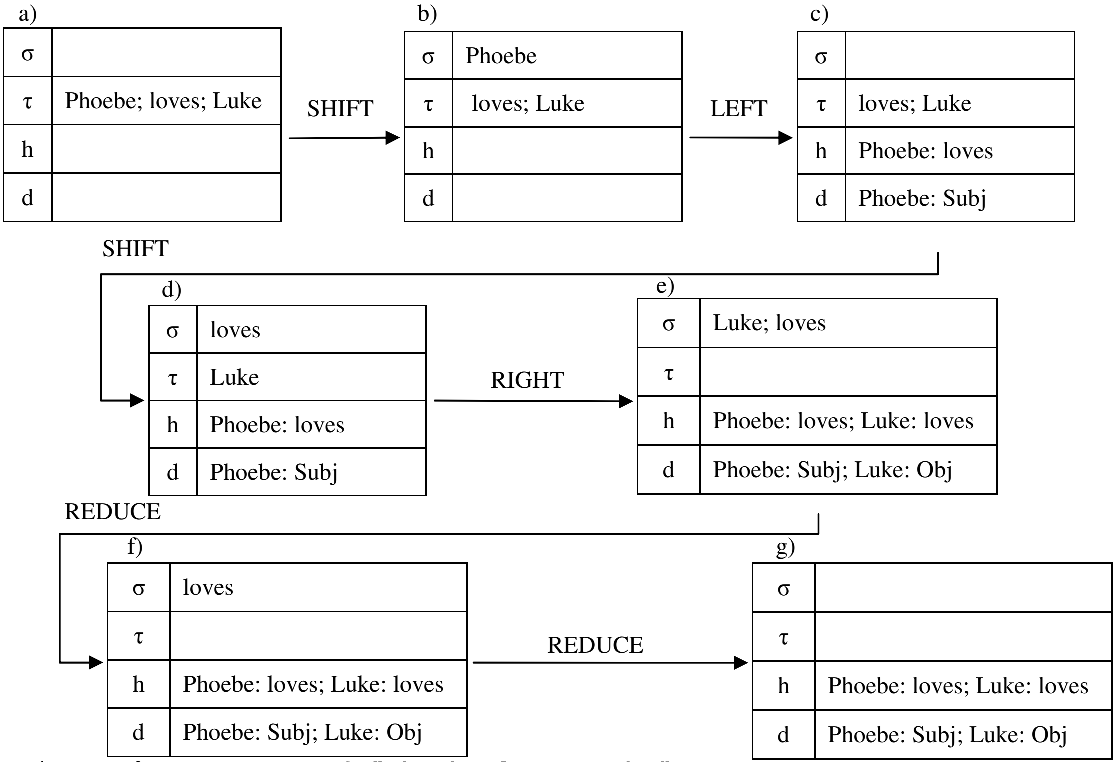
Shift shifts the current word being parsed j onto σ without drawing any arcs.

Reduce pops σ (applies only if the top word has a head)

Figure 2.1 illustrates how a Nivre parser's configuration changes during the processing of the sentence *Phoebe loves Luke*. In the initial configuration (Figure 2.1a) the stack σ is empty, the input list τ contains the full input string, and the h and d functions are empty. Because LEFT-ARC, RIGHT-ARC, and REDUCE are not viable options, the parser must SHIFT the first word *Phoebe* onto the stack, leading to the configuration in Figure 2.1b. A LEFT-ARC from *loves* to *Phoebe* pops *Phoebe* off the stack and adds head information to h (the head of *Phoebe* is *loves*) and arc-type information to d (the arc from *loves* to *Phoebe* could be labeled with the Subject function). Similarly, once the parser is in the state shown in Figure 2.1d, a RIGHT-ARC transition leads to configuration Figure 2.1e, which now defines the Object function of the sentence. Finally, when no inputs are left, the parser uses the REDUCE action to pop the stack until the parser is in the final configuration, with σ and τ empty (Figure 2.1g). What remains is the information in h and d, which is used to draw the dependency analysis for the sentence.

The parser follows Nivre (2004b) in using a generative probability model to rank parser actions. A k-best search algorithm explores the top k=3 parser configurations according to the ranking established by the probabilities. There follows a specification of the model's features. Some features take into account the parts of speech to be connected by potential dependency arcs, while others explicitly model the width or direction of a potential arc:

Configuration The probability of a transition T is the probability of the transition given the current configuration (conditioned by the top three elements in σ and the first element in τ)



a)

σ	
τ	Phoebe; loves; Luke
h	
d	

b)

σ	Phoebe
τ	loves; Luke
h	
d	

c)

σ	
τ	loves; Luke
h	Phoebe: loves
d	Phoebe: Subj

d)

σ	loves
τ	Luke
h	Phoebe: loves
d	Phoebe: Subj

e)

σ	Luke; loves
τ	
h	Phoebe: loves; Luke: loves
d	Phoebe: Subj; Luke: Obj

f)

σ	loves
τ	
h	Phoebe: loves; Luke: loves
d	Phoebe: Subj; Luke: Obj

g)

σ	
τ	
h	Phoebe: loves; Luke: loves
d	Phoebe: Subj; Luke: Obj

Figure 2.1: How a shift-reduce parser processes the sentence *Phoebe loves Luke.* in stages (a) to (g).

Part-of-Speech (POS) Pair The probability of an arc R from word$_i$ to word$_j$ is the probability that POS(word$_i$) heads POS(word$_j$)

Surface Distance The probability of arc R from word$_i$ to word$_j$ is the probability of the number of words between POS(word$_i$) and POS(word$_j$).

Directionality The probability of an arc R from word$_i$ to word$_j$ with direction d (left or right) is the probability that an arc from POS(word$_i$) to POS(word$_j$) is of type d

These features, when weighted, give probabilistic advice to the parser for carrying out a particular action. This can be used to model human garden-path processing. Boston and Hale (2007) show how the parser handles a particular syntactic ambiguity: the subject object ambiguity, an example of which is given here, where *the answer* may initially be attached as the object of *knew* or as the subject of a new VP – *the answer was wrong*.

(2.4) John knew the answer very well.

(2.5) John knew the answer was wrong

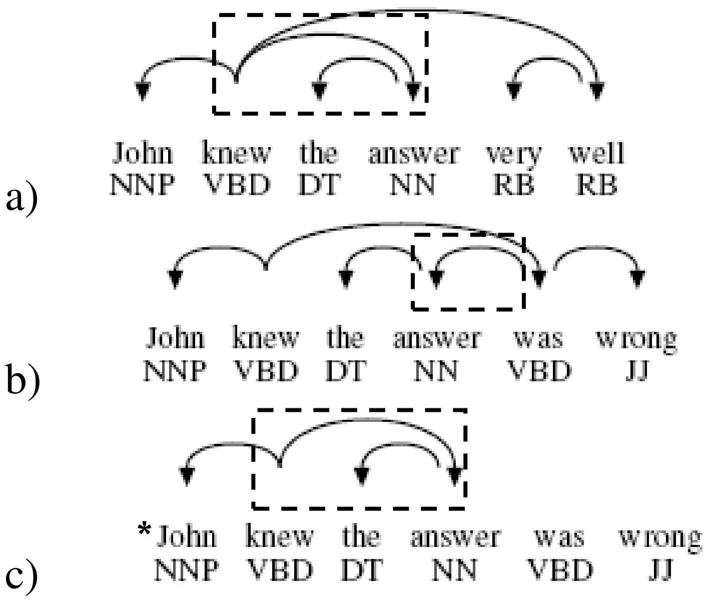

a)
John knew the answer very well
NNP VBD DT NN RB RB
b)
John knew the answer was wrong
NNP VBD DT NN VBD JJ
c)
*John knew the answer was wrong
NNP VBD DT NN VBD JJ

Figure 2.2: How HumDep parses subject-object ambiguity

With reference to the panels (a) to (c) in Figure 2.2, a Subject-Object ambiguity arises when a parser assumes a noun to be the direct object (DO) of a verb (panel a), unaware that further input may reveal a second verb to which the noun can attach as a subject (S) (panel b). The human preference in this case is for the reading in (a), which leads to a garden path in Figure (c). Both the *POS Pair* and the *Directionality* features favour the S reading of the sentence, where the noun *answer* is not immediately attached as an object of the verb *knew*. The *POS Pair* feature rates a SHIFT transition higher than a RIGHT-ARC transition and *Directionality* slightly prefers a LEFT-ARC from a verbal head *was* to a noun *answer*.

However, the *Configuration* feature is able to produce the correct human garden path by choosing a RIGHT-ARC transition to attach the noun *answer* to the verb *knew*. The *Configuration* feature prefers the action that leads to the garden path DO reading over alternatives leading to the S reading. This particular type of ambiguity is not as difficult for the human processor in the sense that both readings of the sentence (a) and (b) are acceptable (Ferreira & Henderson, 1990; Kimball, 1973).

2.4 Metrics of incremental probabilistic parser load

“The point of the working cognitive model is that it should be able to predict when humans will find certain sentences difficult. But, in order to do this, there must be a way to measure parser difficulty." (Boston, 2012, p. 68).

This section describes two such metrics of parser difficulty which are used in this thesis. The models of parsing described above (humdep and tdparse) both yield *surprisal* either with respect to a probabilistic grammar or the dependency equivalent, a list of features that specify the probability of each parser action at a given state in the sentence. TDPARSE also yields *entropy reduction*. Both metrics are used to formulate *linking hypotheses* that link parser actions to cognitive load. The sentence processing theories that embody these linking hypotheses are discussed in Chapter 3.

2.4.1 Surprisal

When a grammar is augmented with probabilities, as in a PCFG, it becomes possible to assign probabilities to sentence parses. This is done by multiplying together the probabilities on all rules that were used to construct that parse. In the event that the sentence being parsed is globally ambiguous with 2 alternative parses, each of the alternatives would be assigned its own probability composed of the product of the probabilities on all the rules used in the particular parse. In this event the probability of a sentence with two alternative parses would be the sum of the probabilities of each parse.

Because a parse is constructed incrementally, it is possible, following the efficiency gains in the work of Stolcke (1995), to examine each state transition in the process, and ask what the probability is of the partial sentence seen so far at that state transition. The probability of a parse of a partial sentence is computed the same way as a parse of a full sentence – by multiplying together the probabilities of the rules that generated the partial string at that state transition.

When we can find what the probability is for a parse of a sentence fragment as it stands after n words have been processed, and when we can find what the probability is for a parse of an extended fragment of the same sentence after the next word $n + 1$ has been processed, we can then ask questions about the contribution of a word $n + 1$ to the probability of the parse that contains it, making reference to the previous probability of that parse at word n. One way to quantify this contribution is to measure the ratio of the probability of the parse at word n to the probability of the parse at word $n + 1$. *Surprisal* is derived from this quantity using the equations provided below. When this quantity is large we can say that word $n + 1$ contributed greatly to the probability of the current parse, and when it is small, we know this is because word $n + 1$ contributed little to the probability of the current parse. Words which contribute little extend parses in ways which are found frequently in the corpus over which the probabilities were computed, whereas words that contribute greatly extend parses in ways which are less common in the training corpus.

Surprisal is computed using two other quantities. These quantities are: (1) the probability of a derivation: a derivation is a set of weighted rule productions that result in the current partial string of input words, such that a sentence fragment with two alternative parses is represented as two derivations; (2) prefix probability: this is the probability of the parse of the fragment seen so far, which is composed of the sum of the probabilities of the two derivations if the fragment is syntactically ambiguous with two alternatives.

Let d be a derivation composed of a series of applications of grammar rules. Let i index these applications so that d_i is the ith application in d, and let j be the total number of applications in the derivation. Then the probability of a derivation is given by the product of the probability of each rule applied in the derivation, thus:

$$\text{probability}(d) = \prod_{i=1}^{j} \text{probability}(d_i) \tag{2.6}$$

Let $\mathcal{D}$ represent the set of all derivations d that are present for the current sentence fragment – when there are two alternative parses available for the sentence fragment seen so far, $\mathcal{D}$ has two elements. Let w be the set of words in the sentence fragment seen so far. Let w_k be the word that the parser encountered most recently at the current state. Let w_{k+1} be the first word of the rest of the sentence. As the parser transitions from its state at w_k to its state at w_{k+1} we can derive a *prefix probability* at w_{k+1} that represents the sum probability of the derivations of the string $w_{1...k+1}$. So the prefix probability of word w_{k+1} with respect to grammar G and the prefix $w_{1...k}$ is given by the sum of the probability of all derivations of the string $w_{1...k+1}$ that the grammar generates.

$$\text{prefix probability}(w_{k+1}, G, w_{1...k}) = \sum_{d \in \mathcal{D}} \text{probability}(d) \tag{2.7}$$

The conditional probability of the next word given a grammar and a prefix is the ratio of the prefix probability of the next word w_{k+1} to the prefix probability of the current word w_k.

$$\text{conditional probability}(w_{k+1}, G, w_{1 \dots k}) = \frac{\text{prefix probability}(w_{k+1}, G, w_{1 \dots k})}{\text{prefix probability}(w_k, G, w_{1 \dots k})} \tag{2.8}$$

The surprisal of the next word given a grammar and a prefix is the negative log of the conditional probability of the next word.

$$\text{surprisal}(w_{k+1}, G, w_{1 \dots k}) = -\text{log}(\text{conditional probability}(w_{k+1}, G, w_{1 \dots k})) \tag{2.9}$$

2.4.2 Entropy reduction

In general, the entropy (Shannon, 1948) of a random variable is the uncertainty associated with that variable. Specifically, for a discrete random variable X with outcomes $x_1, x_2, \dots$ with probabilities $p_1, p_2, \dots$

$$\text{entropy}(X) = -\sum_{x \in X} p_x \log_2 p_x \tag{2.10}$$

When all outcomes are equally likely, entropy is maximal. For example, a fair die has six equally probable outcomes. The probability of a particular outcome, e.g., the probability of rolling a five, is $1/6$. The entropy of the die is then

$$\begin{aligned} \text{entropy}(\text{die}) &= -\sum_{x \in \text{die}} \frac{1}{6} \log_2 \frac{1}{6} \\ &= -\left(6 * \left(\frac{1}{6} * \log_2 \left(\frac{1}{6}\right)\right)\right) \\ &= 2.58 \text{ bits} \end{aligned} \tag{2.11}$$

Putting this in sentence processing terms, let $\mathcal{D}$ be a set of derivations d for a string containing the words read up to the current word k, words $w_{1 \dots k}$. A derivation d is the set of rule productions for that string $w_{1 \dots k}$. Let $\text{pp}(w_{1 \dots k})$ be the prefix probability of the string $w_{1 \dots k}$. The entropy of $\mathcal{D}$ for the string $w_{1 \dots k}$ is given by:

$$\text{entropy}(\mathcal{D} | w_{1 \dots k}) = -\sum_{d \in \mathcal{D}} \text{pp}(w_{1 \dots k+1}) \log_2 \text{pp}(w_{1 \dots k+1}) \tag{2.12}$$

The first k words of a sentence are words $w_{1 \dots k}$. The set of derivations given the string $w_{1\ldots k}$ is $\mathcal{D}|w_{1\ldots k}$. The information I conveyed by the set of derivations for words $w_{1\ldots k}$ is denoted $\mathrm{I}(\mathcal{D}|w_{1\ldots k})$ and defined by subtracting the entropy at the current word from the entropy at the previous word:

$$\mathrm{I}(\mathcal{D}|w_{1\ldots k}) = \mathrm{entropy}(\mathcal{D}|w_{1\ldots k-1}) - \mathrm{entropy}(\mathcal{D}|w_{1\ldots k}) \tag{2.13}$$

The amount of information that a parser gets from the current word is the amount by which uncertainty about the derivation of the string containing the current word is reduced from the uncertainty about the derivation of the string as it was before the current word was processed. Finding this quantity requires us to know what the entropy of the current string is. Equation 2.12 shows us that in order to compute the entropy of the current string we need to do computations over words that have not been seen yet. In Hale's theory, the computations are done over all possible extensions that are consistent with the current sentence fragment, but this requires a fairly small grammar. In Roark's implementation, fairly large grammars are used. To keep the computations over a large grammar tractable, Roark extends the current string by one word only, and calculates the entropy of that set. The limits on what the next word could be are that adding the word has to result in a string that the grammar generates, and that the end of sentence marker </s> counts as a possible next word. Because Roark's measure is calculated from just one additional word beyond the current word, it is an approximation to Hale's conditional entropy of grammatical continuations, which is over complete derivations. In the thesis I use Roark's approximation.

Subtracting entropy at the current word from entropy at the last word gives a positive quantity if entropy at the current word is less than entropy at the last word, and entropy reduction is large to the extent that a word takes a sentence from a state of high uncertainty to a state of relatively lower uncertainty. In the case that entropy at the last word has a higher value than entropy at the current word, the subtraction gives a negative quantity. Hale describes this situation as one in which the new word leaves the processor in a more uncertain state than before. In this case "no progress disambiguating the string has occurred, analogous to pushing against a heavy boulder on an incline, which nonetheless drives the pusher backward" (Hale, 2006, p. 650). Hale's formula for the ER measure given here imposes a lower bound at zero so that the value of ER itself is never negative:

$$ER = max(0, I(\mathcal{D}|w_{1\ldots k})) \tag{2.14}$$

Chapter 3

Theories of sentence processing

This chapter gives the main theories of sentence processing, and their predictions for differential processing difficulty at disambiguating words.

3.1 *Garden-path* theory

The label *garden-path theory* is given to a cluster of accounts that take a similar approach. They are serial models that construct an initial analysis by means of heuristics until it is inconsistent with an input word. This is typically a disambiguating word, in which case these models propose that the initial analysis is revised in line with the information conveyed by the new input word.

Recall the example sentence fragment *The horse raced ....* Perhaps the simplest answer to the question how a word like *raced* (that has more than one way of being integrated into the sentence) should be handled by the parser is offered by Bever (1970). Bever suggests a simple rule that states a preference: *prefer to attach a verb as an active main verb instead of as a passive participle*. This rule would have to be represented somewhere in the parser itself since it is not derivable from the grammar or from the string of words, but it offers a way for the parser to make a choice at the onset of ambiguity.

There are some hidden assumptions underlying a proposal like Bever's that can be brought into light and which are characteristic of this class of approaches to the onset of ambiguity. The statement of a preference for one alternative over another hides an assumption that the parser's response to the onset of ambiguity is to pursue one alternative and not to pursue others. Such a parser can be described as having a narrow beam of width one, because it pursues a single analysis, where a logically possible alternative would be to pursue both analyses in a parser whose beam is of width two. Parsers whose beam is of width greater than one are considered in the section after this one, while the present section deals with a variety of ways that a parser with beam width of one can identify that single analysis at the onset of ambiguity.

One disadvantage of a system like Bever's is that it requires a great many rules additional to those in the grammar: one for each type of ambiguity that the parser could encounter in sentences across the language. For this requirement to be met, the number of individual rules must be very large, and this may be incompatible with human memory limitations. One way for the number of rules to be limited without departing from the commitment to a single analysis at the onset of ambiguity is to express them in a more general form such that a smaller number of rules has the same coverage of sentences that are possible in the language. Whereas Bever's rule for the main clause / reduced relative clause ambiguity is hard-coded for that ambiguity, an ideal more general set of rules would abstract over particular ambiguities and provide the parser with a set of principles for dealing with any ambiguity that it might come across.

There are proposals for such general heuristics in the literature. In order for these heuristics to apply more generally than per-ambiguity rules, they must appeal to some higher order notion that subsumes individual ambiguities under a common banner. There are heuristics that appeal to *minimality* to abstract over particular ambiguities and heuristics that appeal to *locality* to abstract over particular ambiguities.

Minimality heuristics appeal to a notion of syntactic simplicity to abstract over particular ambiguities. The most general form of the minimality heuristics says: *prefer to build a structurally more simple analysis than to build a structurally complex analysis*. Minimality accounts differ as to how they measure structural simplicity. In the Sausage Machine model of Frazier and Fodor (1978) structural simplicity is measured by the number of syntactic nodes that would have to be built to accommodate each alternative analysis. The analysis that mandates the fewest nodes to be built is considered to be the structurally simplest analysis, and choosing this analysis satisfies the assumption that exactly one single parse should be favoured and pursued. The heuristic is called *minimal attachment* in this framework. The way that minimal attachment generalises over particular ambiguities to provide a heuristic is to operate on the number of nodes that must be built, and not on some preference order for one path over the other at some particular ambiguity onset.

Whereas minimality heuristics abstract over particular ambiguities with reference to a count of syntactic nodes, locality heuristics abstract over particular ambiguities by leading the parser to combine nearby constituents in preference to combining farther away constituents. The appeal is to notions of distance, which may be expressed as distance between words, or more abstractly as distance between higher-order constituents in the phrase marker like noun phrases or main clauses. *Late closure* (Frazier, 1978) is a locality heuristic that privileges attachment of the incoming word into the currently considered constituent and disprefers attachment of the incoming word into a new constituent. The effect of applying this principle is to use new material to pursue, depth-first, the existing clause. It prevents the parser from pursuing the alternative and starting a new clause. *Right association* (Kimball, 1973) also says that incoming syntactic material is preferentially construed with recently built structures rather than with long-ago built structures and thus appeals to the same notion of locality as does *late closure*. *Local association* (Frazier & Fodor, 1978) is another guise for the same appeal to locality. The *recency preference* of Gibson (1991) is another locality-based principle.

Within locality heuristics there are two kinds: *windowing* models and *right-to-left* models. Windowing models include the 3 constituent window of M. Marcus (1978) which constrains the number of constituents in the current processing window; and the Sausage Machine model of Frazier and Fodor (1978) which constrains the number of words in the processing window to 5 or 6 words. Locality effects fall out of windowing models because only within-window attachments are available for parsing, so that if a constituent or word can be attached locally, it will be attached locally, regardless of any availability of non-local attachments *outside* the window. Windowing models became unpopular as it became clear that for any window size, counter examples can be found to demonstrate that the given window size is unable to account for empirically-established human parsing preferences. Right to left models include the recency preference of Gibson (1991) in which whenever there are several attachment points for an adverbial phrase, only the most recent is considered. Such models need to stipulate that recency is preferred, in contrast with the way that locality effects fall out of windowing models.

Verbs are particularly important in accounts of the onset of ambiguity since many ambiguities concern the attachment of arguments to verbs. This requires a notion of the verb as the head of a clause, and a notion of the verb having arguments. This is often indicated by using the terminology of θ to represent any word that has arguments - typically a verb which might have an agent and a recipient as well as adverbial clauses that qualify the manner of action. For example in the following sentence:

(3.1) The spy shot the policeman with the rifle.

the theta-assigning word is the verb *shot*, which has a scope (or *domain*) that includes its agent *spy* and its patient *policeman*. Both the agent and the patient are considered to be roles that are assigned by the verb: they are part of its theta-domain. The adverbial clause giving the manner of shooting *with the rifle* is also considered to be part of the theta-domain. The fact that the manner of shooting is considered to be an argument of the verb *shot* shows how a notion of locality can be expressed that operates higher up than the level of the words in the sentence. To see this, consider that the string *with the rifle* is not particularly close to the verb *shot* in this case: the patient of the verb *the policeman* intervenes between the words *shot* and *with the rifle*, but appealing to theta-assignment permits a theory to treat a verb and its arguments as "close", and it is in this sense that such theories can be considered to be locality-based.

An example of a parsing principle that honours theta-assignment relations is Pritchett's *theta-reanalysis constraint* (Pritchett, 1992) that states "syntactic reanalysis that reinterprets a theta-marked constituent as outside of the current theta-domain is costly". While this principle is expressed as a reanalysis constraint, the prior notion of a domain that a constituent is *inside* or *outside* is really a theory of locality expressed over syntactic units rather than words.

Frazier and Rayner (1982) observed two kinds of eye movement behaviour at disambiguation and explained them in a serial reanalysis framework. In Frazier and Rayner's garden-path model, repair is choosing a new analysis of the existing left-context that is compatible with the disambiguation. The first they called *chaos*, where fixations were very long in the disambiguating region of the sentence but "eye movements generally continued in a forward direction through the sentence. Upon reading the end of the sentence, the subject then made a long regression to the beginning of the sentence and reread the sentence" (Frazier & Rayner, 1982, p. 196). This was taken to indicate that subjects had great difficulty understanding the sentence, but that they had “no insights as to what the nature of the processing difficulties were” (Frazier & Rayner, 1982, p. 197). The second pattern *disruption* was associated with long fixation durations or regressions where the “reader fixated on the disambiguation region for an average amount of time and then immediately made a regression back to the ambiguous region of the sentence” (Frazier & Rayner, 1982, p. 197). This was taken to indicate that “the reader was able to reanalyse the sentence and resolve the ambiguity, but it took some additional processing after encountering the disambiguation”.

While minimality and locality heuristics are intended to meet the assumption that a parser must achieve a single analysis at the onset of ambiguity, this is not the only logically possible approach to the onset of ambiguity. An alternative class of accounts shares a different assumption that the parser responds to the onset of ambiguity by generating each alternative. The terms *serial* and *parallel* are often used to describe this difference.

3.2 Decay model

The decay model (Ferreira & Henderson, 1991b, 1998) uses the *minimal attachment* and *late closure* heuristics to make initial attachments. When a phrasal head or other theta-assigner is encountered, all argument structures associated with the theta-assigner are activated in parallel with weights on the structures according to their relative frequency. A thematic role is assigned to a phrase *as soon as the head of that phrase is encountered* (Ferreira & Henderson, 1998, p. 84).

The model makes use of the notion of a *Thematic Processing Domain (TPD)* to explain how reanalysis is initiated. The TPD refers to the scope of the theta-assigner including all its arguments, and the decay of activated theta-assigned roles. Reanalysis is triggered by any of these conditions: when one TPD is embedded inside another; when two TPD’s must interact; or when two theta-assigners are adjacent. If any of these occur, the parser will detect an error at disambiguation, and will retrieve a previously-discarded theta-domain. When the syntactic analysis breaks down, syntactic and thematic reanalysis occurs.

Differential reanalysis difficulty is based on how easily the required argument structure can be retrieved. This depends on both the amount of the weighted initial activation for the structure and the amount of decay that has occurred since it was activated. The claim that thematic roles associated with unadopted argument structures of a theta assigner start to decay once theta roles are assigned seems reasonable: if this was not the case then a very large number of roles would have to be maintained in working memory during sentence processing.

Repair is more difficult in the model when the onset of ambiguity is further back in the sentence because the activation of the structure decays more over the greater number of intervening words than when the onset of ambiguity is nearer to disambiguation. Since thematic assignment has to wait for the head of a phrase to be processed, less decay will have occurred at disambiguation in the case of a head-final NP than in the case of a head-initial NP. So, assuming equal initial activation, reanalysis for ambiguously attached head-initial NPs is predicted to be more difficult than for ambiguously attached head-final NPs, because the role associated with the unadopted argument structure of a head-final NP will be easier to retrieve.

Ferreira and Henderson (1991a) used sentences like the following examples to show differential reanalysis costs based on the length of an ambiguous region and the associated decay of an initially-activated structure. The percentages following them indicate grammaticality judgements, so that 100% would indicate that everyone tested considered the sentence grammatical. The sentences come from two different experiments with different participants.

(3.2) After the Martians invaded the town the people were evacuated. (82%)

(3.3) After the Martians invaded the town that the city bordered the people were evacuated. (64%)

(3.4) After the Martians invaded the town was evacuated. (69%)

(3.5) After the Martians invaded the town that the city bordered was evacuated. (18%)

The first finding to be discussed, using Examples 3.2 – 3.3 and Examples 3.4 – 3.5, is that a longer ambiguous region leads to lower grammaticality judgements. Examples 3.2 – 3.3 are *early closure* sentences and Examples 3.4 – 3.5 are *late closure* sentences and thus not expected to cause difficulty under the *garden-path* model. Examples 3.2 and 3.4 have short ambiguous regions and Examples 3.3 and 3.5 have long ambiguous regions. The effect of lengthening the ambiguous region was to reduce grammaticality judgements. Surprisingly, the effect obtained for the non-garden-path pair as well as the garden-path pair but this aspect was not pursued in this incarnation of the decay model. A simple decay interpretation is that the alternative argument structures made available at the theta assigner, but not immediately used, had longer to decay during a relatively long ambiguous region than during a short ambiguous region, and were thus harder to bring back into the analysis upon disambiguation.

(3.6) If the clerk forgets the customer typically yells. (63%)

(3.7) If the clerk forgets the customer that likes Bobby typically yells. (35%)

(3.8) If the clerk forgets the customer with turquoise shoes typically yells. (35%)

An alternative explanation that Ferreira and Henderson (1991a) sought to rule out is that it is not syntactic complexity that is responsible for the grammaticality judgement gradient. To spell out this alternative possibility, consider that 3.8 requires more nodes than 3.7. Yet the grammaticality judgements did not differ significantly across the two cases. From this, Ferreira and Henderson (1991a) concluded that whatever might be responsible for the disadvantageous effect of lengthening the ambiguous region, it could not be syntactic complexity.

The next finding we will consider uses Examples 3.9 and 3.10.

(3.9) While the boy scratched the dog that Sally hates yawned loudly. (24%)

(3.10) While the boy scratched the big and hairy dog yawned loudly. (51%)

Here ‘dog’ is the head of the misanalysed phrase and ‘yawned’ is the disambiguation. When the head of the misanalysed phrase is further from the disambiguation (as in 3.9) the sentence is perceived to be less grammatical than when the head is closer to the disambiguation (as in 3.10). This explanation also works for Examples 3.2 to 3.5.

Finally we consider effects of argument structure frequency, and see how a modified decay model can accommodate them. We will use Examples 3.11 and 3.12 which both resolve to complements but where the verb ‘knew’ is more frequent with a complement than with an object whereas for ‘saw’ the reverse is true. From Ferreira and Henderson (1991b) :

(3.11) The woman knew the nervous man would leave. (*easy*)

(3.12) The woman saw the nervous man would leave. (*hard)*

Greater difficulty was found for the sentence with *saw* indicating that when resolution is against frequency, comprehension is harder. If argument structures are accessed in parallel and weighted by frequency, then more frequent structures will be more readily accessed at disambiguation than will less frequent structures, accounting for the frequency effect here.

It is only necessary to add this assumption of weighting by frequency to the decay model to extend it to cope with these data. Furthermore, the structure not chosen on the first pass will be a more frequent structure for biased-to-correct than for balanced verbs, where *biased-to-correct* means that, of the available argument structures, one is more frequent than the other, and the more frequent structure turns out to be correct; and where *balanced* means that the available argument structures are equally frequent.

This offers an opportunity for reanalysis to benefit from the greater activation of the not-chosen structure of biased verbs relative to the not-chosen structure of balanced verbs. Support for this comes from Ferreira and Henderson (1991b) who showed that using biased verbs rather than balanced verbs reduced the disadvantageous effect of a long ambiguous region on grammaticality in Early Closure sentences.

Garnsey, Pearlmutter, Myers, and Lotocky (1997) investigated effects of verb bias (called *subcategorisation frequency* in this thesis) in direct object vs sentential complement (DO/SC) ambiguity using sentences like the following, which all require a SC interpretation:

(3.13) The talented photographer accepted the money could not be spent yet. (*DO-bias, plausible DO*)

(3.14) The talented photographer accepted the fire could not have been prevented. (*DO-bias, implausible DO*)

(3.15) The sales clerk acknowledged the error should have been detected earlier. (*EQ-bias, plausible DO*)

(3.16) The sales clerk acknowledged the shirt should have been marked down. (*EQ-bias, implausible DO*)

(3.17) The ticket agent admitted the mistake had been careless and stupid. (*SC-bias, plausible DO*)

(3.18) The ticket agent admitted the airplane had been late taking off. (*SC-bias, implausible DO*)

They found that varying the plausibility of the ambiguously attached NP (e.g., *the money, the fire*) as a direct object of the verb (e.g., *accepted*) resulted in no effect of plausibility in sentences with SC-biased verbs (like *admitted*), but resulted in effects of plausibility for sentences with verbs that had other biases (e.g., DO-bias like *accepted*, or equi-bias like *acknowledged*). They interpreted this finding as evidence that sub-categorisation frequency / verb bias guides readers away from the DO interpretation of the NPs in sentences with SC-biased verbs.

3.3 Monotonicity proposal

Sturt et al propose that a generalized monotonicity constraint can explain differential processing difficulty (Sturt & Crocker, 1996, 1997, 1998). Revisions to the existing representation of the sentence are proposed to be easy in the case where the appropriate revision effects only monotonic changes, and difficult in the case where the appropriate revision requires destructive (non-monotonic) changes. Easy reanalysis is predicted for sentences that can be parsed by the core parser, and difficult reanalysis is predicted for sentences that "do not receive a parse" Sturt and Crocker (1996) from the core parser. The question which sentences can be parsed by the core parser is taken up in this section.

The following sentences will be used as examples in this section. The model predicts that the core parser can process 3.19 and 3.20 but that the core parser cannot process 3.21 (i.e., the core parser "does not assign a parse" to 3.21). To see why the core parser is restricted to handling only some of these examples, it is necessary to give the restrictions on the operations that the core parser is claimed have at its disposal, and then show that 3.19 and 3.20 can be parsed within these restrictions but that 3.21 results in a contradiction within these restrictions.

(3.19) John knows the truth.

(3.20) John knows the truth hurts.

(3.21) While John was washing the dishes crashed on to the floor.

The model is expressed in terms of *trees* in Description Theory (M. Marcus, Hindle, & Fleck, 1983): trees adhere to the the following conditions. Dominance and precedence are both defined as transitive relations. Dominance is reflexive (every node dominates itself) and precedence is irreflexive.

Single root condition there is a single node, the root node, which dominates every node in the tree: $\exists x \forall y \cdot \mathrm{dom}(x, y)$

Exclusivity condition no two nodes can stand in both a dominance and a precedence relation: $\forall x, y \cdot \mathrm{prec}(x, y) \vee \mathrm{prec}(x, y) \leftrightarrow \neg\mathrm{dom}(x, y) \wedge \neg\mathrm{dom}(y, x)$

Inheritance (*no-tangling*) condition) all nodes inherit the precedence properties of their ancestors: $\forall w, x, y, z \cdot \mathrm{prec}(x, y) \wedge \mathrm{dom}(x, w) \wedge \mathrm{dom}(y, a) \leftarrow \mathrm{prec}(w, z)$

The conditions on trees are compatible with many possible implementations of a parser (e.g., Gorrell, 1995). Some additional constraints serve to identify the Sturt and Crocker (1996) implementation. These are:

Strict incrementality each word must be connected to the current tree description *at the point at which it is encountered* through the addition of a non-empty set of relations to the description

Structural coherence at each state, the tree description should obey the conditions on trees: (a) single root condition; (b) exclusivity condition; (c) inheritance.

Full specification of nodes tree-descriptions are built through the assertion of dominance and precedence relations between fully specified nodes. In the current implementation, each node is a triple $\langle Cat, Bar, Id \rangle$, consisting of category Cat, bar-level Bar and an identification number Id. Each of these three arguments must be fully specified once the structure has been asserted.

Informational monotonicity the tree-description at any state n must be a subset of the tree-description at state $n{+}1$. Thus the parser may not delete relations from the tree description.

Obligatory assertion of precedence if two or more nodes are introduced as sisters, then precedence relations between them must be specified.

Grammatical coherence at each state, each local branch of the phrase marker described must be well-formed with respect to the grammar.

Two operations that the parser can perform are *simple attachment* and *tree-lowering*. There are two kinds of simple attachment, *left* and *right* attachment. These are defined here, and illustrated in Figure 3.1:

Left attachment let D be the current tree description, with root node R. Let S be the subtree projection of the new word, whose left-most attachment site, A, is of identical syntactic category with R. The updated tree description is $S \cup D$, where A is identified with R.

Right attachment let D be the current tree description, with the first right attachment site A. Let S be the subtree projection of the new word, whose root, R, is of identical syntactic category with A. The updated tree description is $S \cup D$, where A is identified with R.

The parser is also capable of creating a new attachment site with reference to a verb's argument structure. If simple attachment is not possible, then the parser attempts to perform a second mode of attachment, *tree-lowering*. *Tree-lowering* is defined using the notion of *accessibility*. Both terms are defined here, with an illustration in Figure 3.2:

accessibility Let N be a node in the current tree description. Let W be the last word to be attached into the tree. N is accessible iff N dominates W, and N does not dominate any unsaturated attachment sites.

tree lowering Let D be the current tree description. Let S be the subtree projection of the new word. The left attachment site A of S must match a node N accessible in D. The root node R of S must be licensed by the grammar in the position occupied by N. Let L be the set of local relations in which N participates. Let M be the result of substituting all instances of N in L with R. The attachment node A is identified with N. The updated tree-description is $D \cup S \cup M$.

Tree-lowering will succeed only if the operation is performed on an *accessible* node. Figure 3.3 illustrates successful tree-lowering. Sometimes tree-lowering will be attempted for some node such that checking whether that node can occupy its position R in Fig 3.2 requires subcategorisation information to be retrieved that

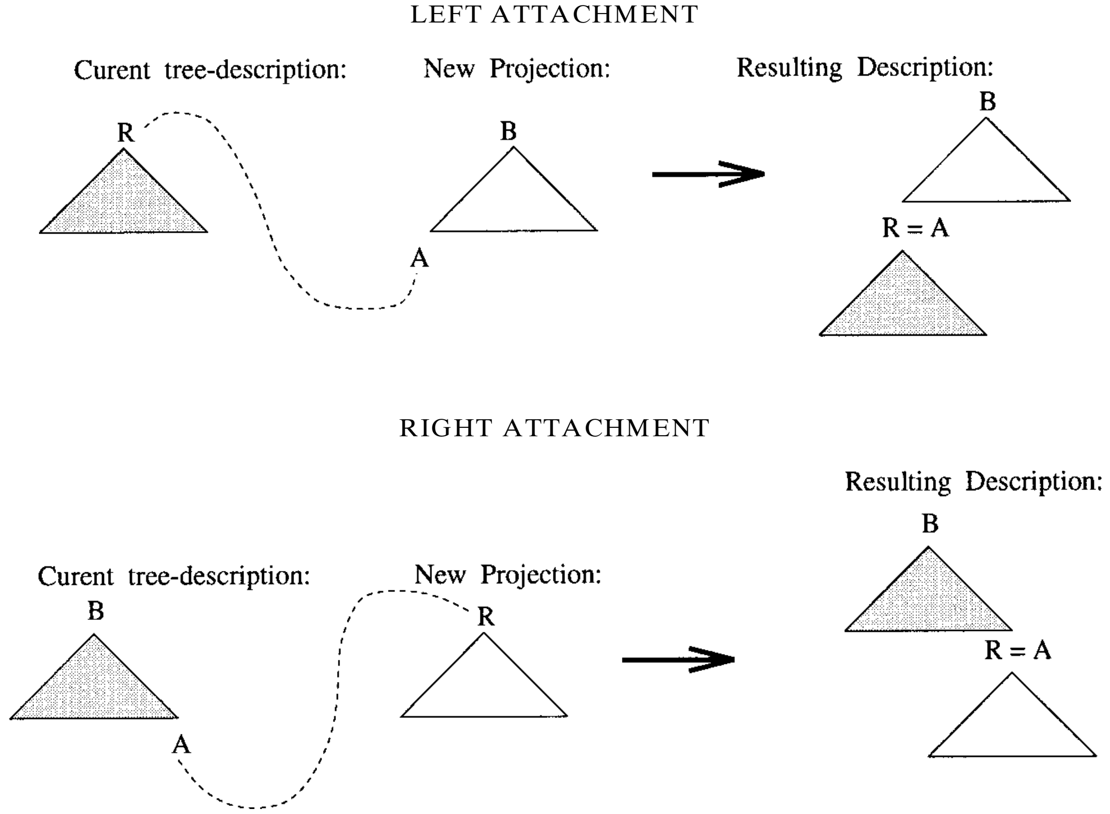


Figure 3.1: An illustration of *left* and *right* attachment in the model of Sturt and Crocker (1996). R is the root node. B is the node that is not the root node. A is the attachment site. Figure taken from (Sturt & Crocker, 1996, p. 461).

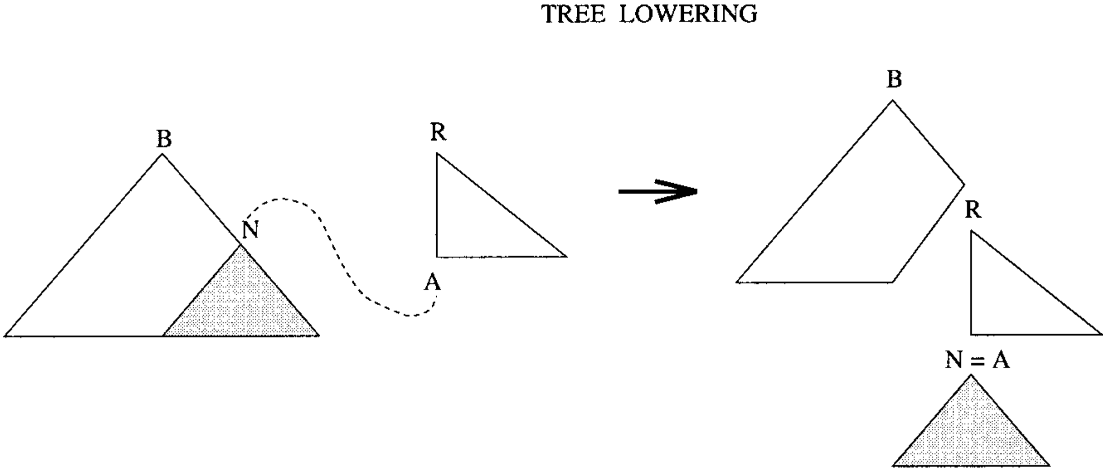


Figure 3.2: Schematic illustration of tree-lowering. The node R must be licensed in the position previously occupied by N. Figure taken from (Sturt & Crocker, 1996, p. 466).

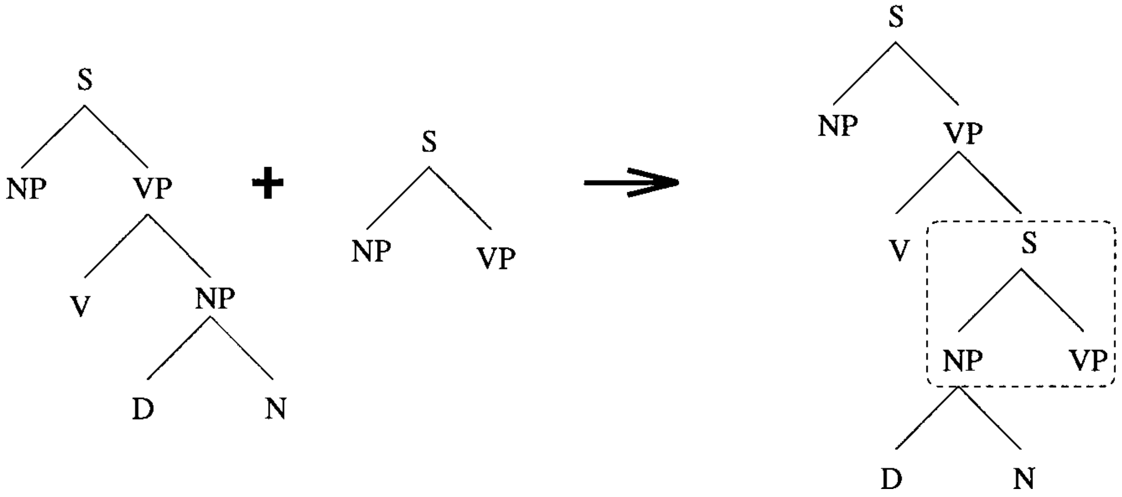


Figure 3.3: Successful tree-lowering, monotonic reanalysis for accessible nodes. Reanalysis as the insertion of one tree inside another. The inserted material is enclosed inside the dotted lines. Figure taken from (Sturt & Crocker, 1996, p. 465)

is associated with some word that is in the left-context, but is no longer itself *accessible* in sense in which the term is used in the model. This thesis takes up the question whether this kind of retrieval of subcategorisation information is reflected in eye movements in reading.

We are now in a position to concisely state how the model of Sturt and Crocker (1996) operates. It is a parser that carries out simple attachment until that fails, whereupon the accessible nodes of the current tree-description are considered until a node is found at which tree-lowering may be applied. The parser successively considers higher and higher nodes until either tree-lowering succeeds, and the parser continues to the next input word, or the current node path is exhausted, in which case the core parser fails, and the string is rejected.

If the parser was attempting to process an ungrammatical sentence then tree-lowering would fail and the string would be rejected, but in the model, tree-lowering also fails for grammatical sentences like Example 3.21, which Sturt et al call *conscious* garden-path sentences to indicate that processing at a higher level than that provided by the core parser is necessary for successful processing of these sentences. The model appeals to the distinction whether a sentence can be fixed by tree-lowering or not to explain the greater human difficulty observed for sentences like 3.21 (which cannot be fixed by tree-lowering) versus sentences

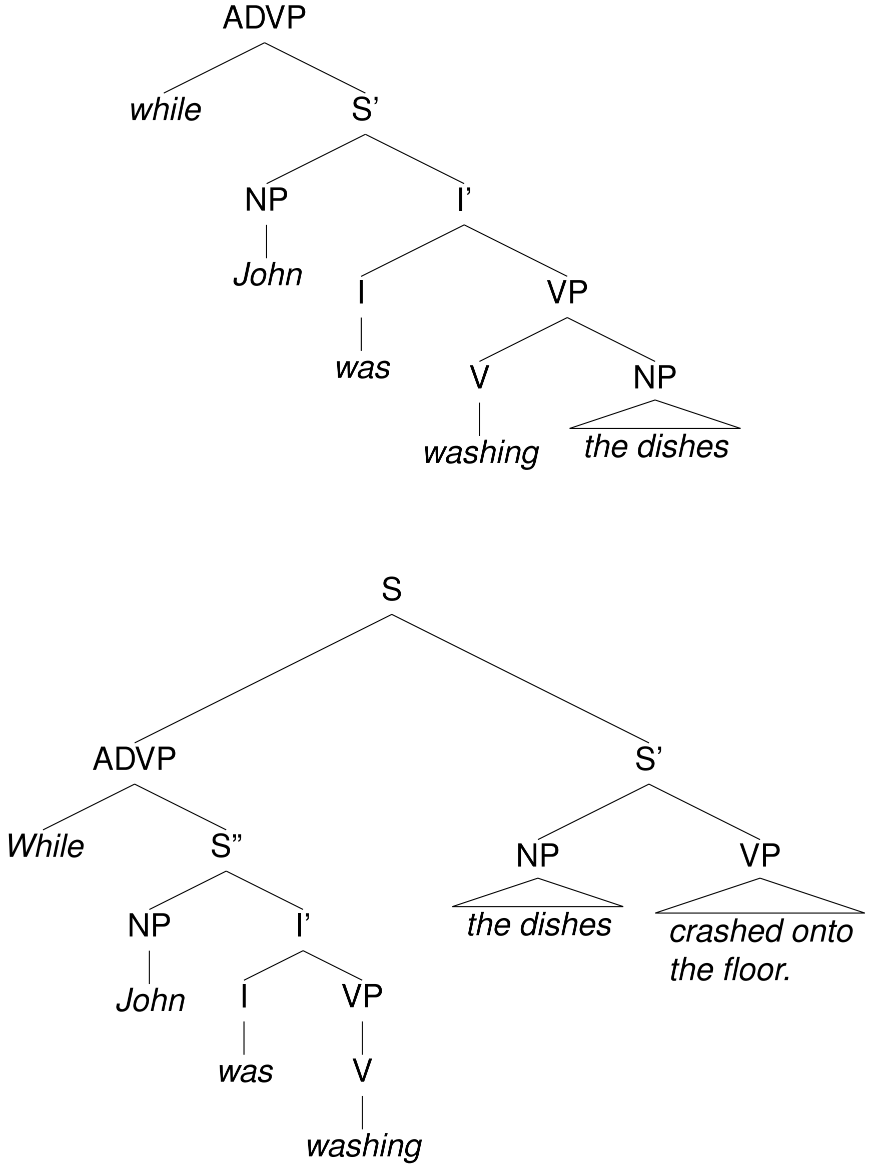


Figure 3.4: Illustration of a case that tree-lowering rejects. Top: an adverbial phrase formed by attaching *the dishes* low, as a sister node of *washing*. Bottom: A phrase-marker for the whole sentence formed by attaching *the dishes* high, as the first constituent of the main clause of the sentence. In Sturt et al’s monotonicity proposal, attempts at *tree-lowering* fail for this type of *conscious* garden-path sentence.

like 3.20 (which can be fixed by tree-lowering).

Figure 3.4 illustrates why tree-lowering fails when the new word cannot be attached into the existing clause (i.e., the new word must be must be attached high). In the representation of the full sentence, it can be derived that the original VP now precedes the NP (through the inheritance condition), but this leads to a contradiction, because we now have $\mathrm{dom}(VP, NP) \wedge \mathrm{prec}(VP, NP)$ which goes against the exclusivity condition. The result of this violation of the exclusivity condition is that the parse cannot be achieved by the core parser.

The way in which the term *monotonicity* is used by Sturt and Crocker (1996) is as follows. Monotonic revisions are those that tree-lowering can implement. Non-monotonic revisions are those that tree-lowering cannot fix. Thus the parser does not offer an explanation of graded difficulty within the set of *conscious* garden path sentences, in contrast with, for example, the diagnosis model described in section 3.4, where gradations of difficulty come about as a result of the varying number of times that a given fixing principle must be carried out.

3.4 Diagnosis model

The Diagnosis model (Fodor & Inoue, 1994, 1998, 2000) explains repair with reference to the grammar and whether the grammar licenses a given attachment. The model treats reanalysis as an extension of the routines that bring about parsing itself.

First we consider the assumptions of the model, as follows. This is a member of the class of limited repair models. The model assumes that syntactic structure is built. This stands in contrast to the view that initially-activated parses are either deselected by entropy or pruned in a parallel parser. The model assumes that repairs are effected. This stands in contrast to the position that there is no reparsing, as in strictly incremental parsers. The observed difficulty at disambiguation is explained as the parser's difficulty in deducing which repairs are called for - whereas the majority view is that the explanation is difficulty in effecting repairs. In the diagnosis model there is no special machinery for revisions. Instead the model invokes existing first pass machinery and calls it as a function repeatedly with different parameters. The model's coverage extends to the possibility that diagnosis may fail, as a way of covering the evidence for re-reading the sentence from the beginning as a consequence of unexpected disambiguation.

Parsing is implemented in the diagnosis model by a principle called Attach, which is described thus:

Attach "On receiving a word of the input sentence, connect it to the current partial phrase marker for the sentence in such a way that the resulting current partial phrase marker is syntactically well-formed, though possibly incomplete at its right edge" (Fodor & Inoue, 1998, p. 103).

In the event that no syntactically well-formed current partial phrase marker can be brought about by the application of Attach, the diagnosis model applies a principle called Attach Anyway, which is described thus:

Attach anyway "Having established that there is no legitimate attachment site in the current partial phrase marker for the current input word, attach the input word into the current partial phrase marker wherever it least severely violates the grammar, and subject it to the usual preference principles that govern Attach." (Fodor & Inoue, 1998, p. 105)

This leaves the current partial phrase marker in an ill-formed state, albeit the least ill-formed state consistent with the Attach principle. The parser now applies a principle called Adjust, which is described thus:

Adjust "When a grammatical conflict has been created between two nodes X and Y in the current partial phrase marker, by either Attach Anyway or Adjust, eliminate the problem by altering minimally (i.e., no more than is necessary for conflict resolution) whichever of X and Y was less recently acted on, without regard for grammatical conflicts thereby created between that node and others in the current partial phrase marker." (Fodor & Inoue, 1998, p. 106).

Note that the Adjust principle is recursive: it can be applied to its own output. This allows for the situation where Adjust leaves the current partial phrase marker in an ill-formed state - the current partial phrase marker in this ill-formed state is itself subjected to a further round, or series of rounds, of application of the Adjust principle, and it is in this way that adjustments can be propagated up through the current partial phrase marker to bring about an ultimately well-formed current partial phrase marker.

A further principle called the Grammatical Dependency Principle (GDP) guides the parser mechanistically to the source of the ill-formedness in the current partial phrase marker:

Grammatical Dependency Principle "When a grammatical violation has been created in the current partial phrase marker by an action on node n in accord with Attach Anyway or Adjust, attempt to eliminate the problem by acting on a node that is grammatically incompatible with n" (Fodor & Inoue, 1998, p. 109).

In the Diagnosis model of Fodor and Inoue (1998) the difference between NP/S and NP/Z sentence types is critical, and the model offers reasons why ambiguous NP/S sentence types are found easier to process than ambiguous NP/Z sentence types (these two cases have been shown to differ in difficulty, by Sturt et al. (1999)). The diagnosis explanation is cast in terms of *stealing* - the disambiguating word *steals* the initially misattached noun away from the first clause and into the second. There are two types of stealing: *capture* and *theft*. Descriptions of each term follow.

Capture In the case that a verb has multiple subcategorisation frames, and in the event that the parse involving the wrong subcategorisation frame is initially pursued, then, upon disambiguation, the parser can reinspect the subcategorisation frames of the verb and select an alternative. Crucially, because there is a grammatical link between the material that has been wrongly attached and the verb that it has been wrongly attached to, the parser can traverse this link and directly find the verb whose subcategorisation frames must be reinspected. The diagnosis model claims that sentence types for which the verb has an alternative subcategorisation frame can be fixed more quickly by the parser than sentence types for which the verb does not have an alternative subcategorisation frame. The term *capture* is used to describe the process by which the wrongly attached material is *stolen* into a new clause in sentences where the verb does have at least one alternative subcategorisation frame. Below there is a worked example of the NP/S sentence type that uses capture, using Example 3.22.

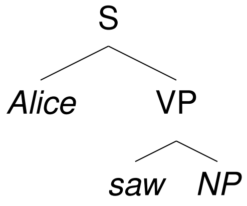
Theft In the case that the verb does not have an alternative subcategorisation, the wrongly attached material can still be *stolen* into a new clause, but it cannot be attached as a complement of the verb. The parser has to project a lot of new structure in cases like this, both to demote the initial part of the sentence to an adverbial clause, and to nest this clause into the new representation of the sentence by projecting a new main NP and VP. The term *theft* is used to identify cases where the verb does not have an alternative subcategorisation frame, and to describe the process by which the wrongly attached material must be *stolen* into a new clause in sentences where the verb does not have any alternative subcategorisation frames. Below there is a worked example of the NP/Z sentence type that uses theft, using Example 3.29

A worked example of NP/S ambiguity There follows a worked example of how the diagnosis model parses a sentential complement ambiguity:

(3.22) Alice saw Bob limped.

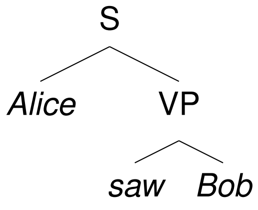
The example starts with processing *Alice saw*. The verb *saw* has two subcategorisation frames, for a NP and for a sentential complement. Each frame has a different attachment site for an NP at word 3. The direct object attachment site for a noun in position 3 is:

(3.23)



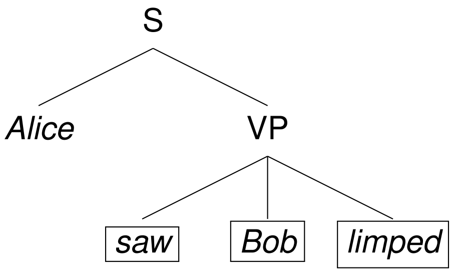
The ambiguity typically resolves to a direct object reading. The parser is assumed to pursue the direct object reading first. Word 3 is *Bob*. The parser attaches *Bob* in NP position using attach:

(3.24)



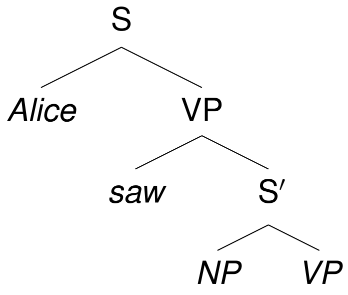
Word 4 is *limped*. The parser tries to use attach, but because there is no valid attachment site for *limped*, the attach anyway principle is used to attach the word in the place of least violation, which is as a sister node of *saw* and *Bob*, as shown in the example below, where the words that are affected by the attachment are boxed:

(3.25)



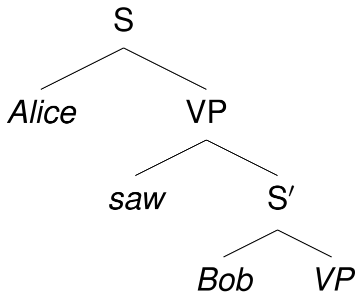
The parser is led to the words which are affected by the bad attachment. The parser looks up alternative subcategorisation frames for *saw* and recalls that *saw* can take a sentential complement, illustrated below:

(3.26)



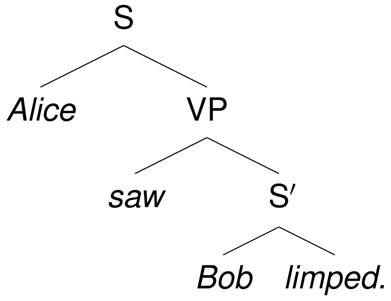
In applications of the adjust principle, the parser checks whether the constituents required by the new subcategorisation frame unify with constituents available among the words affected by the bad attachment. The new subcategorisation frame requires a NP and a VP from among the words in red. The NP requirement is satisfied by *Bob*:

(3.27)



The VP requirement is met by *limped*. At the full stop the structure is complete:

(3.28)

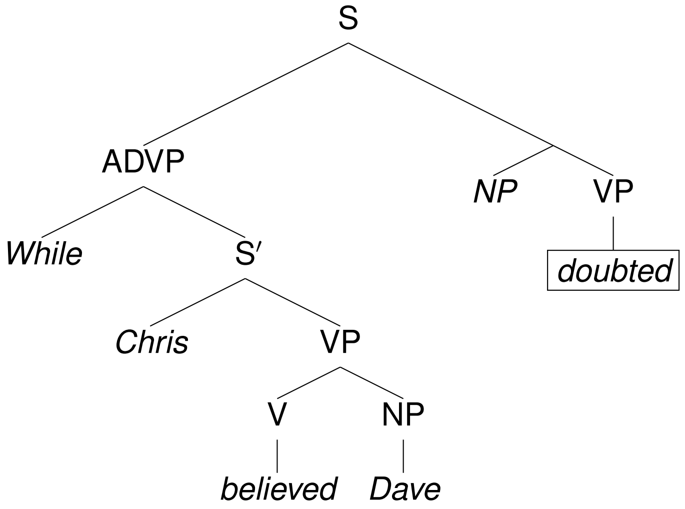


A worked example of NP/Z ambiguity Now we turn to NP/Z ambiguity and see how the diagnosis model handles it, using example 3.29:

(3.29) While Chris believed Dave doubted.

At *While Chris believed Dave doubted* there are several possible subcategorisation frames for *believed*. One frame is for *believe* to take a direct object as a complement, as in example 3.30 below.

(3.30)

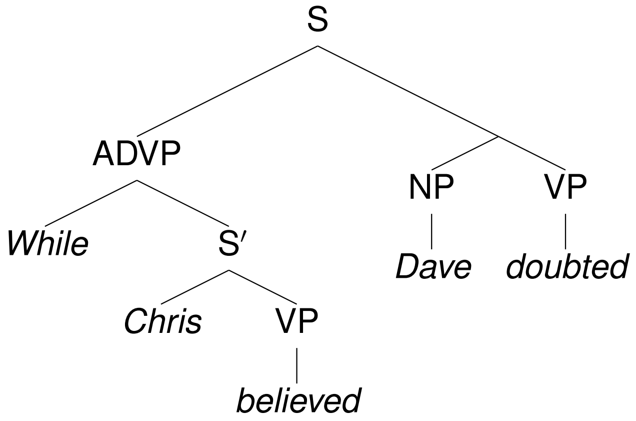


This is the more common frame and it is pursued first. The next word is *doubted*. It is first 'attached anyway' in the place of least violation as part of the VP (boxed in example 3.30). This means leaving the NP constituent empty which the GDP forbids. The parse attempts to fill the NP constituent but the GDP also forbids reaching across to the NP in the subordinate clause (Fodor & Inoue, 1998, p. 125). The parser must resort to searching the partial string of terminals rather than constituents in this case. The authors describe this as follows:

> “Because there is no grammatical dependency along which the parser can travel from the matrix to the lower clause, the parser is forbidden by the GDP to steal the object NP node or make any other structural change in the lower clause. But it could try stealing at a different level where access to structural facts is not at issue. That is, it could grab some nearby words in the word string to fill up the hole in the matrix clause where the subject ought to be. This is to steal linearly and very superficially from the word string, rather than structurally and legitimately from the tree” (Fodor & Inoue, 1998, p. 125)

Once *believed* is considered, the parser retrieves the alternative subcategorisation where *believed* takes no complement. Then contents of the NP *Dave* are available for consideration as components of the main NP, and this succeeds. Eventually the correct representation using the “no complement" frame for *believed* is achieved as in example 3.31 below.:

(3.31)



The crucial difference between (1) *Alice saw Bob limped.* and (2) *While Chris believed Dave doubted.* in the Diagnosis model is that in (1) the process of integrating *limped* can be done at constituent level whereas in (2) the GDP prevents the integration of *doubted* being done at constituent level and it must be done at terminal level by terminal search instead, which takes longer than constituent search in the model.

In the model, the cost of implementing terminal search is claimed to grow with the number of terminals. However, the cost of implementing constituent search does not grow with the number of constituents. This is why the model predicts an interaction between sentence type and head-position: early-head capture is no harder than late-head capture but early-head theft is harder than late-head theft due to increasing search size in theft only.

It is because the parser has a mechanistic way to reach the source of the initial error in the current partial phrase marker that the diagnosis model parsimoniously avoids postulating a repair agent that must be endowed with the capacity to reason abstractly about the current partial phrase marker in order to calculate what changes would bring about a well-formed current partial phrase marker. Consider briefly what such an agent would have to be able to do. It would have to be able to identify that the current partial phrase marker was ill-formed. It would have to be able to calculate what changes needed to be made. It would have to be able to establish that those revisions do lead to a well-formed current partial phrase marker.

In summary, the diagnosis model assumes that the parser must expend effort in order to revise an ill-formed current partial phrase marker. This effort is quantified in terms of how many times Adjust needs to be deployed before the current partial phrase marker is rendered well-formed. Some revisions can be achieved by few deployments of Adjust, but others will require many deployments of Adjust before the original misattachment is rectified. The model distinguishes itself from models that assume that certain types of structure are harder to build than others (e.g., those involving raising rather than lowering a node) by attributing the observed differential difficulty of different revisions to differences in how many times a simple mechanistic principle needs to be applied to reach and rectify the original misattachment, implementing the intervening revisions as it goes along. In this way the diagnosis parser is able to predict graded difficulty for revisions.

Eye movements in capture and theft The Diagnosis model offers an explanation of the differences in reading time for capture and theft sentence types. It does not explicitly link the two sentence types to any particular distribution of regressive saccades. However one extension of the model that is tested in the thesis makes the link as follows. Disambiguation in capture sentences can be done with the benefit of the syntactic links between terminals and constituents established on the first pass. In this case the proposed link with eye movements is that in capture the eyes are directed straight to a particular word representing the ambiguously attached NP or the verb that it attaches to. The regressive scan path should have few elements, and the elements present should be over-distributed across the the ambiguously attached NP and the verb that it attaches to. Disambiguation of theft sentences requires that those syntactic links between terminals and constituents be established again in a different way. In this case it is proposed that the link with eye movements is that the eyes are directed back over all the previous terminals, with no privileged direction towards words by reference to the constituent that they represent, since the links from terminals to constituents must be built again differently than they were on the first pass. The regressive scan path should have more elements, and the elements should be randomly distributed over all previous terminals, weighted by distance from the launch site so that short-distance movements are more likely.

3.5 PDP accounts

An example of the class of Parallel Distributed Processing (PDP) parsers is the parser in Tabor, Juliano, and Tanenhaus (1997). Their model considers parsing as a dynamical system. A metric space contains *attractors* and potential parses can be described in terms of how near they are to the attractors. The model integrates many sources of information to carry out its computations, from the syntactic level but also from the lexical level. The model also computes conditional probabilities based on verb argument frequencies.

Tabor et al. (1997) specify that their model is not a repair model "[The current model] is *not* equivalent to serial garden-path models, which assume that slowed reading times following an ambiguous region of a sentence typically reflect an incorrect first parse commitment followed by a time-consuming revision. Rather long reading times often reflect a competition process in the spirit of most constraint-based models." (Tabor et al., 1997, p. 263).

3.6 Bayesian accounts

An example of the class of parsers that use dynamically-updated conditional probability rather than just static frequencies is the parser in Jurafsky (1996); Narayanan and Jurafsky (1998, 2002).

Jurafsky's model operates on conditional probabilities computed over *constructions*. It integrates information from several layers of linguistic processing (lexical, syntactic). It is a rank-and-prune parallel parser. Ranking is done by computing conditional probabilities for given constructions. Pruning is done by removing from the candidate set constructions whose probability falls below a *threshold confidence ratio* (Narayanan & Jurafsky, 1998). The threshold confidence ratio is the ratio of the probability of the currently-considered parse to the probability of the best parse in the currently-considered set. Applying this threshold removes from consideration all constructions whose posterior probability falls enough below that of the best parse.

3.7 Competition-Integration

The competition-integration model (McRae, Spivey-Knowlton, & Tanenhaus, 1998) is a model of a sub-part of sentence processing. It is tailor-made for resolving the main clause / reduced relative ambiguity. It does this by satisfying multiple constraints acting in concert. Although implemented for a specific ambiguity, and thus not readily generalisable (it would need one such model per ambiguity), the model is well-enough specified that, provided we work with an implementation, we can examine its workings step by step and generate its predictions. The model is also sufficiently complex that its predictions must be generated (Green & Mitchell, 2006) and not merely inferred (van Gompel, Pickering, Pearson, & Liversedge, 2005). The model operates bottom-up, considering multiple sources of information at the same time, distributing activation over two discrete alternative parses, with the most highly-activated parse pursued. A systematic diagram of the model is in Figure 3.5.

3.8 Retrieval

Retrieval is a breadth-first theory. A central notion in breadth-first responses to the onset of ambiguity is the notion of *activation*. The notion of activation is actually present implicitly in the depth-first approaches where activation is all-or-none: either the parse in question is the parse being pursued or it is not. Its activation is one if it is the parse being pursued and zero if it is not. In a breadth-first parser activation can be a more subtle business. If such a parser has three candidate analyses in play then each can have a different level of activation. This makes the candidates capable of being ranked by activation. This means that the ordering

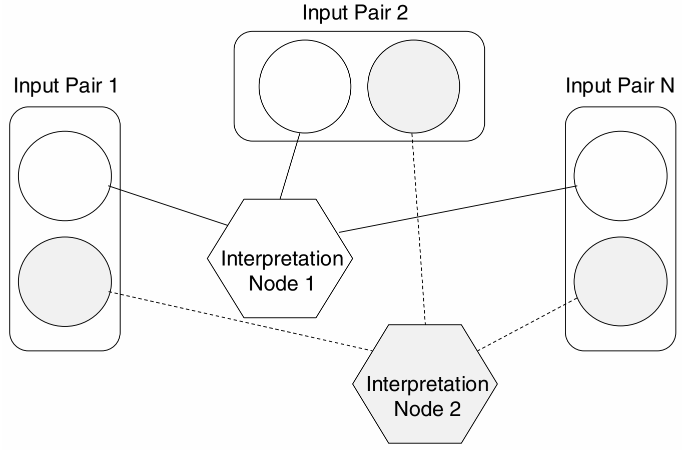


Figure 3.5: Schematic diagram of the competition-integration model. A simplified generic diagram of the competition-integration model. Different kinds of constraint are introduced to the system by a series of paired input units. Within each pair, one unit activates Interpretation Node 1 and the other supports its competitor. The number of input pairs and their participation order are details that vary from instantiation to instantiation.

of candidates may change over the course of a sentence such that a dispreferred parse may become the preferred parse at some later point. This provides a way for parallel parsers to select, at disambiguating input, a newly preferred parse by appeal to changed activation levels. The means by which activation may change at new input are central to such accounts.

One factor that plausibly affects the degree of activation at a given point in a sentence is the distance between the current input word, for example a disambiguating word, and some earlier point in the sentence, for example the onset of ambiguity. Another factor that plausibly affects the degree of activation of a candidate constituent at some given point in a sentence is decay, which may be thought of as the rate of loss of activation. Long-ago activated constituents may be expected to have less activation than recently activated constituents with the same initial degree of activation. Another influence on activation levels is interference. This is a result of the size of the candidate set. The activation of individual candidate constituents in large sets (like nouns) is diminished by the large number of competitors. In contrast if the set is small (like determiners) then the activation of its constituents will be less diminished because of the relatively few competitors.

The *retrieval* account of Lewis and Vasishth (2005) appeals to decay and interference to explain fluctuating levels of activation of candidate parses in a cohort. The model explains sentence processing difficulty at each word in a sentence as the difficulty of retrieving structures from working memory. The link with working memory constraints is achieved by embedding the account in the general cognitive framework Adaptive Control of Thought-Rational (ACT-R) (Anderson, 2005; Anderson & Lebiere, 1998). High demand on working memory motivates a prediction of how much reading time a word will involve, with higher demands leading to longer reading times.

Parsing in retrieval is accomplished by condition-action pairs generated with reference to a phrase structure grammar. A series of memory buffers stores elements in short-term and long-term buffers. Parallel associative retrieval (McElree, Foraker, & Dyer, 2003), fluctuation of activation of elements already in a memory buffer, and retrieval interference as a function of similarity are combined to predict the amount of time that it takes to read a word (Vasishth, Brüssow, Lewis, & Drenhaus, 2008).

A word's activation is based on two quantities: the baseline activation of the word, which is taken to decay given the passage of time; and the amount of similarity based interference with other words that have been parsed. The baseline activation B for a word i is given in Equation 3.32 (taken from Lewis & Vasishth, 2005; Patil, Vasishth, & Kliegl, 2009), where t_r is the time since the rth retrieval of the word, the summation is over all n retrievals, and d is a decay factor set to 0.5 as in other ACT-R models (Anderson, 2005). The equation tracks the log odds that a word will need to be retrieved, given its past usage history. It yields not a smoothly decaying activation from initial encoding to the current time, but a "series of spikes corresponding to the retrieval events" (Lewis & Vasishth, 2005).

$$B_i = \ln \left(\sum_{r=1}^{n} t_r - d \right) \tag{3.32}$$

The overall activation A for word i is given in Equation 3.33 from Lewis and Vasishth (2005). In this equation, B_i is the fluctuating baseline level of activation for word i which is subject to time-based decay. This quantity B_i is yielded by Equation 3.32. In the model, a *goal buffer* contains retrieval cues for integrating the current word. Overall activation A for word i is found by adding to the baseline activation for word i an associative activation boost received from retrieval cues in the goal buffer that are associated with i. The variable j indexes those retrieval cues in the goal buffer. W_js are weights on the retrieval cues in the goal buffer. The weight on a retrieval cue represents the proportion of the total activation available for the whole goal buffer that is assigned to the particular retrieval cue j in the goal buffer. S_{ji}s are the strengths of association from each retrieval cue j of the goal buffer to word i. This equation is effectively adding to the baseline activation an activation boost received from retrieval cues in the goal buffer.

$$A_i = B_i + \sum_j W_j S_{ji} \tag{3.33}$$

The amount of similarity based interference is estimated by the weighted strengths of association between the word to be retrieved and retrieval cues from other words already parsed and with a trace in memory. In Equation 3.34, word i is the current word, and retrieval cue j is from a word that is similar to word i, with reference to its part of speech tag, so that nouns interfere with other nouns but not with verbs. If retrieval cue j is similar to word i then the amount by which retrieval cue j interferes with word i varies according to how many words have already been associated with retrieval cue j. The array of words that is associated with retrieval cue j is considered to form a fan so that fan_j gives the number of words in the fan for cue j. The constant S refers to the maximum associative strength of 1.5 (Lewis & Vasishth, 2005). This equation is effectively reducing the maximum associative strength S by the log of the "fan" of cue j, that is, the number of items associated with j.

$$S_{ji} = S - \ln(fan_j) \tag{3.34}$$

The mapping from activation level to retrieval latency is given by equation 3.35. F is a scaling constant set to 0.14 in Lewis and Vasishth (2005). A_i is the word's activation from Equation 3.33 and e is Euler's constant. T_i is retrieval latency for word i.

$$T_i = Fe^{A_i} \tag{3.35}$$

Having given the details of the computations involved in normal parsing, the focus now moves to how the model copes at the onset of ambiguity. Lewis and Vasishth explicitly state that the model implements “serial, probabilistic repair parsing” (2005, p. 389). Lexical access is achieved by ordered access modulated by frequency and context, and competition results in one candidate being proposed. This follows from equations 3.32, 3.33, and 3.34.

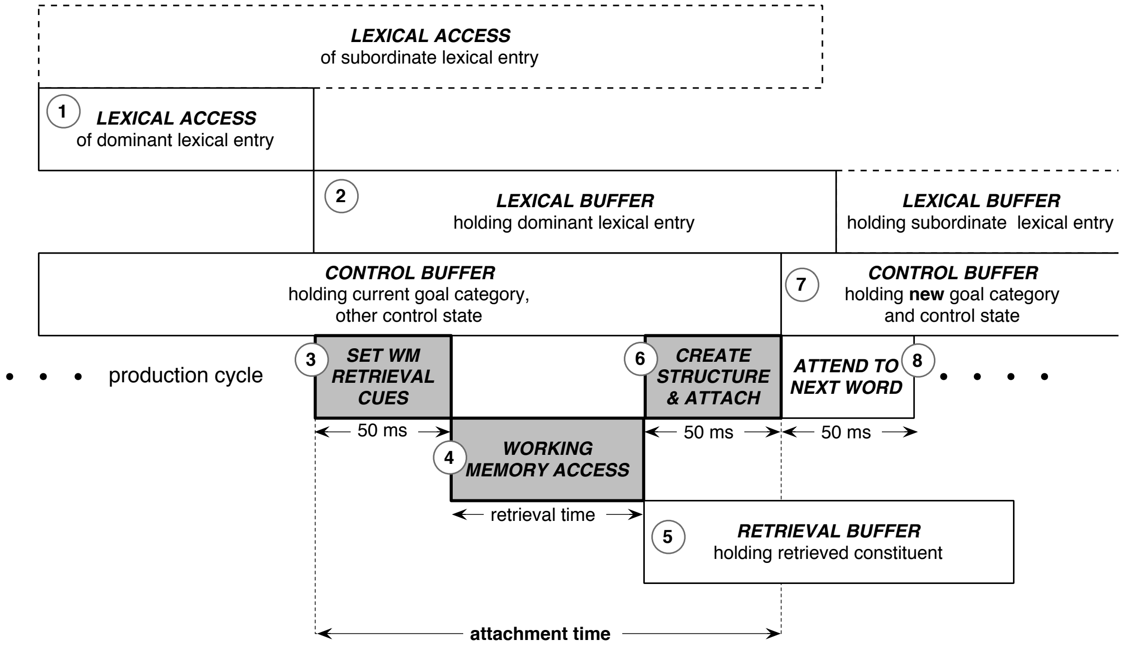


Figure 3.6: Figure taken from Lewis and Vasishth (2005, p. 383). Overview of the model, showing the critical focus buffers (control buffer, lexical buffer, and retrieval buffer) and processing dynamics (time flows left to right). The three key working-memory processes are shown in grey: (3) a production rule encoding grammatical knowledge sets cues for retrieval of a prior constituent; (4) a prior constituent is retrieved from working memory via parallel associative access; and (6) a second production rule creates the new structure and attaches it to the retrieved constituent.

Structural ambiguity is resolved by a probabilistic process that combines working memory factors (recency) and the rational production choice rules of ACT-R. Multiple possibilities for attachments are locally generated in parallel, but then a single structural analysis is pursued. Associative retrieval interference mitigates against maintaining multiple alternatives that have similar structure.

In the event that the parser pursues an analysis that later becomes untenable, limited recovery is possible. Recovery is construed as the reactivation of structures that were initially activated as possible attachments but which lost the competition on the first pass. These structures remain in memory but their activation has decayed since they were first activated: the decay contributes to their difficulty of resurrection.

Figure 3.6, from Lewis and Vasishth (2005), gives a high-level overview of the retrieval model, showing the critical buffer usage and production rule firings unfolding over time. The typical processing cycle is as follows; the numbers refer to the circled numbers in the figure:

(a) A word is attended and a lexical entry is accessed from declarative memory (1) containing syntactic information, including argument structure. The lexical entry resides in the lexical buffer (2).

(b) Based on this syntactic goal category (a kind of syntactic expectation) and the contents of the buffers, a production fires (3) that sets retrieval cues for a prior constituent to attach to.

(c) The working-memory access takes some time (4), and eventually yields a single syntactic chunk that resides in the retrieval buffer (5).

(d) Based on the retrieved constituent and lexical content, a production fires (6) that creates new syntactic structure and attaches it to the retrieved constituent. The control buffer is also updated with a new syntactic prediction (7).

(e) Finally, other productions fire that guide attention to the next word.

3.9 NL-SOAR

NL-SOAR (Lewis, 1993) is a language comprehension model based on the Soar cognitive architecture (Newell, 1990).

NL-SOAR is a single-path model: it pursues only one interpretation at a time. However the machinery of attachment is temporarily parallel in the sense that two syntactic alternatives may co-exist for a short time.

As the model proceeds through a sentence it maintains two kinds of representation of the sentence: an *utterance model* that represents the syntactic relations in the sentence so far; and a *situation model* that represents the semantics of the discourse, where the discourse extends beyond the current sentence so that the model can take advantage of semantic information that comes from previously-processed sentences in order to process the current sentence where syntax may be ambiguous.

The utterance model uses two sets of operators: *constructors* and *destructors*. The main constructor operator is *link*. The main destructor operator is *snip*. In this section I will describe how these operators generate the utterance model in simple cases.

Nodes are indexed by the syntactic relations that they can enter into. These relations form a set of relation *assigners* and relation *receivers*. The term *A/R set* is used to refer to these sets of relations and the A/R set is perhaps the central notion in NL-SOAR's account of parsing. Links between nodes are formed by matching assigners and receivers that share the same relation.

A node can only have one parent in the tree. As a result of this constraint, once a link is made, the receiving is removed from the set of available receivers. In contrast, the assigning node is maintained in the set. The reason for maintaining the assigning node in the set is so that it can play a role in subsequent semantic analysis. If the assigning node were removed from the set once a link is made, it would then be unavailable in working memory to any subsequent processes. By this rule, the assigner set can continue to provide access to partially completed syntactic structures for the benefit of interpretation processes, and this will be called upon later in the discussion of how NL-SOAR deals with ambiguity resolution. For now the focus is on how NL-SOAR implements the process that follows lexical access through to syntactic construction.

Lexical access provides a set of nodes that correspond to all the terminal positions (or leaf nodes) that the identified lexical entity can occupy. This provides coverage of simple lexical ambiguity such as the word *square* which may be either the head of an adjectival phrase (*the square box*) or the head of a simple noun phrase (*a square is not a circle*). Lexical access in NL-SOAR is then both context-free, since a word is considered without reference to words that came before, and parallel in that all the possible lexical entries for a word are considered at the same time.

Syntactic construction proceeds by linking nodes by means of the application of the operator *link*. Link operations are constrained. Constraints include number- and person- agreement, left-to-right order, subcategorisation (for verbs and their arguments), and grammatical case.

Parsing in NL-SOAR is then both bottom-up and head-driven. Lexical heads are identified in lexical access (which indicates how the word may enter into syntactic relations), and phrasal nodes are projected from these possible relations.

There are no explicit phrase structure grammar rules in NL-SOAR. Instead the utterance model follows from the interaction of Xbar structures projected from lexical heads, and independently specified constraints (the constraints on agreement, order, subcategorisation and case). Bottom-up parsing is useful for considering fragments of a sentence. Head-driven parsing helps to avoid what Lewis calls the "spurious ambiguity" (1993, p. 86) engendered by the presence of multiple grammar rules that contain the fragment so far. Such ambiguity is considered to be spurious because it is not motivated by the actual string of words that has been encountered part of the way through a sentence.

The mechanism proposed so far provides a way to link structures that attach as projections (structures that occupy specifier or complement positions). Other types of attachment include adjunction, which is taken to include all nonprojective attachments. Adjunction involves the creation of new nodes which is simply handled in NL-SOAR by the link constructor, which creates the additional node necessary.

Because lexical access is parallel in the model, temporary structural ambiguity can arise. However because NL-SOAR is a single-path comprehender, this structural ambiguity must be resolved quickly. In the following sentence ambiguity is local in scope and lexical in origin.

(3.36) The square table is large.

When the parser encounters *square*, both NP and AP nodes are created and co-exist in the problem space. At this time the determiner *the* is attached to neither. However, the determiner's syntactic relation is present in the A/R set (as spec-NP, the specifier of a noun phrase). At this time there are three structures present in working memory: the orphan determiner *the*; the NP structure for *square*; and the AP structure for *square*.

Next the determiner is attached into the NP structure in spec-IP position forming the NP *the square*, and leaving as an orphan in working memory the AP structure.

Then the word *table* appears and is projected to NP. Now there are three structures in memory: NP [the square]; AP [square]; and NP [table]. Next the AP [square] is adjoined to *table* forming NP [square table]. This means that there are now two structures in working memory that both incorporate *square*. Both of these structures are well-formed, but they are mutually incompatible because *square* cannot play both roles simultaneously.

The model cannot leave things like this because it is committed to pursue a single path. This follows from the architecture of the model that does not permit two *interpretations* (as distinct from structures) to coexist.

The operator *snip* is available: it breaks a link formerly established by the *link* operator. Snip is triggered immediately by the presence of syntactic structure that is attached to both senses of the same lexical token.

As a way of determining which structure should be remedied, snip exhibits a preference to preserve recent structure and to operate on non-recent structure. Since the AP adjunction of *square* to *table* is most recent and therefore preserved, snip operates on the determiner attachment and breaks the NP [the square].

This leaves in memory the orphan NP [square]; the orphan determiner *the*; and the AP [square table]. The link operator then attaches the determiner in the specifier position of the AP [square table] to form the NP [the square table].

The snip operator considered as a repair mechanism has the following properties that Lewis (1992) describes as *simple destructive repair*. Firstly it operates only on the structures already present in working memory: it does not make any appeal to previously built structure. Secondly its scope of operation is limited to destruction: it breaks links but does not create any.

This accounts for the attachment and repair of structural ambiguity that is lexical in origin. Next an account is given of how Lewis's NL-SOAR model processes syntactic ambiguity, using the following example of noun phrase complement / sentential complement ambiguity, and with reference to Figure 3.7.

(3.37) John knows Shaq is tall.

At the word *knows*, both of its possible subcategorisation frames are available in the A/R set. These two frames are: subcategorisation for a nominal complement, representing the interpretation *John is familiar with Shaq.* of the full sentence *John knows Shaq.*; and subcategorisation for a sentential complement, representing the interpretation *John knows (that) Shaq has some property . . .* for the partial sentence *John knows Shaq is . . . .* In NL-SOAR jargon, it can be said that *knows* permits adjunction (adjoin-V') of a sentential complement as well as the attachment of a nominal complement (comp-V').

ASSIGNERS	spec-IP:	[$_{IP}$ *is*]
	adjoin-V':	[$_{V'}$ *knows*]
	comp-V':	[$_{V'}$ *knows*]
RECEIVERS	comp-V':	[$_{CP}$ *is*], [$_{CP}$ *John knows*]

↓ LINK

ASSIGNERS	spec-IP:	[$_{IP}$ *is*]
	adjoin-V':	[$_{V'}$ *knows*]
	comp-V':	[$_{V'}$ *knows*]
RECEIVERS	comp-V':	[$_{CP}$ *John knows*]

↓ SNIP

ASSIGNERS	spec-IP:	[$_{IP}$ *is*]
	adjoin-V':	[$_{V'}$ *knows*]
	comp-V':	[$_{V'}$ *knows*]
RECEIVERS	comp-V':	[$_{NP}$ *Shaq*], [$_{CP}$ *John knows*]
	spec-IP:	[$_{NP}$ *Shaq*]

↓ LINK

ASSIGNERS	spec-IP:	[$_{IP}$ *is*]
	adjoin-V':	[$_{V'}$ *knows*]
	comp-V':	[$_{V'}$ *knows*]
RECEIVERS	comp-V':	[$_{CP}$ *John knows*]

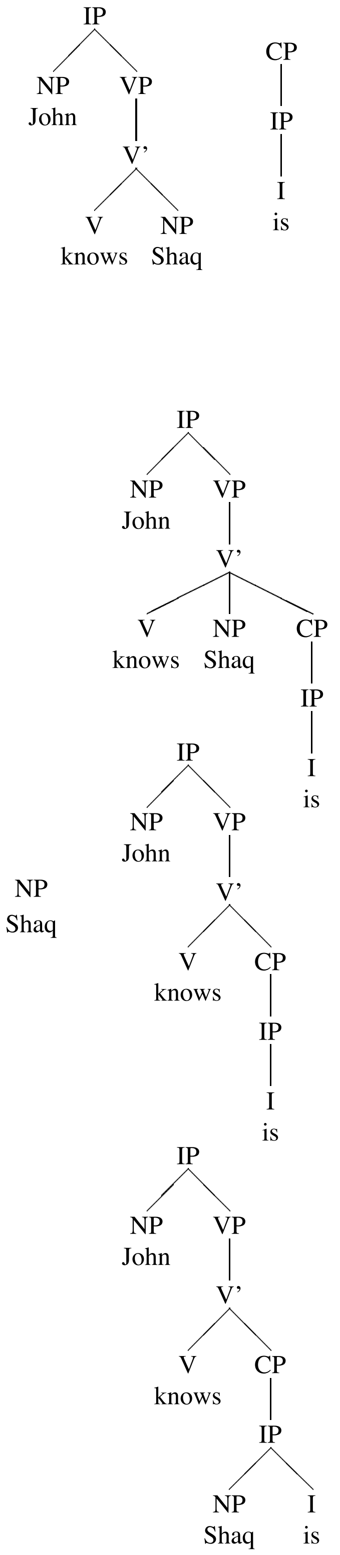


Figure 3.7: From Lewis (1993, p. 95). Figure illustrates repair in NL-SOAR for noun phrase complement / sentential complement ambiguity

Shaq is initially attached as the nominal complement of *knows*. When *is* arrives, it is projected to an IP and a CP. Lewis follows Pritchett (1992) in proposing a CP in the absence of an overt complementiser. Because *knows* is still in the assigner set, and one of the permitted relations is adjoin-V' the CP is attached in the complement position of *knows*. At this point the V' unit [knows Shaq is] is well-formed because *knows* permits each of the nominal and sentential complement relations. The utterance model at this point exhibits momentary parallelism. The parallelism takes the form of a kind of super-position of two syntactic structures: the nominal complement structure and the sentential complement structure. In other words the nominal and sentential complements are both claiming the same structural attachment position as *comp-V'* complement of *knows*.

The presence of two structures both claiming the same relation triggers the snip operator in the same way that the two interpretations of *square* resulted in mutually incompatible structures both attached to *square*. To make this clearer, although *knows* licenses a nominal complement and a sentential complement, it does not license both at the same time. This contrasts with a verb like *give* which licenses two different complement types at the same time, as in *She gave the ball to the boy*, where *the ball* represents the *source* of the transfer event and *to the boy* represents the *goal* of the transfer event, and both relations apply at the same time.

The snip operator protects the most recent attachment (the CP attachment), it operates instead on the non-recent attachment of *Shaq* to *knows*. This frees up *Shaq* as an orphaned unit, which makes it available to the link operator. Link attaches *Shaq* in spec-IP position of *is*, as its syntactic subject. This application of the link operator completes the repair of the temporary structural ambiguity, leaving exactly one well-formed syntactic structure (with no orphans) in working memory as required by the single-path nature of the NL-SOAR architecture.

So far I have given two ways in which the snip operator may apply. The first was in the description of the resolution of the lexically-based structural ambiguity in *The square table*; the second in the subcategorisation ambiguity in *John knows Shaq is tall*. In fact these are the only two ways in which snip may apply. Each is a particular example of a general pair of rules for the application of snip. The two general cases are as follows:

snip 1 When incompatible projections of one lexical item are both simultaneously attached to other lexical structure: multiple incompatible structures triggers snip, as in the multiple attachment of *square*.

snip 2 When inconsistency is detected local to a particular maximal projection: multiple incompatible attachments locally into the same maximal projection (or node) triggers snip, as in the two attachments into the *knows* maximal projection.

Although it would be possible to propose a snip operator that would apply more generally, limiting the scope to these two general cases is necessary within the NL-SOAR architecture. This is for two reasons.

Firstly, an unconstrained snip operator would give the potential for a large number of partial detached structures in working memory. Lewis shows how this affects the properties of the model's *knowledge search*: it leads to exponential match cost in the model's recognition memory (Lewis, 1993), and this would prevent the real time performance of the model.

Secondly, unconstrained snip would lead to a too-large problem space which would make the *problem search* intractable, even for relatively simple syntactic structures. This too would make the model's real time performance impossible, especially for complex syntactic structures.

By constraining the snip operator to the two general cases outlined above, Lewis provides a repair mechanism that he claims avoids the worst extremes of impossibly-large search, while still accounting for human performance on his set of target ambiguities.

3.10 Surprisal theory

The sentence processing theory that claims that differential disambiguation difficulty is related to the information theory metric *surprisal* is given in Hale (2001) and Levy (2008). Hale directly proposed *surprisal* as an analogue of human reading times. The surprisal framework does not commit to the assumption of repair. Instead, the observed human difficulty at disambiguation is construed as "the work incurred by resource reallocation during parallel, incremental, probabilistic disambiguation in sentence comprehension" (Levy, 2008, p. 1126).

For practical purposes, the surprisal value for each word in a sentence is generated as a by-product of the Earley parser algorithm (Earley, 1970) which is

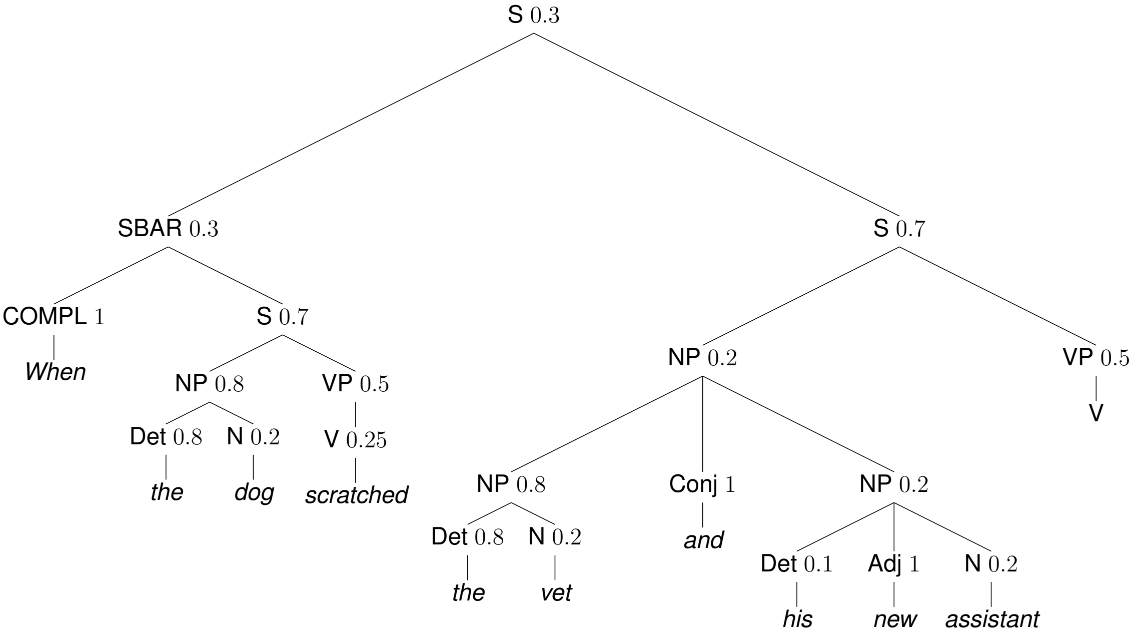


Figure 3.8: Main clause analysis (partial, annotated with probabilities for each application of a rule from the grammar)

available in several implemented forms (since Stolcke, 1995).

Applying the probabilities in the PCFG grammar in Table 2.1 to the sentence fragment *When the dog scratched the vet and his new assistant*, the derivation involving the tree representing the main clause analysis of *the vet and his new assistant* (Fig 3.8) has probability 0.174, and the derivation involving the tree representing the direct object analysis of *the vet and his new assistant* (Fig 3.9) has probability 0.826. Taking in word w_{k+1} which is *removed* results in surprisal of 4.85 bits (Levy, 2013). Part of this is due to *removed* taking out the direct object reading from the set of derivations because there is no combination of rules in the grammar that generates a direct object reading of the string $w_{1...k+1}$ so it does not contribute to the prefix probability at the new word w_{k+1}.

Surprisal offers a theoretical reason why a given word in a given sentence should vary in comprehension difficulty on the basis of knowledge from a probabilistic grammar. The approach is to "model processing difficulty as a logarithmic function of the probability mass eliminated by the most recently added word" (Patil et al., 2009). Surprisal is a measure of the information value of the most recently processed word, as rated by the grammar's probability model.

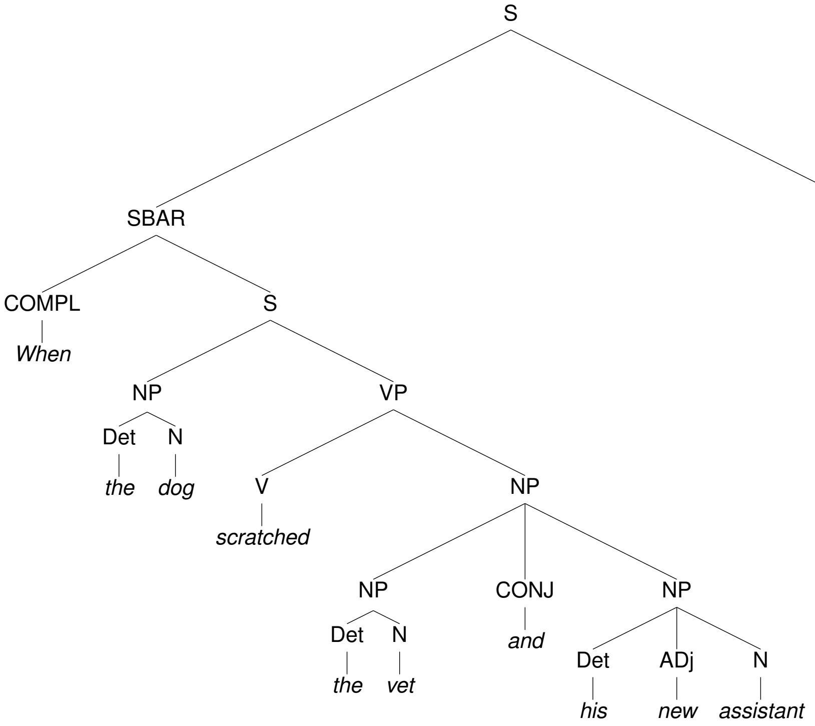


Figure 3.9: Direct object analysis (partial)

The surprisal model frames incremental sentence processing as step-by-step disconfirmation of the possible analyses of the sentence, and this casts cognitive load as the amount of work taken to disconfirm structures, where this load is greater for high-probability structures than for low-probability structures (Levy, 2013).

So far the focus has been on surprisal with respect to probabilistic phrase structural grammars. However, surprisal can also be computed over the state transitions in a Nivre-style (Nivre, 2004b) probabilistic dependency grammar in just the same way that it is computed over a probabilistic phrase-structural grammar's state transitions. In fact, surprisal is representation-agnostic, and can be computed over any grammar that represents a parse of a sentence as a sequence of state transitions and associated probabilities. In this thesis, surprisal measures are obtained both for a phrase-structural grammar, using Roark's TDPARSE (Roark, 2013), and a dependency grammar, using Boston's HUMDEP (Boston, 2013).

There is some empirical data from studies (e.g., Boston, Hale, Kliegl, Patil, &

Vasishth, 2008; Boston, Hale, Vasishth, & Kliegl, 2011; Demberg & Keller, 2008) comparing surprisal with corpora of eye movements in reading e.g., Potsdam sentence corpus (Kliegl, Nuthmann, & Engbert, 2006), Dundee eye movement corpus (Demberg & Keller, 2008). These studies show that surprisal is a good predictor of (normal) reading times in such corpora. This thesis contributes tests of surprisal against *regressive* eye movement patterns generated from experimental sentences focussing on syntactic disambiguation, rather than from corpus sentences.

3.11 Entropy reduction hypothesis

Hale's Entropy Reduction Hypothesis (Hale, 2004, 2006) states that "a person's processing difficulty at a word in a sentence is directly related to the number of bits signalled to the person by that word with respect to a probabilistic grammar the person knows." (Hale, 2004, p. 1). The link from entropy reduction (ER) to human performance is made explicit in (Hale, 2006, p. 650): "ER is positively related to human sentence processing difficulty".

3.12 Dependency theory

For Dependency theory, sentence processing is establishing word to word relations. These relations are head→dependent pairs. The main verb in a sentence has special status and is denoted *root*. Thematic subjects and objects depend on the main verb. Other constituents depend on their head (e.g., a determiner depends on its noun).

Dependency theory has ancient roots: it was used to describe 4th century B.C. Sanskrit by Panini (cited in Cardona, 1998). Modern dependency theory dates back to Tesnière (1959) who described sentences as sequences of "word-to-word connections". Work by Hays (1964) and Mel'čuk, I (1988) led to a full scale dependency theory of language, both explanatory and descriptive.

All theories of sentence processing use the notion of dependency – they differ in how much structure they claim exists over and above word-to-word dependencies. For example, phrase structural theories claim that sentences have phrasal nodes as well, whereas dependency theory does not claim phrasal nodes. Using dependency grammar for modeling human sentence processing allows models to focus on what different formalisms have in common.

The simplicity of the dependency theory approach makes it suitable for computational linguistics. For example, it was the syntactic component in early machine translation Hays (1964). It is well-suited to representing non-configurational (free word order) languages like Czech and therefore more generally applicable. Dependency grammar relations can be derived from existing syntactically parsed corpora (like the Penn Treebank) so models can take advantage of previous work.

Dependency grammar relations can be established incrementally, allowing for comparison with incremental metrics of human sentence processing, including a direct comparison for words of special psycholinguistic interest (like the disambiguating word in a sentence with temporary syntactic ambiguity). Probabilities for establishing a particular dependency relation where several are possible can also be automatically computed. So dependency theory offers incremental and probabilistic predictions of human sentence processing difficulty.

This fact, that a parser operating over a dependency grammar makes incremental probabilistic predictions, makes the approach suitable for empirical evaluation. Boston et al. (2011) showed that surprisal and retrieval computed over a dependency grammar predict fixation durations in the Potsdam Sentence Corpus (an eye movement corpus obtained from participants reading a corpus of German newspaper text) and therefore that such a dependency parser does capture some part of human sentence processing difficulty. Boston et al. (2008) found that dependency grammar predictions accounted for different types of difficulty than phrase structure grammar predictions in the Potsdam Sentence Corpus.

The existing evaluations have been done over all the words in the sentences in the corpus. In this thesis I will focus on comparing predictions for the disambiguating word in sentences with temporary syntactic ambiguity against regressive eye movements made at the same word by human participants in eye tracking experiments.

3.13 Underspecification

The *good-enough language processing* account of parsing (Ferreira & Patson, 2007) takes as its starting point the claim that readers often under-specify parses for sentences and thus never resolve many of the ambiguous sentences that are commonly held to yield information about ambiguity resolution.

This underspecification claim represents an unusual stance on incrementality in the human parser. Most models of parsing assume that the parser is incremental, with words representing increments, and that the parser seeks to do as much processing as it possibly can, making the most of new information straight away. However the good-enough account suggests that integration may not happen incrementally, even by the end of the sentence.

Weinberg's (1988; 1993) Minimal Commitment parser is incremental because the processing is done straight away at new input - however it does not make all the commitments that it possibly could as a result, restricting itself to making only some.

In Construal (Frazier & Clifton Jr, 1995) the loose attachment of non-primary adjuncts is a kind of underspecification.

Marcus's (2013) treelet hypothesis is a kind of underspecification account.

3.14 Reprocessing model

The reprocessing model Grodner, Gibson, Argaman, and Babyonyshev (2003) described here is not a repair parser. Grodner et al. (2003) give what they describe as a potentially simpler alternative to the proposition that reanalysis proceeds by repair. This alternative is that "reanalysis proceeds by reprocessing (some portion of) the input using just those grammatical operations available to first-pass parsing" (Grodner et al., 2003, p. 144). They make it clear that the term *reprocessing* is intended to cover *reparsing* and *reranking*, phenomena that they consider to be separable from a commitment to *repair*. They use evidence from extreme garden paths to show that reparsing is a necessary part of a model of sentence processing. In contrast, they argue, repair is not a necessary part of a model of sentence processing. Sentences like the following must typically be re-read several times, demonstrating that reparsing is necessary for a model of the human parser.

(3.38) The horse raced past the barn fell. (Bever, 1970)

(3.39) Tom told the children the story scared a riddle. (Frazier, 1978)

Sentences like the following (from Frazier and Rayner (1987); MacDonald (1993)) where the alternative readings are either dissimilar in syntax or dissimilar in semantics constitute cases where the parser cannot benefit from the preservation of representations: in such cases reprocessing is necessary because the correct structure is not derivable from structure built so far at disambiguation.

(3.40) The warehouse fires cause a lot of damage.

(3.41) The warehouse fires many employees each spring.

Chapter 4

Spatio-temporal distribution of regression path fixations

This chapter focuses on some theoretical issues that are not answered by the existing literature and which the rest of the thesis will take up. The chapter covers the following: the purpose of regressive eye movements; and the question whether disambiguation proceeds by repair or replacement.

4.1 Introduction

There is a special class of eye movements made during reading, i.e., regressive movements, for which it is not clear whether the immediacy and eye-mind assumptions hold, but which are used in a vast body of literature since Frazier and Rayner (1982) as though they indexed syntactic load. Frazier and Rayner (1982) took a position on regressive eye movements that has become canonical in work on parsing. Their position, dubbed *selective reanalysis*, was that regressive eye movements were tightly coupled to parsing. They identified two main classes of such regressive eye movements, which they dubbed *chaos* and *disruption*. In Frazier and Rayner (1982, pp. 196–197), they describe the chaos pattern, which can be summarised as long fixations on the disambiguation together with regression from the end of the sentence in order to re-read. The second class they dubbed *disruption*. Disruption took two forms, which can be summarised as (1) long fixations on the disambiguation, and (2) averagely long fixations in the disambiguation followed by a regression to the ambiguous region.

Some basic facts about regressions set the section in context. Baseline regression rates in reading are estimated between 10% and 15% (Buswell, 1922; Rayner et al., 2012; Vitu & McConkie, 2000). There is some evidence that regression rates increase in the disambiguation regions of syntactically ambiguous sentences (Frazier & Rayner, 1982; Meseguer, Carreiras, & Clifton Jr, 2002; Rayner, Carlson, & Frazier, 1983; Traxler, Pickering, & Clifton Jr, 1998; Trueswell et al., 1993; van Gompel et al., 2001). As for the targets of these regressions, there is evidence that readers regress more to words of low predictability with no effect of an additional manipulation of frequency (Kliegl, Grabner, Rolfs, & Engbert, 2004; Rayner et al., 2004), although see also Bicknell and Levy (2011) who note that these studies did not control for word skipping.

4.2 Models of eye movement control in reading

There are some good models of eye movements during reading that model early processes in reading and predict first pass reading times well: SWIFT (Engbert, Nuthmann, Richter, & Kliegl, 2005); GLENMORE, (Reilly & Radach, 2006); EZ-READER, (Reichle, Warren, & McConnell, 2009), MR. CHIPS (Legge, Klitz, & Tjan, 1997), EMMA (Engelmann, Vasishth, Engbert, & Kliegl, 2013), a machine learning model, (Nilsson & Nivre, 2009, 2010). There are fewer that model late processes in reading (only EMMA does), and none that focus on the spatio-temporal properties of regression path fixations.

Most eye-movement control models focus on relatively low level phenomena (e.g., the extent of the perceptual span in reading). There is only some prior work that uses eye movement data to model parsing (Binder, Duffy, & Rayner, 2001; Ferretti & McRae, 1999; Just & Carpenter, 1992; Konieczny & Döring, 2003; Spivey & Tanenhaus, 1998; Tanenhaus, Spivey-Knowlton, & Hanna, 2000; Vasishth et al., 2008). Because the prior work is of limited extent, there are plenty of fundamental issues that still need addressing when it comes to the use of eye movement data. For example, when the eyes regress, are the eye movements under linguistic guidance, or are the movements better characterised as spatially constrained under oculomotor control?

4.3 Metrics for regressions

There are some metrics of the spatio-temporal properties of fixations in regression paths. These start with qualitative descriptions such as those in Frazier and Rayner (1982), and include the regression signatures of Mitchell, Shen, Green, and Hodgson (2008), and scan path similarity (von der Malsburg & Vasishth, 2012),

4.3.1 Qualitative descriptions

Frazier and Rayner (1982) observed patterns which they gave descriptions as follow. One pattern was *chaos*, where fixations were very long in the disambiguating region of the sentence but "eye movements generally continued in a forward direction through the sentence. Upon reading the end of the sentence, the subject then made a long regression to the beginning of the sentence and reread the sentence" (Frazier & Rayner, 1982, p. 196). This was taken to indicate that subjects had great difficulty understanding the sentence, but that they had "no insights as to what the nature of the processing difficulties were" (Frazier & Rayner, 1982, p. 197). The second pattern *disruption* was associated with long fixation durations or regressions where the "reader fixated on the disambiguation region for an average amount of time and then immediately made a regression back to the ambiguous region of the sentence" (Frazier & Rayner, 1982, p. 197). This was taken to indicate that "the reader was able to reanalyse the sentence and resolve the ambiguity, but it took some additional processing after encountering the disambiguation".

4.3.2 Regression signature

Regression signatures are described in Mitchell et al. (2008, p. 274). A regression signature is the distribution of landing site regions for the set of regressive saccades launched from a given region on the first pass. The signature gives a value for each landing site region, which is the proportion of first pass regressive saccades that a given participant made that landed in that region. Because the regression signature is a proportional distribution, it has the virtue that it allows comparison between conditions with very different frequencies. The regression signature metric has some disadvantages too. It tends to be sparse in the sense that not all trials result in a regression; and in the sense that if a participant fails to make a regression in any of the conditions being compared, then that participant's data are removed from the other conditions too regardless whether they made a regression in those other conditions: this constitutes a major reduction of the data in the service of the requirement for balance in classical ANOVA. In order to achieve a balanced analysis, a lot of rare events are being discarded. This is a rather undesirable state of affairs where rare movements are the focus of investigation. In the next section we will discuss an alternative framework for analysing the regressive movements that we care about which has a different cost-benefit profile, and which does not discard data in the same way.

4.3.3 Scan path similarity

Like regression signature this measure addresses eye movements at the level of the scan path. It considers the pattern of spatio-temporal movements across whole sentences or parts thereof to be the unit of analysis. Scan path similarity analysis is a spatio-temporal edit-distance method (von der Malsburg, 2010; von der Malsburg & Vasishth, 2011, 2012). The method is new, and there are still some unresolved issues, but it is a method well-suited to studying reanalysis processes.

A scan path is a sequence of symbols. The symbols represent fixations, which are spatial locations with associated temporal durations. The scan path has properties of its own, at a level above the level of the symbols, and these properties are due to how the fixations form patterns. There are several metrics designed to capture the properties of sequences as patterns, mostly due to research in areas other than psycholinguistics. For example, in bioinformatics, it is of interest to find the best alignment between fragments of DNA, and this involves comparing sequences of symbols (Durbin, 1998). In psycholinguistics it is of interest to compare scan paths launched in response to linguistic events that are hypothesised to be qualitatively different. If two such patterns of fixations can be shown to be different, according to some measure of similarity, this can help to resolve theoretical psycholinguistic questions empirically.

If two scan paths s and t have the same number n of fixations, and fixate the same locations in the same order, but with different temporal dynamics, then the sum of the absolute difference of their fixation durations dur is a natural way to express the difference between them. This is the *Manhattan metric*, or

$Minkowski - 1$ metric of s^{dur} and t^{dur}. Using k to index the scan path's elements:

$$d_{dur}(s,t) = \sum_{k=1}^{n} |s_k^{dur} - t_k^{dur}| \tag{4.1}$$

If s and t have the same number n of fixations and the same temporal dynamics but some of the fixations are in different locations, then a simple metric is the number of fixations that fixate different locations. This is *Hamming distance* (Hamming, 1950). When the fixations in such sequences target locations that are far from each other, we might want the similarity measure to penalise this to a greater extent than the case where the different locations are close. This can be achieved by summing the Euclidean distance between the fixations:

$$d_{loc}(s,t) = \sum_{k=1}^{n} d_{euclid}(s_k^{loc}, t_k^{loc}) \tag{4.2}$$

The possibilities mentioned so far (i.e., Manhattan metric and Hamming distance) share the same limitations: (1) they are only applicable to sequences that are of the same length; and (2) they are not sensitive to the similarity of sub-sequences within the sequences being compared if those sub-sequences occur at different positions in the larger sequences. Both of these are serious limitations for the application to scan paths in reading, which can be of arbitrary length, and where repeated patterns that are shifted within the larger sequence are likely to hold information about the cognitive process, parsing, and therefore to be of value.

Another class of sequence dissimilarity (or distance) metrics is the class of *edit-distance* metrics. An edit-distance is the minimum number of edit operations that must be made in order to transform one sequence into another. Permitted operations usually include deletion, insertion, and substitution of symbols. An example of an edit-distance is the *Levenshtein metric* (Levenshtein, 1966). Analyses based on this edit-distance have been applied to scan paths outside the area of reading research. For example Brandt and Stark (1997) analysed eye movement patterns in response to checkerboards with a Levenshtein metric; Josephson and Holmes (2002a, 2002b) used a Levenshtein metric to analyse scan paths in response to viewing web sites; and Cristino, Mathôt, Theeuwes, and Gilchrist (2010) used a Levenshtein metric to analyse eye movements in a visual search task. The Levenshtein metric has the advantage that it can cope with sequences of arbitrary length. It does this by aligning sequences through the use of null elements. In the same way, similar patterns that occur shifted within their sequences can be aligned and their similarity recognised, using for example, the Needleman-Wunsch algorithm, which uses a dynamic programming technique to find those optimal alignments that minimise the dissimilarity of two sequences (Needleman & Wunsch, 1970). However, while the Levenshtein metric yielded a suitable measure for the studies in Brandt and Stark (1997), Josephson and Holmes (2002a), Josephson and Holmes (2002b), and Cristino et al. (2010), where the sequence of locations of fixations was at issue, it is not ideal for reading research where the duration of fixations is equally important. In fact it is possible, but impracticable, to make the Levenshtein metric take account of durations by representing a long fixation with a series of repetitions of the same location. However, in order to achieve sufficient temporal resolution, the number of symbols involved in the repetition method becomes extremely large. Unfortunately the algorithm's scaling properties are $O^{(m*n)}$ where m and n are the length of the two sequences, with the consequence that the approach is impracticable in reasonable time.

Another drawback of the Levenshtein metric for reading analysis is that it penalises spatial difference irrespective of how great the difference is. However, in reading studies, small versus large divergence in fixation location can represent a discrete difference in cognitive processes: for example, this could be the difference between a saccade to the next word that indicates normal processing, and a regressive saccade to the beginning of the sentence that indicates processing breakdown.

A further drawback of the Levenshtein metric for reading research is that, by applying an all-or-nothing penalty to location differences, it does not take account of the gradient of acuity across the human field of vision, due to: the higher density of photoreceptors in the central part of the retina; differences in how photoreceptors converge on retinal ganglion cells; and how they in turn converge in the projection to the lateral geniculate nucleus (Grill-Spector & Malach, 2004). This yields an effect dubbed *cortical magnification* whereby foveally fixated objects are represented in the cortex with higher resolution than parafoveally fixated objects, and thus with greater cognitive acuity. Cortical magnification was measured by Daniel and Whitteridge (1961) and is approximated by the function m^{δ}, where δ is the eccentricity in degrees and m is 0.83. For reading research, a good metric of scan path similarity should take account of the degree of spatial divergence of two fixations in the projection to the cortex, and not merely on the screen: doing this would take account of the difference in visual acuity at the two locations.

The scan path similarity, or SCASIM, metric developed by von der Malsburg (2009, 2010); von der Malsburg and Vasishth (2007, 2008, 2009, 2011, 2012) is a promising candidate for a better metric for reading studies. The method is implemented in the R function SCASIM (von der Malsburg, 2010) which computes scan path similarity. The measure is a specialised kind of edit-distance. It is specialised for spatio-temporal patterns, particularly scan paths, and can cope with scan paths of different temporal and spatial dynamics, unlike other edit distance measures.The measure computes a dissimilarity matrix for all the scan paths in the set.

It is possible to define the distance between two scan paths that have the same length as a function of their fixation locations and durations, as a weighted sum of two terms: the magnitude of the difference between their fixation durations $|s_k^{dur} - t_k^{dur}|$; and the sum of their fixation durations $s_k^{dur} + t_k^{dur}$. The first summand corresponds to the case where the fixations target the same location for different durations, and the second summand corresponds to the case where the fixations target very far apart locations: it represents the cost associated with removing s_k and replacing it with t_k. The weights are derived from the distance between the two fixations in the approximation to visual cortex, using the function m^δ. If $\delta(x, y)$ is the angle between two fixation targets $x = s_k^{loc}$ and $y = t_k^{loc}$, then the overall distance between two scan paths s and t is then:

$$d(s,t) = \sum_{k=1}^{n} m^{\delta(s_k^{loc}, t_k^{loc})} |s^{dur} - t^{dur}| + (1 - m^{\delta(s_k^{loc}, t_k^{loc})})(s_k^{dur} + t_k^{dur}). \tag{4.3}$$

When fixations are far apart, then the longer the fixations are, the larger the penalty should be. If the fixations are both short, the penalty should be small. Summing the fixation durations allows the small penalty for short far apart fixations, and the large penalty for long far apart fixations.

Once a dissimilarity value has been obtained for each scan path, the value can be submitted to conventional analysis of variance like any other scalar measure. However, because participants make the scan paths of interest quite rarely, it is often difficult to achieve a balanced design for conventional analysis of variance. scan path similarity is also amenable to cluster analysis which does not need to assume balance in the same way. A cluster in similarity space is a set of self-similar scan paths that correspond with a reading strategy (von der Malsburg

& Vasishth, 2011). This allows the researcher to ask whether a given reading strategy is overrepresented in a particular experimental condition.

In order to analyse the distribution of reading strategies, it is first necessary to reduce the dimensionality of the pairwise dissimilarity matrix yielded by the SCASIM function. To this end, the dissimilarity matrix returned by scasim is submitted to multidimensional scaling. Multidimensional scaling seeks to represent the large matrix in a lower-dimensional form that is still faithful to the patterns in the original matrix. There are (at least) three options for doing the multidimensional scaling in R: CMDSCALE; ISOMDS; and SAMMON. There are some restrictions which determine which method to use. These restrictions are on the form of the input, which is a distance structure of the kind returned by SCASIM: some methods do allow distances of zero, representing exactly equivalent scan paths, and some do not.

Once a lower-dimensional representation has been achieved, it can be submitted to cluster analysis to yield reading strategies. There are (at least) two options for doing the cluster analysis in R: KMEANS; and MCLUST. For KMEANS, the number of clusters must be specified. However, the function MCLUST computes the optimal number of clusters with reference to Bayesian Information Criterion (BIC), and because it is not obvious what the optimal number of clusters is in this case, the MCLUST function was used here. The function MCLUST computes the optimal number of clusters to fit the data by generating a selection of models, and choosing the one that maximises the BIC. The technical description of the optimal model given in the function's help file is: *The optimal model according to BIC for EM initialized by hierarchical clustering for parameterized Gaussian mixture models... The Expectation-maximization algorithm can be used to compute the parameters of a parametric mixture model distribution... A 'best' model can be estimated by fitting models with differing parameterizations and/or numbers of components to the data by maximum likelihood, and then applying a statistical criterion for model selection. The Bayesian Information Criterion or BIC (Schwarz, 1978) is the model selection criterion provided in the* MCLUST *software. It adds a penalty term on the number of parameters to the log likelihood. For details of model-based clustering, see McLachlan and Peel (2000) and Fraley and Raftery (2002).*

Once reading strategies have been identified by the cluster analysis it then becomes possible to ask what the distribution is of the strategies over the experimental conditions. In the same way one can ask whether a particular reading strategy is over-represented in a particular experimental condition.

4.4 Hypotheses about the purpose of regressions

There are some hypotheses about the purpose of fixations in the regressive path. These are selective reanalysis (Frazier & Rayner, 1982), time out (Mitchell et al., 2008), a perceptual uncertainty account (Levy, Bicknell, Slattery, & Rayner, 2009), and a combined retrieval and surprisal account (Engelmann et al., 2013).

4.4.1 Selective reanalysis

The selective reanalysis hypothesis is the proposal that readers make use of a strategy that enables the sentence processor to exploit "whatever information it has available about the type of error it has committed to guide its reanalysis attempts" (Frazier & Rayner, 1982)

4.4.2 Time-out hypothesis

The time-out hypothesis (Mitchell et al., 2008) proposes that regressions are the result of the linguistic system halting the progress of the eyes into new material before existing material has been fully integrated, in the event that the currently processed material takes longer than usual to integrate. The parser instigates a time-out allowing the integration to be carried out before moving the eyes into new material. The eyes are inhibited from going forwards, and so saccades can only be intra-word fixations on the launch word or inter-word regressions. The time-out hypothesis predicts that the landing sites of these saccades should not be patterned, but should instead follow a distribution over the words preceding that launch site, with higher probability of landing on a word close to, but not further than the launch site.

4.4.3 Perceptual uncertainty account

A different sort of explanation for regressions comes from Bicknell and colleagues (Bicknell & Levy, 2010a, 2010b, 2011; Levy et al., 2009). Their explanation is that there is perceptual uncertainty with respect to what has been read. Regressions in tho framework are made in order to gain information about words that are represented with higher uncertainty, in the face of disambiguation.

4.5 Linguistic and oculomotor co-ordination

An open question is whether we can use the distribution of fixations in regressive paths to adjudicate between theories. The Diagnosis model predicts different processing of NP/S and NP/Z disambiguations and this prediction implies different eye movement distributions in regressive paths. It is also possible to sketch a proposal for how the operations proposed by replacement theories might manifest in the distribution of fixations in the regression path. With these two proposals turned into claims about the distribution of fixations on the regressive path for NP/S and NP/Z sentence types, it is possible to carry out an evaluation against regression paths collected in experiments that manipulate NP/S versus NP/Z disambiguations. There are some assumptions that underly the claims given above about the purpose of regressions. These are the **tight versus loose coupling** possibilities, and the **overt versus covert repair** alternatives. These are discussed next.

4.5.1 Tight coupling

The assumption of tight coupling is widespread, that the linguistic system and the eye movement system are in lock-step, and what is fixated is what the linguistic system is processing. Just and Carpenter treated this as part of their eye-mind assumption when they wrote "the eye remains fixated on a word as long as the word is being processed . . . [S]o the time it takes to process a newly fixated word is directly indicated by the gaze duration" (Just & Carpenter, 1980, p. 331)

4.5.2 Loose coupling

While Frazier and Rayner (1982) saw regressive eye movements as tightly coupled to parser operations, the assumption of tight coupling has been challenged in the literature. For example, Mitchell et al. (2008) argue that a loose coupling explanation can account for the data too. They offer the *time-out hypothesis*, under which the fixation locations in regressive eye movements are partly decoupled from linguistic operations. Such a loose coupling account needs an explanation for the distribution patterns of regression fixation locations. The time-out hypothesis provides a possible explanation: the patterns are held to be spatially determined, with location a function of distance from the launch site with random noise. The time-out hypothesis was considered in Vasishth, von der Malsburg, and Engelmann (2013) where it is used to model short regressions in ACT-R architecture. There are two possible extremes for the role of spatial layout. (1) Spatial layout could have no influence on eye movements in regressions. (2) Influences of spatial layout could overwhelm linguistic influences on eye movements. Because the influence of spatial layout on regressions is largely unknown, common measures that are taken to index parsing processes might turn out to reflect effects of layout instead unless linguistic influences are shown to exist independently of layout effects. Consider that if RPD was shown to be more sensitive to layout than linguistic manipulation, then the whole body of literature that uses RPD to index parsing would become subject to doubt. There is some evidence to be found in the fundamental properties of eye movements in reading of loose coupling between linguistic drivers of eye movements and eye movements themselves (Rayner et al., 2012). For example, spillover effects, where the consequences of disambiguation are not realised in eye movements until the following word or words, indicate that the eyes sometimes move past the word that carried disambiguating information without acting on that information, and also indicate that the word that launches the regression may not be the word that licensed or motivated the regression. Evidence for parafoveal preview benefit indicates that the eyes take in information from a word that is not yet fixated foveally, and that eye movements can be influenced by that word's linguistic properties (e.g., semantic anomaly) which in itself breaks the notion of strict tight coupling. The fact that words are often skipped as a function of the frequency of the word being skipped indicates a further breaking point for the notion of tight coupling.

4.5.3 Overt and covert repair

Another way the link between the linguistic system and the eye movement system can be challenged is by saying that perhaps reanalysis might the covert - that is to say done without observable behaviour. Under these circumstances no amount of eye movement studies would yield information about linguistic processing. There are suggestions of decoupling between syntactic reanalysis and eye movements, as well as reanalysis by covert repair (Lewis, 1998, p. 253). *Overt* repair is the term used for repair when it has an observable behavioural phenomenon to indicate that repair has taken place - typically a set of regressive eye movements - whereas *covert* repair is the term used to indicate that while repair has taken place, it has taken place in the absence of observable behavioural phenomena - for example, a search through memory that is done without moving the eyes. These suggestions serve to raise doubt that regressions are under linguistic control and determined by target-seeking. Such doubts are amplified in Mitchell et al. (2008). If proposals of covert repair and decoupling hold, then modeling regressions in theories of sentence processing is fundamentally misguided and misleading. However it is more likely that control of eye movements at disambiguation is shared between linguistic and non-linguistic controllers. Examples of non-linguistic control include evidence that eye movements are influenced by the spatial details of the print; evidence that the frequency of a word influences the duration of fixations on it; evidence that the overall discourse coherence of the text influences eye movements; evidence that the nature of the task demands faced by the participant influences eye movements, such as whether the participant is proofreading or reading for comprehension (Rayner et al., 2012).

4.5.4 Are the eyes seeking targets or avoiding new information in regressions?

Regressions in the selective reanalysis framework are linguistically guided and target-seeking. Regressions in the time-out hypothesis are subject to particular limited influences from the linguistic system and are launched *away* from the initiating word rather than directed *at* some target. The influence of the linguistic system under the time-out hypothesis is limited to the right to veto and sanction progressive eye movements as part of the decision to take in a new word, a decision shared by the low-level linguistic drivers that respond to, for example, word frequency and which themselves can impose delays. The notion of the parser licensing a progression finds a home in Yang (2006) where a proposal is made that fixation duration is basically the elapsed time while a progression to the next word is inhibited by the linguistic system in the service of time needed to process a word with particular consequences for the linguistic system. The time-out framework can be described as both overt repair and covert repair. Overt because while it withholds the sanction to move on it manifests in the eye movement record as increased duration and possibly increased frequency of regression: covert because the target of regressions is uninformative for the nature of the parsing processes involved - regressing away is considered to be buying time for the parser, and the landing sites of these regressions away are distributed as a function of distance from the launch site with some noise and therefore the identity of the targeted word is irrelevant for information about parsing.

In work on anaphor resolution there is evidence for target-seeking regressions (Inhoff & Weger, 2005; Kennedy & Murray, 1987; Murray & Kennedy, 1988; Weger & Inhoff, 2007), but the task faced by the participant has only superficial similarity with the task faced by the reader of syntactically disambiguating material. Anaphor resolution work also provides examples of under- and over-shooting in target-directed regressions. This finds parallels with the ‘stepping stone’ strategy identified by Mitchell et al. (2008) which they considered to be essentially range-finding in regressions for syntactic disambiguation This suggests that a syntactic mechanism might be target-oriented without necessarily *succeeding* in hitting the target, at least on the first regressive fixation. To explore this Mitchell et al. (2008) also considered all subsequent regressive fixations in the same regression sweep looking for evidence that, after a little range-finding, the eyes succeed in picking out some target. In this analysis they reported some weak evidence that regressions are target-seeking. The evidence for range-finding error implies that a measure that focuses on landing-sites is not likely to find strong evidence for a target-seeking mechanism even if the driver of the movements *is* actually destination-based and target-seeking. One escape from this bind is offered by scan path similarity analysis which abstracts over the precise landing sites to reveal self-similar *patterns* of regression that should still hold up in the face of range-finding error.

In the collection of regression paths for analysis in the evaluation it is important to use materials that control the placement within the sentence of the ambiguous and disambiguating parts. Using materials in which the disambiguation region is placed adjacent to the ambiguous region makes it impossible to distinguish between fixations that are made in the ambiguous region, and regressions that are made in the region adjacent to launch. Interposing material between the onset of the ambiguous material and the disambiguation region allows this distinction to be made.

4.5.5 Can we distinguish repair from replacement theories using regressions?

One response to problematic disambiguation is *repair*. Repair accounts are characterised as those that make use of previously-built structure to effect a new analysis. They start from the analysis that disambiguation reveals to be a wrong analysis, and make changes to that analysis until it fits the new input. There are several variants of the repair proposal, according to how it is proposed that the repairs are identified and applied. Constraints applied to the repair process vary across particular accounts. These accounts are treated separately below.

Another response to problematic disambiguation is *replacement*. Replacement parsers are those that maintain several parses in parallel, and update the representation of the sentence so far in response to disambiguating input by replacing the top-ranked parse with another from the set.

Does the human sentence processing mechanism carry out repair or replacement to fix a temporarily syntactically ambiguous partial parse that turns out to be inconsistent with disambiguation? A sign that repair is used would be if eye movements at disambiguation include purposeful visual search for text that has already been read. Such movements are unnecessary, and difficult to account for, under a replacement hypothesis. If there are no such patterned eye movements at disambiguation this suggests that linguistic operations at disambiguation can be accomplished without taking in again any of the previously-parsed material, and this would be consistent with a re-ranking of partial parses that are still available to the linguistic system.

4.5.6 The link from regressions to disambiguation

Most models of eye movement control in reading operate below the level of syntactic processing. This makes it difficult to make inferences from eye movements to parsing processes. In fact the link between syntactic disambiguation and regressions is not well demonstrated. There is theory for the linguistic operations that might be involved at disambiguation. For example, various linguistic theories propose that parse repairs are mandated at unexpected disambiguation (Ferreira & Henderson, 1991b, 1993; Fodor & Inoue, 1994, 1998; Sturt, 1998; Sturt & Crocker, 1998; Sturt et al., 1999; Van Dyke & Lewis, 2003). This theory work does not specify how repairs are implemented in regressions, although the following papers do take up this question: Frazier and Rayner (1982); Meseguer et al. (2002); Mitchell et al. (2008); von der Malsburg and Vasishth (2011).

Regressions in the selective reanalysis framework are linguistically guided and target-seeking. Regressions in the time-out hypothesis are subject to particular limited influences from the linguistic system and are launched *away* from the initiating word rather than directed *at* some target. The influence of the linguistic system under the time-out hypothesis is limited to the right to veto and sanction progressive eye movements as part of the decision to take in a new word, a decision shared by the low-level linguistic drivers that respond to, for example, word frequency and which themselves can impose delays. The notion of the parser licensing a progression finds a home in Yang (2006) where a proposal is made that fixation duration is basically the elapsed time while a progression to the next word is inhibited by the linguistic system in the service of time needed to process a word with particular consequences for the linguistic system. The time-out framework can be described as both overt repair and covert repair. Overt because while it withholds the sanction to move on it manifests in the eye movement record as increased duration and possibly increased frequency of regression: covert because the target of regressions is uninformative for the nature of the parsing processes involved - regressing away is considered to be buying time for the parser, and the landing sites of these regressions away are distributed as a function of distance from the launch site with some noise and therefore the identity of the targeted word is irrelevant for information about parsing.

4.5.7 Subpopulations of regressions

An interesting possibility and a further complication for using regressions as evidence of parser operations is raised by Mitchell et al. (2008, p. 270), where they point out that “It is perfectly possible for one sub-population of regressive movements to be generated by one kind of mechanism and the rest by another”. One consequence of this possibility is that researchers who treat regressions as a homogeneous population might be falling into the trap of treating a multi-modal distribution as a uni-modal one. Under these circumstances such researchers might commit a ‘failure to find’, or Type 2 error when their statistical test cannot detect differences in the data because the test is rendered insensitive due to the multi-modality. It is worth considering that the probability of making a regression (of any type), which is a measure in common use, cannot distinguish these subpopulations, and that the Frazier and Rayner qualitative descriptions of subpopulations in regressions did not give rise to a quantitative measure that has the potential to provide empirical underpinning. An important recent development that seeks to provide quantification, empirical underpinning, and associated statistics for the consideration of subpopulations of regressions is the work of von der Malsburg (2010); von der Malsburg and Vasishth (2011, 2012) on a scan path similarity measure called SCASIM. Here its important characteristics are that it is an edit-distance metric that quantifies the similarity of two scan paths, and which therefore permits a clustering of scan paths to be identified in a data-driven way. If the proposal holds that there are coherent and distinct subpopulations of regressions, then given sufficiently non-sparse data, then these clusters should identify the subpopulations. Before the SCASIM scan path similarity measure had been developed, Mitchell et al. (2008) could not avail themselves of it, and relied instead on examining the distribution of the landing sites of regressions. This was done in an attempt to test whether the patterns of target-seeking predicted by selective reanalysis were present in the data. The logic of this approach is that if regressions are indeed targeting particular locations then this would support destination-oriented target-seeking behaviour like that indicated in the selective reanalysis hypothesis. In the other outcome where the distribution of landing sites did not pick out particular locations and regressions were instead distributed as a function of distance from the launch site with some random error component, the this would support launch-based regression in the manner predicted by the time-out account.

Chapter 5

Statistical analysis tools

This chapter introduces the statistical tools used in the thesis. This charts the recent move away from using classical Analysis of Variance (ANOVA) with separate by-participants and by-items analyses towards using a unified regression model (LMER) that includes the random effects of participant and item as well as other covariates (e.g., word length, word frequency) in a single model. Advantages and disadvantages of the LMER framework are laid out.

5.1 Introduction

Consider the data that are yielded by a typical eye tracking study of reading. Such experiments measure the performance of individual participants in some sentences from each of four cells derived from the factorial manipulation of two treatment variables. Across an experiment, each sentence is responded to the same number of times in each of the four conditions by the same number of readers. No one reader is exposed to the same sentence in more than one condition. In practice this is achieved by implementing a *split plot* design that was familiar to farmers in the nineteenth century who sought from it a way to rotate crops over fields in the most productive way. Data from these split plot designs have a multilevel structure that can be decomposed into systematic and random components.

The systematic structure is the structure imposed by the factorial manipulation of independent variables. It represents the notion that in a 2 x 2 factorial design there are four treatment conditions, and that the data can be grouped according to which condition yielded them. The systematic structure is responsible for some part of the variability observed in behaviour. For example in a temporarily ambiguous sentence, the disambiguating word will yield different behaviour as a function of whether there is still syntactic ambiguity at the point at which the word is processed. A model that is faithful to the systematic structure of the data may be given as follows. Where X_{ijk} represents any observation, μ is the grand mean of the observations, α_i is the effect of Factor $A_i = \mu_{A_i} - \mu$, β_j is the effect of Factor $B_j = \mu_{\beta_j} - \mu$, $\alpha\beta_{ij}$ represents the interaction effect of Factor A_i and Factor B_j, and ϵ_{ijk} represents the unit of error associated with observation X_{ijk}

$$X_{ijk} = \mu + \alpha_i + \beta_j + \alpha\beta_{ij} + \epsilon_{ijk} \tag{5.1}$$

Within a given cell of the systematic structure, there is variability in behaviour. This is attributable to the differences between the individuals who contributed the behaviour, and to differences between the sentences that yielded the behaviour, and to other sources. Such variability is described in the random structure of a statistical model. In a typical psycholinguistic reading experiment, participants and items are crossed so that each participant reads one version of all sentences, but never the same sentence in more than one version. Similarly each item yields responses from all participants, but an item in a given version is responded to by only some readers, and never by the same reader more than once.

Until the advent of more advanced techniques, such data were submitted to ANOVA. Doing so violated at least one of the assumptions of ANOVA: that for conditional independence of observations. The problem alluded to is when observations in any given cell of the design are treated as though they were independent of observations in some other cell whereas in fact some of them come from the same sentence or participant. Any experiment that uses repeated measures violates the assumption of independence of observations. The traditional response has been to aggregate over the dependent cases to yield single values that are then conditionally independent. This is undesirable because the variability lost in the aggregation can be properly accounted for by modern methods, and may be of direct theoretical significance.

Extra errors are committed when data that are binomially- or poisson- distributed are transformed to proportions and then treated as gaussian-distributed, ignoring the hard floor at zero and the hard ceiling at 1. One problem with this is that confidence intervals for low mean values can encroach into negative proportions which are undefined, or into proportions greater than one which are similarly undefined. Another is that backing off from zero and one in order to remedy this problem by arcsine transformation and then modeling the arcsine value is unlikely to achieve homogeneity of variance in conditions in practice. Logistic regression is one preferred framework. We shall see how this approach can be extended to cope with the random structure of our data.

The statistical models implemented in LMER are well-suited to analysing eye tracking data. LMER copes naturally with data that constitute samples from any distribution (binomial, poisson, gaussian). This enables e.g., binomial outcomes to be appropriately modelled in the raw without suffering by aggregation. The systematic structure of the data is readily represented in these models by using terms for each main effect and each interaction effect. The random structure is also readily represented in such models (where observations are nested within individuals for example). A further benefit is that all this can be done simultaneously in LMER without recourse to separate F_1 and F_2 analyses.

The history of the analysis of variance has important developments brought about by approaches described in articles dated 1973 and 2008. Clark (1973) demonstrated a way round the problem of the fallacy of language as fixed effect, achieved by modeling by-participant variance and by-item variance separately and then taking $minF'$ as a composite measure. The next big shift in the history of the analysis of variance is well captured in papers combined in the 2008 special issue of Journal of Memory and Language (Forster & Masson, 2008). This shift used a single analysis to model variance arising from any of several different sources, and at any of several different levels. Thus it was possible to account separately, but in the same model, for the variance in a given measurement across trials that was due to the item presented on that trial, as well as variance that was due to the participant being presented the trial, leaving cleaner accounts of the variance due to treatment effects. Research into this methodology is approaching maturity at the time of writing. While it is currently fit enough for purpose that it is used in the thesis, we will also see that it is sufficiently immature that there are still serious disagreements on some fundamental practical issues.

Early history Stigler (1986) indicates that astronomers before 1800 had developed methods (i.e., the *personal equation*) to identify and reduce observational errors in astronomical measurements, where errors were those introduced by observers. This can be taken as the origin of the analysis of variance because errors were identified in terms of the extent to which an observation varied from others of the same phenomenon. The argument that significance testing should be carried out for experiments in verbal behaviour was reiterated by Coleman (1964). His reasoning was that it was desirable to know how far beyond the linguistic sample the findings from experiments could be generalised, and that without significance testing the findings from contemporary experiments on verbal behaviour could not be generalised beyond the sample of linguistic materials employed. He gave advice in terms of contemporary statistical methods, including recommending F'.

ANOVA and the F-ratio A standard method used to analyse eye movement data is ANOVA (Fisher, 1918, 1921, 1925; Howell, 2009), adapted to take separate account of both subjects effects and items effects by Clark (1973). ANOVA compares the means of different experimental conditions and decides whether to reject the null hypothesis that the means do not differ, given the observed sample variances within and between the conditions. Simple one-way ANOVA can be described as follows. The F statistic is calculated as the ratio of the variance explained by the model to the variance unexplained by the model.

$$F = \frac{\text{explained variance}}{\text{unexplained variance}} = \frac{\text{variance between subjects (or items)}}{\text{variance within subjects (or items)}} \tag{5.2}$$

The formula for assigning values to these quantities is as follows. Where $\bar{Y_i}$ is the sample mean in the i$^{\text{th}}$ group, n_i is the number of observations in the i$^{\text{th}}$ group, $\bar{Y}$ is the overall mean of the data, Y_{ij} is the j$^{\text{th}}$ observation in the i$^{\text{th}}$ out of K groups, K is the number of groups, N is the overall sample size:

$$F = \frac{\sum\limits_{i} n_i \left(\bar{Y_i} - \bar{Y}\right)^2 / (K-1)}{\sum\limits_{ij} \left(Y_{ij} - \bar{Y_i}\right)^2 / (N-K)} \tag{5.3}$$

Limitations of ANOVA There are several reasons why ANOVA may be unsuitable for statistical inference over eye movement data, particularly over regression signature data, but also over the standard measures.

Firstly, for regression signature data, which are categorical data (whereas the standard measures constitute continuous data), we are dealing with proportions calculated over a categorical variable, that being the proportion of a binary valued presence / absence of a fixation in a given word, when we are considering the location of the very *first* regressive fixation of a regressive sequence; and the proportion of a number of discrete instances of a fixation in a given word, when we are considering the location of *all* regressive fixations in a regressive sequence. For continuous data, the interpretation of mean values, variance, and confidence intervals can be clear. For categorical data however, confidence intervals can extend beyond the permissible values in the data of zero and one. It is for this reason that Jaeger (2008) argues that ANOVA may lead to spurious results, or at the very least, problems of interpretation.

Secondly, eye movement data exhibit crossed random effects of participant and item. ANOVA can cope with this in Latin square designs by providing separate by-participant F_1 and by-item F_2 analyses of aggregated measures. LMER improves on this by accounting for both participant and item level variance in the same model.

Thirdly, the assumption of independence between observations (Barr, Levy, Scheepers, & Tily, 2013; Coco, 2011) presents a problem. This is violated in the case of repeated measures in any given cell of the design. Traditional ANOVA deals with this by aggregating each observation in the cell and then submitting the aggregate value to analysis. In LMER the data reduction caused by the aggregation is unnecessary provided that the random effects structure in the design is specified in the model.

Fourthly, the multilevel structure of the data (Richter, 2006); (Coco, 2011, p. 29) presents a problem. This has more impact in the example where students are nested in schools that are themselves nested in states, all of which exert an effect at their own level that contributes to the value of the measurement at the lowest level of the student.

The need to account for multilevel effects Following on from Clark (1973), work in an age of vastly increased desktop computer power has yielded a further framework for statistical analysis of data from psycholinguistic designs, where linguistic items are sampled from a population of linguistic items, and participants are sampled from the population of people. This framework is set out for psycholinguists in papers collected in a special issue of the Journal of Memory and Language (Forster & Masson, 2008). The framework is variously called *mixed effects modeling, hierarchical modeling; multilevel modeling*, all referring to the notion that sources of variance exist at more than one level in the data. For example, when a given trial in particular levels of the treatment variables yields a response time with a particular value v, the value depends on the particular participant, the particular linguistic item, and the treatment effect(s) with some error attributable to the natural variability of response times. In the case that the reader is a slow reader, his slowness will increase the value of v independently of the treatment effect(s), so to attribute the values of repeated measurements of v to the treatment effect(s) is to attribute too much to the treatment effect(s). A proper treatment would take into account item-level and participant-level variance in the estimation of coefficients for the levels of the treatment effect(s). So far, $minF'$ fits the bill, but multilevel models go further. To see why, we will need to refer to the details of the General Linear Model, to see how it can be extended to cover the demands set out in this paragraph.

5.2 LMER

In this section I deal with the recent introduction of linear mixed effects models into psycholinguistic data analysis.

In recent years there has been a movement in psycholinguistics away from ANOVA and towards linear mixed effects regression models (LMER). This movement is captured in the Journal of Memory and Language Special Issue on Emerging Data Analysis (Forster & Masson, 2008). In this thesis I use the Generalized Linear Mixed Effects Regression Model (GLMM) framework to carry out inferential statistical analysis, particularly the Linear Mixed Effect class of models, with random effects for subjects and items (Baayen, Davidson, & Bates, 2008; Pinheiro & Bates, 2009), implemented in the R package LME4 (Bates, Maechler, & Bolker, 2011) that provides the function LMER.

The simplest model of regression is a linear model assuming a linear relationship between the observed response variable (e.g., regression path duration) and the explanatory variables (e.g., ambiguous / unambiguous). The explanatory variables are expressed as regression coefficients β which give information about the strength and direction of their linear effect on the response variable y. A linear model is a collection of coefficients β_i, one for each explanatory variable i and one term for each of i's interactions. This yields

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_i x_i \tag{5.4}$$

where β_0 is the mean of the response variable and β_i is the contribution of the ith variable to the value of the predicted variable. The sign of the coefficient is a measure of the direction of the effect the predictor exerts on the predicted variable: a positive coefficient signals that the predictor increases the value of the response variable, and a negative coefficient signals that the predictor tends to reduce the value of the response variable. The value of the coefficient is a measure of the size of the effect, with larger values signalling larger effects.

In a GLM, the linear relationship between the response variable and the explanatory variables can be established for different distributions by using different link functions (e.g., Poisson, logistic, etc). When dealing with continuous variables (the standard scalars), a gaussian link function is appropriate: when dealing with regression signatures, which are proportions calculated over binary outcomes, a logit link function is appropriate. The logit function is the logarithm of the odds of the response variable

$$logit(y) = \log(\frac{y}{1-y}) \tag{5.5}$$

where the odds of event a is an expression of the ratio of the probability of a occurring to the probability of a not occurring. In regression signatures, a is a fixation in the word or region under consideration. The logit is the odds on a logarithmic scale, yielding a value that may be treated as normally distributed. LMER has the virtue that it can distinguish between random and fixed effects in a way that GLM cannot. Participants and items constitute random effects in our experiments, and LMER treats them appropriately. This yields

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_i x_i + b_1 x_1 + \ldots + b_j x_j + \epsilon \tag{5.6}$$

where β_0 is a term for the intercept, β is a fixed effect and b is a random effect, i indexes participants, j indexes items, and ϵ is the error term that we seek to minimise.

The justification for the use of LMER is that LMER allows us to quantify the strength and direction of the contribution towards the value of the response variable of each explanatory variable, taking proper account of whether the explanatory variable exerts a fixed or a random effect, and taking proper account of different multilevel components.

One scenario in which LMER is useful is when there are very many factors that represent candidates for inclusion in the best model of the data. In this case, it is important to have a principled way to decide which terms should be retained in the eventual model. Complex experimental designs yield many different models, and many of these can be equally good at fitting the data. Under these circumstances, one (flawed) approach is to set the explanatory variables in a pattern that fits the experimental design. However, to do this is to test a unique hypothesis, and to assume wrongly that no other hypotheses are possible. An alternative approach is to assume that we are sampling from a space of models, and through a bottom-up exploration of the data we bootstrap the best hypothesis.

There are two main approaches to this bootstrapping in model selection, whereby explanatory variables are included or removed one at a time. In forward selection, the final model is built from a minimal model to which explanatory variables are added if they are significant until no further improvement in goodness of fit is yielded. In backward model selection, one starts with a fully-specified model, and removes predictors that exert non-significant effects to yield an efficient model.

In both cases a measure of goodness-of-fit is required. LMER uses the log-likelihood test to provide the measure. A series of pairwise comparisons tells us when to stop adding / taking way predictors. We start by removing random effects first, then fixed effects, then interactions.

LMER can also be used in cases where the model specification is known in advance. The design of a psycholinguistic experiment represents the terms that will be fixed effects, and the participants and items provide the random effects at each level of which the fixed terms may be arranged systematically. In these cases the purpose of LMER is to tell us whether our fixed effects exerted significant influences on a response variable bearing in mind the contributions to the response variable of idiosyncrasies of participants and items.

In the years intervening between the JML special issue (Forster & Masson, 2008) and the time of writing, there has been some inconsistency in the specification of LMER models, noted by Barr et al. (2013). In the worst cases, no model specification is given by the authors, which means that we have no basis on which to evaluate claims of significance. Even in cases where model specification is given, the specified model is often a random intercepts only model characterised by increased power without increased protection against Type 1 error (i,e., the error made when one incorrectly rejects a null hypothesis that is actually true). This leads to much the same situation described in Clark (1973) where we cannot be sure if the null hypotheses rejected in these studies were correctly rejected. Since it is still not common practice to publish raw data (although see `http://read.psych.uni-potsdam.de/pmr2` for a psycholinguistics repository of raw data and analysis scripts), it is not possible to formulate more appropriate models to facilitate testing whether null hypotheses were incorrectly rejected.

This leads to the reasonable question “How should we choose a multilevel model specification that increases power *without* increasing Type 1 error rates?”. This is the question taken up by Barr et al. (2013). They report a series of permutation tests of several model specifications, and compare their performance on simulated data in terms of their respective Type 1 error rates at several given levels of α. This gives us an empirical basis on which to prefer a given model specification. Here I note that the best performing model specification that emerges from this testing is the *maximal model*, and that random-intercepts-onlyspecifications perform worse even than F_1 in isolation.

The maximal model is defined as a model in which all sampling units get a random intercept; any factor gets a by-unit random slope if it is both within-unit and has multiple observations per level per unit; and any interaction term gets a by-unit random slope in the case that all factors are within-unit, and there are multiple observations per unit or cell, where a ‘cell’ is a combination of factor levels (Barr et al., 2013).

Advantages of maximal models Maximal models survive the intense scrutiny of Barr et al. (2013), who identify several reasons for preferring maximal LMERs to ANOVA, listed here: (1) They have greater power than $minF'$, especially for data with relatively small variance. (2) They can better handle unbalanced data. This is a real benefit when the observation at issue is of a phenomenon that occurs rarely in the data, such as measures conditional on a regressive eye movement having been launched from a disambiguating word, where ANOVA would insist on data being *discarded* in the case that they are not balanced in respect of the design. (3) They offer a principled alternative to the improper application of ANOVA to count data and categorical data (like PREG): maximal LMERs model such data properly by allowing values only to approach zero and one. A thorough elucidation of this point may be found in Jaeger (2008). (4) Continuous predictors can be accommodated naturally, in contrast to ANOVA.

Interpreting LMER coefficient significance The LMER function in R's LME4 package does not compute p values for the fixed effects in a mixed model. The function's author Doug Bates defends this choice on the grounds that relevant necessary information is not precisely available in a LMER analysis. This leaves us with the question how we should evaluate the fixed effects for significance.

One approach is to estimate the degrees of freedom: this is the approach taken by Kuznetsova, Brockhoff, and Christensen (2013). This makes use of Satterthwaite's approximation for degrees of freedom.

Another approach is to compare a model that includes the term against a model that is identical in all other respects but lacks the term of interest. This constitutes fitting the model with and without the term at issue and seeing whether the model with the term included is better with reference to some information criterion that penalises appropriately the additional free parameter in the model with the term at issue included. Suitable information criteria include Bayesian Information Criterion (BIC); Akaike's Information Criterion (AIC). Barr et al. (2013) found that model comparison outperforms the t as z strategy, described in the next paragraph, in their permutation testing.

Another approach is to treat model t as though it was model z. This constitutes treating the t-statistic as if it were a z-statistic, i.e., using the standard normal distribution as a reference. An absolute t-value of 1.96 or greater is taken to indicate statistical significance at $\alpha = 0.05$. The t-values in a mixed-effects models are only approximations because determining the exact degrees of freedom is non-trivial (Gelman & Hill, 2007). Boston et al. (2008) used the t as z approach, citing Gelman and Hill (2007) as their authority. The analyses in this thesis contain a fairly large number of participants and items, and only a few fixed and random effects are estimated. Following Angele and Rayner (2012) I assume that the distribution of the t values estimated by each LMER approximates the normal distribution. Therefore I use the two-tailed criterion $|t| \geq 1.96$ which corresponds with a significance test at the 5% level of α. In the case of binomially distributed dependent variables, the z values are interpreted in exactly the same way.

Chapter 6

Overview of experiments

This chapter describes the type of syntactic ambiguity that is used throughout the thesis. Next there is an overview of each of the seven experiments in the thesis that shows how the syntactic ambiguity is manipulated in each experiment, together with a description of other manipulations used in the experiments.

The experiments in this thesis are all concerned with some form of complement ambiguity (shown in Table 6.1).

Experiment One uses the NP/Z variant to provoke regressions caused by resolution to the dispreferred Z option. These regressions are constrained to be launched from different spatial locations on screen. A strict interpretation of the Selective Reanalysis hypothesis predicts that the difference in spatial location should not influence patterns of regression from disambiguation since it is the linguistic content that is purported to guide regressions under that hypothesis. The experiment shows that layout affected regression probability and scan path distribution but not the other standard measures of parsing. This experiment draws its raw data from Experiment 1 in Mitchell et al. (2008) but the LMER analyses and scan path analyses are novel.

Experiment Two uses the same materials as in Experiment One (the NP/Z form), but implements a different manipulation on layout. Spatial layout affected probability of a regression and distribution of regressive scan paths, but not the other standard eye movement measures measures of parsing.

Experiment Three uses the same difficult NP/Z form of the complement ambiguity. The spatial location of the misanalysis area was manipulated so that there were sentences with the misanalysis area located late, near to the disambiguation, and sentences with the misanalysis area located early, and thus further away from the disambiguating verb. A strict interpretation of the Selective Reanalysis hypothesis predicts that regressions should be drawn to the linguistic content of the misanalysed words, and thus that there should be a misanalysis location effect on regressive scan paths. There was only a non-significant tendency for regressive scan path shape to be influenced by the location of misanalysis. This experiment drew its raw data from experiment 2 of Mitchell et al. (2008) but the LMER analyses and scan path analyses are novel.

Experiment Four introduces verbs like *noticed* that have NP/S form and compares them with verbs that have NP/Z form). These verbs with NP/S form are paired with overt complementisers to produce pre-disambiguated forms that select the sentential complement reading. Commas are used to pre-disambiguate the NP/Z forms to Z as before. This experiment shows that the extra difficulty of disambiguating the NP/Z forms shows up in longer reading times, greater regression frequencies, and also the relative over-dispersal of scan path patterns that target the onset of ambiguity.

Experiment Five maintains the comparison between NP/Z forms and NP/S forms that yielded the relatively greater disambiguation effects in the previous experiments. This experiment focuses on the NP that serves as the alternative possible structure in the comparison. The NP is extended and its contents are manipulated. The NP is expanded to include a qualifier which is placed either before or after the bare noun at the head of the NP. On a repair view, lengthening the NP like this increases the amount of material that must be moved as a unit to effect a successful parse following an initial misparse. On a replacement view, lengthening the NP like this means the ultimately-correct parse is unlikely to be highly-ranked at disambiguation.

In this expanded NP, placing the noun before the qualifier effectively increases the distance from the noun the head of the misattached noun phrase to the disambiguating verb, and placing the noun after the qualifier decreases this distance from disambiguation. The experiment shows that this relatively small manipulation of distance from disambiguation does not exert a consistent enough effect

Table 6.1: Overview of the experiments in the thesis[a]

	verb and licensed complements				ambiguity	sentence type	other manipulation
1	watched (,)	**Z**	NP	-	✓	×	layout
2	watched (,)	**Z**	NP	-	✓	×	layout
3	watched (,)	**Z**	NP	-	✓	×	misanalysis location
4	saluted (,)	**Z**	NP	-	✓	✓	-
	noticed (that)	-	NP	**S**			
5	saluted	**Z**	NP	-	×	✓	head position
	noticed	-	NP	**S**			
6	saluted	**Z**	NP	-	×	✓	head position
	noticed	-	NP	**S**			

[a] Abbreviations. Header line: *verb and licensed complements* = example verb from each condition followed by all the complement-types that the verb licenses - boldface is used to indicate the complement relation that was used in the correct analysis; *amb* = whether temporary syntactic ambiguity was manipulated; *sentence type* = whether sentence type was manipulated; *other manipulation* = if there was another manipulation it is given here. Abbreviations: in table: *Z* = no complement; *NP* = noun phrase direct object complement; *S* = sentential complement. (,) indicates that disambiguation was by punctuation, (that) indicates that disambiguation was by complementiser; [...] (,) indicates that some material intervened between the verb and the comma

to achieve significance, although there are trends towards an over-dispersion of search-for-onset regressions.

Experiment Six takes as its starting point the observation that the previous experiment found non-significant trends for increasing the distance from disambiguation to increase indicators of reading difficulty, and asks whether the manipulation was not extreme enough to demonstrate any real effect that there might be of the manipulation. This experiment tests whether those non-significant trends are significant when the manipulation of distance from disambiguation grows more extreme. distance from disambiguation is increased by lengthening the qualifier, and varied by putting the qualifier before or after its head noun. Similar trends were observed but did not reach significance even when the manipulation was more extreme, suggesting either that this is not a real effect, or that the effect varied so much over subjects and items that its significance was washed out.

Chapter 7

E1: Spatial layout affects regression behaviour

This chapter presents novel computational parser simulation results for the materials used in the eye tracking study reported in Experiment One in Mitchell et al. (2008). The eye tracking data from that experiment are submitted to mixed effects statistical models that were not in common use at the time of the original experiment. The data are also submitted to scan path analysis, revealing an over distribution of linguistically targeted regression scan paths when disambiguation is on the same line as the ambiguous material versus on the line below.

7.1 Introduction

In work on regressions in reading, a common assumption is that selection of saccadic landing-sites is linguistically supervised. This assumption is implicit in the *selective reanalysis* hypothesis first set out in Frazier and Rayner (1982). A logically possible alternative is that, in regressions, what determines saccadic landing-sites is a combination of low-level spatial properties of the landing sites, such as their proximity to the launch site. This alternative is set out as the *time-out* hypothesis in Mitchell et al. (2008) and later in Vasishth et al. (2013).

The importance of the distinction between selective reanalysis and time-out lies in the consequences of the widespread uncritical acceptance of the selective reanalysis hypothesis. The standard eye movement metrics were premised on this hypothesis. Particularly, the regression path duration measure assumes that regressions are under linguistic control. Regression path duration has been used as a yardstick for syntactic integration difficulty to evaluate theoretical claims in the literature. If the assumption is flawed, that would mean that conclusions drawn from regression path duration comparisons in the past would be unsound. So it is important to test the null hypothesis that regressions are governed by spatial properties of the text, something which had not been done prior to Mitchell et al. (2008).

A problem with the materials in Frazier and Rayner (1982) was that the onset of ambiguity was also launch-adjacent. In the example below, the onset of ambiguity at *a mile* is adjacent to the disambiguating word *seems*:

(7.1) Since Jay always jogs a mile seems like a very short distance to him.

A problem with these materials is that returns to *a mile* could be due to its linguistic properties under selective reanalysis (it is the ambiguous region) or, under time-out, could be due to its spatial properties i.e., proximity to the disambiguation. In order to have power to discriminate between the two hypotheses, materials must separate the onset of ambiguity from launch-adjacent material. This was done by interposing an unambiguous relative clause as below. Unambiguous versions (disambiguated by punctuation) were used to serve as a baseline for measuring ambiguity effects.

(7.2) While those men hunted(,) the moose that was sturdy and nimble hurried into the woods and took cover.

With these materials it is possible to distinguish between regressions to the onset of ambiguity at *the moose* under selective reanalysis and regressions away from the launch word *hurried* in time-out.

A strong version of the selective reanalysis hypothesis predicts that since regressions are under linguistic control, regressions should not vary as a function of line break, but only as a function of ambiguity. A strong version of the time-out hypothesis predicts that, since regressions are governed by spatial layout, they should be influenced both by ambiguity, and also by proximity to the launch site. Selective reanalysis predicts returns to *the moose*. Time-out predicts returns to the relative clause, with relatively more returns aimed at *nimble*, the most launch-adjacent word.

A first experiment showed some support for the time-out prediction of an interaction ambiguity x line break: the overall probability of making a regression in the ambiguous versions changes significantly, by as much as 4 times, as a function of line break position, the way in which the material is laid out.

The presence of ambiguity effects mitigates against the time-out hypothesis - if regressions are not under linguistic guidance, how could ambiguity affect their landing sites?

In a second experiment, we varied the location of the onset of ambiguity so that it was either *early* – near the start of the sentence, or *late* – disambiguation-adjacent. In both cases, the disambiguating word is preceded by words with the same number of characters with a tolerance of two characters.

(7.3) While those men hunted, the moose that was sturdy and nimble hurried
into the woods and took cover

(7.4) One sole hiker spotted that while those men hunted, the moose hurried
into the woods and took cover

The results showed that regressions can be placed under some kind of linguistic control (against time-out, there was a tendency for saccades to land in the misanalysis area, wherever that was) but also that a strict selective reanalysis (saccades directly to a linguistically informative word) was not borne out at conventional levels of significance. The Mitchell et al. (2008) experiments found some evidence of a *stepping stone* strategy whereby regressive saccades did not target critical words on the first saccade, but landed on intermediate words on the way back to landing on the critical word. This can be thought of as an intermediate position between selective reanalysis and time-out, and this position was recommended by the authors. Under this proposal, regressions are launched in response to the linguistic system vetoing the move to the word after disambiguation on the grounds that first-pass parsing has not been successful in time to license a move to new material. Spatial factors influence the landing sites of regressive saccades but are not the sole driver of saccade landing site selection.

This chapter uses the data from Experiment One of Mitchell et al. (2008). I submit the data to a mixed effects analysis rather than the F_1 and F_2 analyses that were reported in that paper. I also report novel simulation results from running implemented computational parsers on the same materials that the human participants read. The computational simulations used the following parsers: the HUMDEP version (3.0) parser (Boston, 2013) and the TDPARSE parser (Roark, 2013).

7.2 Method

This section gives details of participants, apparatus, materials and procedure.

Participants This is quoted from the original paper: “Twenty-eight volunteers participated in this study. All were students or employed at the University of Exeter and aged between 18 and 38. Six were male and 22 were female. They were either paid 3 to 6 pounds or granted equivalent course credit for their participation if they were first year Psychology students. In total, eight of the participants were paid in cash. The whole experiment normally took up to 30–50 min depending on the reading speed of different participants.”

Apparatus This is quoted from the original paper: “The apparatus was an EyeLink II eye tracker developed by SR Research Ltd., connected to two Dell computers. The eye tracker has a head mounted system with two miniature cameras mounted on a comfortable padded helmet and an extra camera in the middle of the helmet to determine the central position of the head. The two eye cameras allowed binocular eye tracking with built-in illuminators in each of them. The screen was set to the resolution of 800*600 pixels and the top left of the first letter of each sentence was located at screen coordinate (6, 218). The experimenter’s computer was equipped with the EyeLink II setup and control programme so that all the calibration and validation could be controlled through this screen.”

Materials The materials concentrate on the disambiguation of complement ambiguity. An example of the materials follows below (7.5): the disambiguated version includes the comma that appears in parentheses. A correct analysis of the sentence is given in Figure 7.2 below. A typical misanalysis where the initial clause is extended too deeply is given in Figure 7.3 below.

(7.5) While the mob watched (,) the juggler who was gifted and nimble
swallowed a silver sword that was very sharp.

The design crossed two factors. These were *ambiguity* (the sentence either was or was not disambiguated by a comma immediately following the first verb) and *layout* (the disambiguating verb either appeared at the end of the first line or at the beginning of the second line).

Under the selective reanalysis hypothesis (Frazier & Rayner, 1987) the manipulation of layout should not have any effect on regressions from disambiguation, because it is the linguistic properties of the earlier words that attract regressions under that hypothesis and not their spatial position. Under the assumption that oculomotor constraints exert effects on sentence processing, it is expected that regressions will be less attractive as a strategy for the parser if the regressions involve a move to the line above rather than a move leftwards within the same line. Under this proposal one would expect that regressions to the line above are less frequent even for the same linguistic content.

```
While the mob watched the juggler who was gifted and nimble swallowed
a silver sword that was very sharp.

While the mob watched the juggler who was gifted and nimble
swallowed a silver sword that was very sharp.

While the mob watched, the juggler who was gifted and nimble swallowed
a silver sword that was very sharp.

While the mob watched, the juggler who was gifted and nimble
swallowed a silver sword that was very sharp.
```

Figure 7.1: Illustration of how the spatial layout manipulation was implemented. The conditions were as follows, in the same order as the illustration: *ambiguous, late line break*; *ambiguous, early line break*; *pre-disambiguated late line break*; *pre-disambiguated early line break*

Figure 7.1 indicates how the materials were broken over lines. The full list of materials is given in Appendix A. The actual materials used a monospaced font so the illustration of layout below uses a monospaced font and is faithful to the relative positions of the words in the materials. The disambiguating word is underlined. The materials were all NP/Z ambiguities that resolved to the Z reading. The experimental materials were included with 48 filler sentences that were not ambiguous and an equal number (24) of foil sentences that were NP/Z ambiguities that resolved to the NP reading.

Procedure This is quoted from the original paper: “On entering the test cubicle, each participant was asked to put on the eye-tracking helmet. One of the EyeLink cameras was directed at the participant’s right pupil. At the beginning of the session approximately 10 min was set aside for tracker calibration. The experiment proper was started only after calibration and validation was classi-

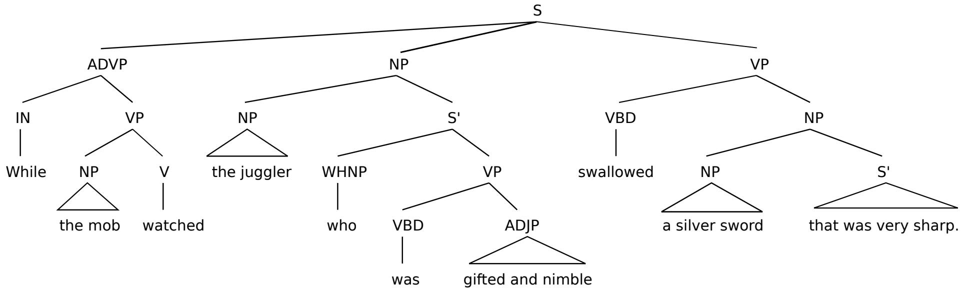


Figure 7.2: A correct analysis of Example 7.5

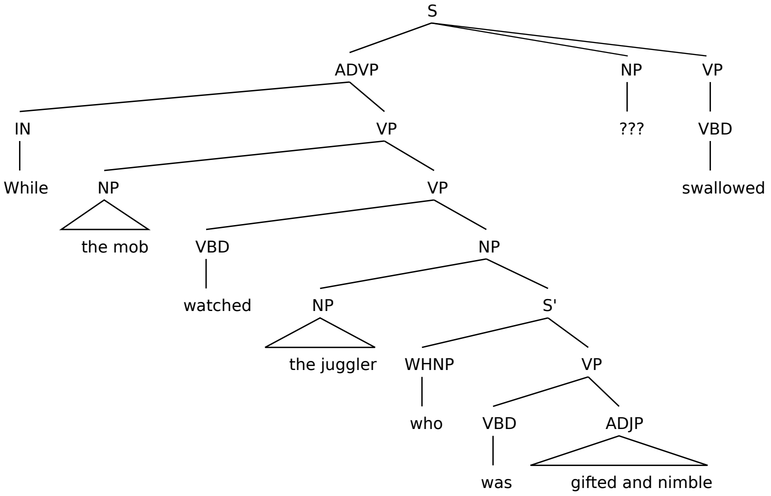


Figure 7.3: A typical misanalysis of Example 7.5. The attachment of swallowed is unsuccessful because its syntactic subject is missing - denoted by ???

fied as being “Good” within the standard EyeLink system. Once this had been achieved, participants were presented with five practice sentences. In each case the display was initiated when the participant pressed a button to advance to the next trial. A further press triggered the display of a question (where this was included). Comprehension questions were answered by pressing game controller buttons marked either “Yes” or “No”. Calibration and validation were freshly adjusted between the presentations of successive sentences. A fixation point was displayed at pixel (10,200) to mark the starting point of the sentence. Participants were instructed to press the Advance button once they had focused on this dot. Provided the tracker returned the same coordinates at this point (modulo small pre-specified tolerances), the dot display was removed and replaced immediately by the display of the new sentence. In cases where the discrepancy threshold was exceeded, there followed an automatic recalibration prior to the display of the new text. Participants were invited to ask questions during the practice session. On its completion, the experimenter left the test cubicle, allowing the participant to work their way through the full experimental session.”

7.3 Simulation data

In this section I report the outcome of computational simulations on materials of which an example is Example 7.5. There were four such simulations: two based on the HUMDEP parser (Boston, 2013) and two based on the TDPARSE parser (Roark, 2013). In the subsections below each simulation is described in detail. The value of each parser’s metric of difficulty is given in Table 7.1. The approach to comparing eye movement data with the output of surprisal and retrieval parsers has a parallel in Patil et al. (2009).

Dependency surprisal The parser was given as input an example of an ambiguous and a disambiguated version of each of the 24 materials. The surprisal value for the comma was added to the preceding word. Figure 7.4 shows that although the parser was sensitive to differences at the verb at the onset of ambiguity (word 4 which appeared either without a comma in the ambiguous conditions or with a comma in the disambiguated conditions), the metric had settled down by the time of disambiguation and there was no difference between the ambiguous and disambiguated versions of the sentences at the disambiguating word, word

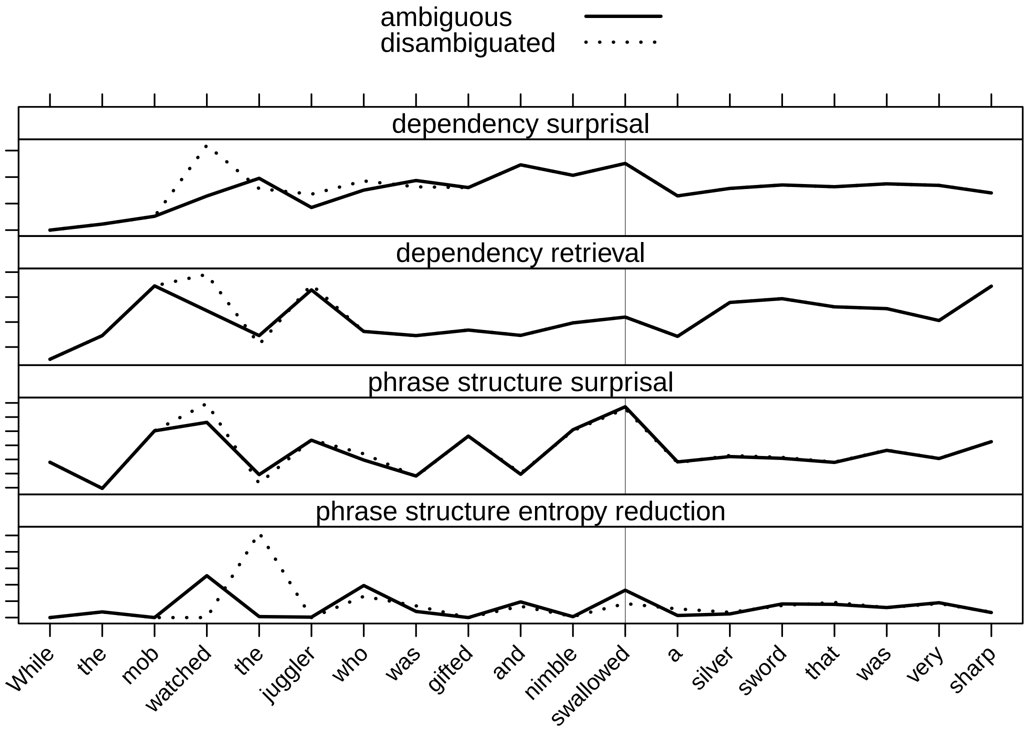


Figure 7.4: Computational measures ambiguity effect across a whole sentence, with a vertical line indicating the disambiguating word

12. In order to test this formally, an LMER model was constructed for dependency surprisal at the disambiguating word 12, with the output metric dependency surprisal as the dependent variable and the following independent variables: word length, word frequency, ambiguity, line position and the interaction between line position and ambiguity[1]; and terms for the random effects of participant and item, as well as terms for the main effects of ambiguity and line position and their interaction at each level of item[2]. The interaction effect over items had to be dropped to achieve model convergence but a term for item was retained. There was very little variation in dependency surprisal at the disambiguating word. Such variance as there was is accounted for by the parser's sensitivity to grammatical number in items earlier in the sequence than disambiguation[3]. This yielded a spurious significance for variation at the disambiguating word in word length, with longer words leading to more surprisal ($\beta < .01$, $SE < .01$, $t = 3.6$) and word frequency, with more frequent words leading to more surprisal ($\beta < .01$, $SE < .01$, $t = 5.9$), but there was no variance left over to be explained by ambiguity and line position, all of whose t values were 0.0.

Dependency retrieval Figure 7.4 shows that this measure was again sensitive to the presence of the disambiguating comma for values of words immediately following the comma (words 3 and 4) but that the difference had evened out by the disambiguating word 12. The formal model that tested this apparent insensitivity at the disambiguating word showed that there was no sign of a difference. The interaction effect over items had to be dropped to achieve model convergence but a term for item was retained. The model indicated a spuriously significant coefficient for word length ($\beta < -.01$, $SE < .01$, $t = -9.5$) with no variation accounted for by frequency ($t = .3$) or ambiguity and line position or the ambiguity x line position interaction (all of whose t values were 0.0).

Phrase structure surprisal Figure 7.4 shows that this measure was also sensitive to differences at the comma. The differences at the disambiguating word are very small but the formal test revealed that the small differences at the dis-

[1] The term for line position is included in the model to facilitate comparison with the human data which are described later in the chapter.

[2] These models did not include random effects for subject because surprisal only varied over items.

[3] The grammatical number of the first noun in the sequence at word 3 was sometimes singular and sometimes plural - a distinction to which the parser is sensitive.

Table 7.1: Computational measures at the disambiguating word[a]

ambiguity	line break	DSURP	DTIME	TSURP	ER
ambiguous	early	2.51	219.86	13.45	1.67
ambiguous	late	2.51	219.86	13.45	1.66
disambiguated	early	2.51	219.86	13.17	0.85
disambiguated	late	2.51	219.86	13.17	0.85

[a] Please see page 14 for the abbreviations used in the table

ambiguating word were significant. This marks a departure from the pattern for the dependency based measures discussed immediately above where there were no differences at all at the disambiguating word. The interaction effect over items had to be dropped to achieve model convergence but a term for item was retained. Effects of word length and word frequency were non-significant: ($\beta = .10, SE = .62, t = .17$ and $\beta = -.20, SE = .31, t = -.64$ respectively). There was a significant effect of ambiguity with ambiguous sentences attracting more surprisal than disambiguated controls ($\beta = .29, SE = .07, t = 4.539$). The effect of line position was not significant and nor was the ambiguity x line position interaction: ($\beta < .01, SE = .06, t = -.06$ and $\beta < .01, SE = .07, t = -.11$ respectively).

Phrase structure entropy reduction Figure 7.4 shows that there were again differences near to the comma for this metric. However, for this metric differences persist throughout the sentence. The plot suggests that there was a big difference at the disambiguating word for this measure. The formal test supports this impression. The interaction effect over items had to be dropped to achieve model convergence but a term for item was retained. Effects of word length and word frequency were non-significant: ($\beta = .09, SE = .24, t = .38$ and $\beta = .14, SE = .11, t = 1.21$ respectively). There was a significant effect of ambiguity with ambiguous sentences attracting more entropy reduction than disambiguated controls ($\beta = .81, SE = .02, t = 46.21$). The effect of line position was not significant and nor was the ambiguity x line position interaction: ($\beta < .01, SE = .02, t = -.15$ and $\beta < .01, SE = .02, t = .05$ respectively).

Table 7.2: Human measures at the disambiguating word[a]

ambiguity	line break	FFD	FPRT	RPD	TSR	PREG
ambiguous	early	216	369	463	94	0.07
ambiguous	late	278	347	657	311	0.30
disambiguated	early	227	344	415	71	0.05
disambiguated	late	231	292	435	143	0.19

[a] Please see page 14 for the abbreviations used in the table

7.4 Eyetracking data

In this section I submit the data from Experiment One of Mitchell et al. (2008) to scan path analysis. I show that layout affects scan path patterns when spatial and temporal properties of the regressive scan paths are taken into account. This represents additional information over the analyses presented in that paper because the original analyses were only able to deal with spatial properties of the regressive scan paths (using *regression signatures*), or temporal properties (e.g. using regression path duration) separately. I also report new analyses of the standard eye movement analyses from the data from the original experiment. These analyses differ from those presented in the original paper because here I use mixed effects models to simultaneously account for effects of participant and item in a single analysis whereas in the original paper traditional separate by-participants and by-items analyses were reported. Furthermore, because the single mixed model approach is robust with respect to missing data, some of the limitations of those original analyses are overcome. For example in the original analyses data had to be discarded from participants if they did not contribute data to one of the conditions – here I can take advantage of the data that they did contribute without violating the assumptions of the statistical test. In this section I give the results of the new analyses (not just reporting the original results). Mean times and probabilities are given in Table 7.2.

The results from human behavioural measures are given first, followed by the results of model evaluation against these results. The behavioural measures are: first fixation duration; first pass reading time; regression path duration; time spent regressing; probability of regression; scan path analysis; distribution of scan path behaviours at disambiguation; distribution of regression strategies at disambiguation. The computational parsers give their account of processing difficulty in the following measures: surprisal from a dependency grammar parser; retrieval time from a dependency grammar parser; surprisal from a phrase structure grammar parser; entropy reduction from a phrase structure grammar parser.

For each experimental sentence I examined fixation time on the disambiguating word. If a fixation was less than 80 ms and located within 1 degree of visual angle of another fixation, it was merged into (added to) that fixation. If there were two such neighbouring fixations, the short fixation was merged into the larger of the two neighbours. If there were no such neighbouring fixations, the short fixation was deleted. In another pass over the data, any remaining fixations that were less than 80 ms or greater than 800 ms were deleted.

The data analyst must decide how to treat a value of 0 for reading time data. This value is yielded when the critical word is skipped. I take the view that zero values for reading time measures are properly transformed into missing values denoted NA in R. However, since words are frequently skipped, the transformation of zeroes into missing values yields a data set where some combinations of subject and item contain only missing values. In some cases this can cause the maximal model to fail to converge properly because slopes are computed at each level of the interaction between ambiguity and line break separately over each participant and separately over each item. All such cases are noted in the text, and I state how the model was adapted to cope (either by leaving the zeros in the data, or by relaxing the model specification). As it turns out, the probability of skipping a word was approximately equal across the linguistic manipulations, with the consequence that leaving the zeros in the data rarely affected the direction or significance of any of the results reported below. Nevertheless the reader may have confidence that unless otherwise stated, each model is maximal, and skipped words are treated as yielding missing values for reading time measures.

Four fixation based measures and one non-time based measure were computed, including standard measures described in Rayner (1998) and Rayner (2009).

For each of the five measures a multilevel linear model was fitted, with terms for the low level covariate word properties word length and word frequency, which were both centred; fixed effects for ambiguity and line break and their interaction; and random intercepts for participant and item; as well as random slopes for the fixed main effects and interaction effect at each level of participant and item. Such a model may be described as a *maximal* model (Barr et al., 2013).

The mathematical designation of such a model is

$$y = \beta_0 + \beta_1 + \beta_2 + \beta_3 x_1 + \ldots + \beta_i x_i + b_3 x_1 + \ldots + b_j x_j + \in \qquad (7.6)$$

where β is a fixed effect and b is a random effect and $\in$ is the error. The R syntax for such a model is: `measure ∼ length + frequency + effect + (effect|subj) + (effect|item)` where *effect* is a contrast matrix that specifies the main effects and their interaction.

First fixation duration The model of FFD showed no effect of word length ($\beta = 4.5$, $SE = 5.2$, $t = .86$); no effect of frequency ($\beta = -2.9$, $SE = 2.5$, $t = -1.73$); no main effect of ambiguity ($\beta = 10.2$, $SE = 6.4$, $t = 1.6$); and no main effect of layout ($\beta = -20.1$, $SE = 12.5$, $t = -1.6$); but a significant ambiguity x layout interaction ($\beta = -15.7$, $SE = 7.7$, $t = -2.0$).

First pass reading time The model of FPRT showed no effect of word length ($\beta = -13.2$, $SE = 15.7$, $t = -.84$); no effect of frequency ($\beta = 4.2$, $SE = 7.7$, $t = .55$); and no main effect of layout ($\beta = 15.4$, $SE = 15.8$, $t = .98$). There was a significant main effect of ambiguity ($\beta = 20.1$, $SE = 10.3$, $t = 1.96$); but no ambiguity x layout interaction ($\beta = -9$, $SE = 9.2$, $t = -1.0$).

Regression path duration The model of RPD showed only a main effect of ambiguity ($\beta = 65.4$, $SE = 27.9$, $t = 2.3$). The other effects were not significant: word length ($\beta = -25.3$, $SE = 40.4$, $t = -.6$); frequency ($\beta = -9.3$, $SE = 19.8$, $t = -.47$); layout ($\beta = -49.0$, $SE = 31.7$, $t = -1.55$); ambiguity x layout ($\beta = -44.3$, $SE = 28.3$, $t = -1.57$).

Time spent regressing The model of TSR showed only a significant main effect of layout ($\beta = -64.1$, $SE = 29.3$, $t = -2.184$). Other effects were not significant: word length ($\beta = 4.38$, $SE = 33.9$, $t = .13$); word frequency ($\beta = -19.55$, $SE = 16.7$, $t = -1.17$); ambiguity ($\beta = 43.6$, $SE = 27.6$, $t = 1.57$); ambiguity x layout ($\beta = 35.4$, $SE = 29.3$, $t = -1.21$).

Probability of a regression The model of PREG converged only after dropping the interaction term over items. Only the main effect of layout was significant

($\beta = -6.5$, $SE = 1.33$, $z = -4.882$, $p < .001$). Other effects were not significant: word length ($\beta = -.49$, $SE = .4$, $z = -1.2$); word frequency ($\beta = -.29$, $SE = .19$, $z = -1.48$); ambiguity ($\beta = 1.48$, $SE = 1.0$, $z = 1.4$): ambiguity x layout ($\beta = .21$, $SE = .98$, $z = .22$).

Scan path analysis In this section I present analyses of the spatio-temporal dynamics of scan paths that were launched from the disambiguating word and that ended just before a fixation in new material. When one considers behaviour at the disambiguating word, one can categorise this behaviour into several types. These are: (1) *skip*: the word was skipped on the first pass; (2) *progress*: the word received a single first pass fixation that was followed (immediately and without any intervening fixations) by a fixation in new material to the right; (3) REFIXATE: the word received a first pass fixation and was then refixated, possibly more than once, before taking in new material to the right; (4) REGRESS: the word received one or more fixations that were followed not by a movement to the right to take in new material, but by a movement to the left to take in material already passed through, regardless whether those words fixated to the left had been fixated or not on the first pass. This classification is complete and without redundancy, in the sense that it results in every trial receiving exactly one value. First the distribution of these behaviours over the treatment conditions is considered. This permits answers to questions like “Were people more likely to *skip* the word in some condition?”; “Did some condition have a tendency to result in simple *progress* behaviour?” Next I will concentrate on the *regress* cases and further subdivide them into various types of regression. This will allow questions like “Did the ambiguous theft conditions make people more likely to re-examine the words intervening between disambiguation and the upstream onset of syntactic ambiguity, as opposed to making a direct movement from disambiguation to upstream ambiguity onset?” First I consider the distribution of coarse behaviours at disambiguation, plotted in Figure 7.5. The *regress* behaviour appears to be over represented in the ambiguous late line break condition.

Next the regress cases were separated off for further analysis.

The analysis starts by computing for each regression scan path its pairwise dissimilarity from all other scan paths. This involves making several choices about parameters which are detailed here. I chose the formula fixation duration as a function of word number rather than using the raw x,y coordinates. This is because the main interest in this analysis is to compare the words visited in

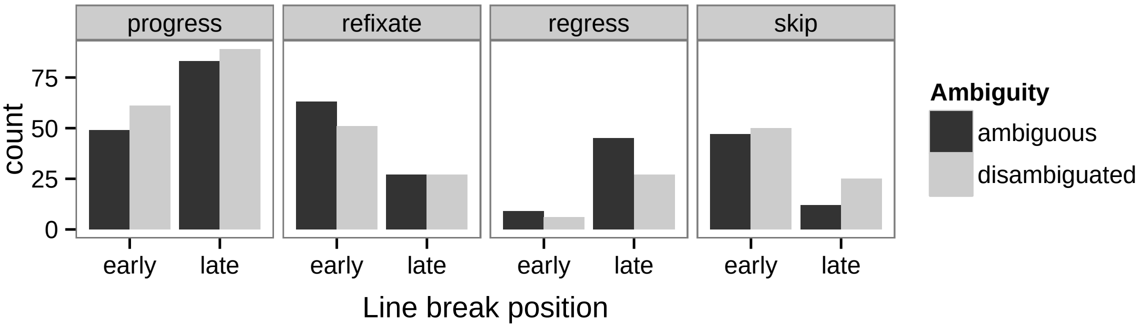


Figure 7.5: Distribution of coarse behaviours at disambiguation

each condition, rather than the abstract spatio-temporal pattern of the movements themselves. I set the modulator to zero in order to bypass the machinery for taking into account foveal eccentricity. This is because when the participants' heads are free to move, as in a head mounted eyetracker, it is not possible to get a reasonable estimate of viewing distance for each trial, and so the modulator would in effect be being applied inconsistently is it was used. I normalized the raw scasim values by fixations to avoid arbitrary effects of number of fixations in a scan path (the alternative is to normalize by durations).

The first run of scasim is done to check for outliers: none were observed using a 2.5. sd threshold. This leaves a corpus of 87 regressions for more detailed analysis.

Fig 7.6 plots stress against number of dimensions and number of clusters. This plot is used to choose a dimensionality for the scaled-down data. I chose 6 dimensions as a reasonable compromise between aiming for a low stress (good fit) without allowing too many clusters (5 in this case) in the solution.

The WHICH.MEAN function in the SCANPATH package allows identification of a single scan path that lies nearest the centroid of a given cluster. Applying this method yielded the scan paths plotted in Fig 7.7. Pattern D is the only one that takes in the onset of ambiguity - it goes back as far as the head of the misattached noun phrase. There were 34 such regressions[1]. I constructed a LMER model of this cluster by recording for each trial whether it resulted in a cluster D type scan path, and then asking whether a trial's membership of this cluster was significantly predicted by ambiguity and layout, taking proper account of the contributions of

[1] Because there were so few regressions, it is possible that this might have led to problems with R's backing off algorithms where proportions approach zero.

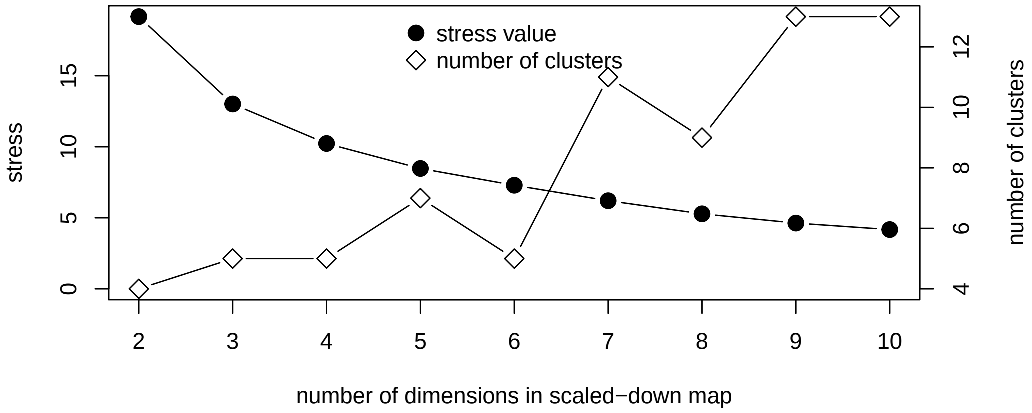


Figure 7.6: Choosing a dimensionality, experiment one

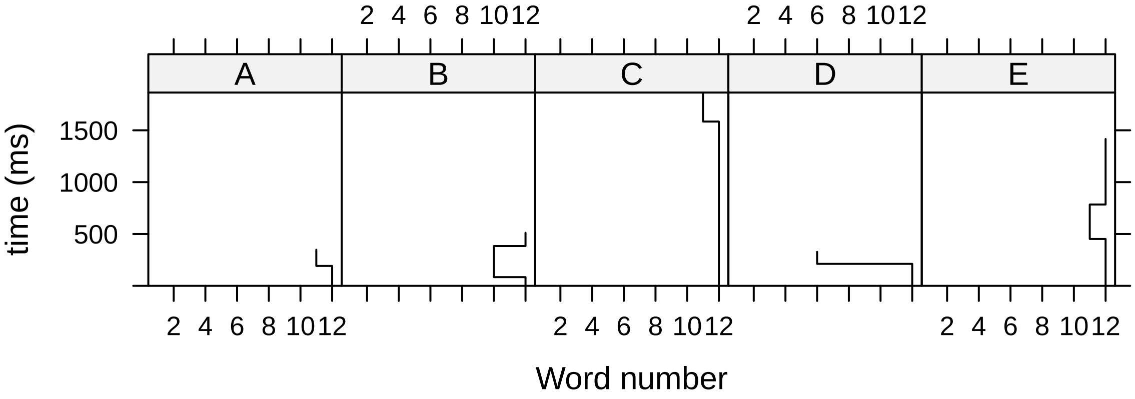


Figure 7.7: Regression shapes for Experiment 1

individual subjects and items. The term for the interaction of ambiguity and layout over each level of item had to be dropped so that the model converged. This left a model that had terms for ambiguity and layout and the ambiguity x layout interaction, as well as separate terms for the random effects of subject and item, and for the ambiguity and line break main effects over subjects, and the ambiguity x line break main effects and interaction over each level of item. The R syntax for the model was:

(7.7) `lmer(data=mydata, isD ~ ambiguity*line break + (ambiguity*line break|subject) + (ambiguity+line break|item), family ="binomial")`

The result of the model indicated that the cluster D regressions were more likely to be made in the late line break condition (probability of coming from early conditions = .02; probability of coming from late conditions = .08, $\beta = -1.16, SE = .45, z = -2.6, p < .05$). The ambiguity main effect was non-significant ($\beta = .41, SE = .45, z = .9, p = .34$) as was the ambiguity x line break interaction ($\beta = -.57, SE = .43, z = -1.4, p = .17$).

7.5 Discussion

The new analyses of the Mitchell et al. (2008) data reveal that the spatiotemporal properties of regressions launched from disambiguation vary as a function of how the materials are arranged on screen. This means that regressions can only be at best partly guided by purely linguistic factors.

The data also show that people seek out material that is relevant for repair purposes – that is to say that there were enough regressions of the D type to form a distinct cluster. In this case people either regressed one or two words (which is a pattern that is compatible with time-out), or, in a distinct cluster (D), sought out the word *juggler* which is the head of the misattached noun phrase and the unit that must be moved away from being the object of the subordinate clause to the head of the matrix clause in a repair operation.

The results of the simulations with computational parsers show that only the phrase structure parsers were sensitive to differential difficulty at the disambiguating word. The dependency grammar parsers were sensitive to differences at the onset of ambiguity but these differences did not persist into the disambiguating word, at which the same difficulty was predicted regardless of whether ambiguity persisted up to the disambiguating word. An explanation for the dependency parser's lack of coverage here could be due to differences in how a dependency parser and a phrase structure parser handle punctuation. Punctuation is considered as a terminal in its own right in the dependency grammar, and dependency arcs are assigned to the relations between the punctuation terminals and the lexical terminals. The phrase structure parser also treats the punctuation as a terminal in its own right, but makes much more out of the information provided by the punctuation. Since the phrase structure parser explicitly assigns hierarchical structure to the input where the dependency parser does not, it is better placed than the dependency grammar parser to benefit from the clause boundary information that is inherent in the punctuation but speaks to a level higher than the terminal level.

In summary, at disambiguation, people either made time-out consistent small regressions, or they made regressions that are compatible with a repair based explanation that requires the parser to have kept track of points at which more than one analysis was available such that they can be sought out for repair. The data favour a repair based account of disambiguation as well as a parser that operates over a phrase structural grammar.

Did the LMER and SCASIM analyses shed any additional light on the data over and above that shed by the original ANOVA and Regression Signature results from Mitchell et al. (2008)? I will take LMER and ANOVA first, for the standard eye movement measures at the disambiguating word, and then compare scan path similarity with the regression signature analysis.

The original ANOVA analyses detected the following effects at the disambiguating word (where an effect is required to reach significance on both by-participant (F_1) analysis and by-item (F_2) analysis): ambiguity main effects in RPD, PREG and TRT; and a layout effect in PREG. LMER found the most important of these effects too: ambiguity main effects in RPD and TRT[1], and the layout main effect in PREG. LMER missed the ambiguity main effect in PREG, and also the layout effect in TRT. LMER found an ambiguity x layout interaction in FFD that ANOVA missed, but this seems unlikely to play a theoretically relevant role in interpretation. LMER found an ambiguity main effect in FPRT that ANOVA missed.

For the standard measures then, the original ANOVAs and the new LMERs tell more or less the same story which can be summarised thus: (1) people were more likely to regress from disambiguation on the same line than from disambiguation on the line below and (2) people spent longer regressing from disambiguation on the same line than from the line below in the ambiguous conditions.

However, SCASIM analysis outperformed the original Regression Signature analyses, as follows. SCASIM revealed a cluster of regressions back to the linguistically relevant head of the ambiguously attached NP that was over distributed in the late line break conditions. This indicates that people were able to make linguistically targeted regressions better when disambiguation was on the same line versus the line below. The original Regression Signature analyses were not directly comparable because those analyses had to omit participants and items for which a regression was not made in one of the four conditions. This meant that

[1] this measure TRT is not computed for most of the analyses in the thesis but it was computed for purposes of comparison for this experiment. The results were an ambiguity main effect ($\beta = 87.39, SE = 16.45, t = 5.314$), but no layout main effect ($\beta = 22.53, SE = 24.26, t = 0.929$)

it was not possible to carry out orthogonal comparisons with enough data. The response to this in the Regression Signature analyses was to enlarge the corpus of regressions to include those from spillover region as well as the disambiguating word itself, and to restrict the analyses to compare only the ambiguous early and ambiguous late conditions. The closest that the Regression Signature Analyses could come to the LMER result was the finding, for all first-landing-site regressions, of a main effect of Landing Site that showed that the distribution over all words was uneven. More targeted analysis on individual words in ambiguous and unambiguous conditions failed to show significant differences at any word except the word before disambiguation. It is the SCASIM analysis that provides the clearest evidence for targeted repair operations based on the Mitchell et al. (2008) Experiment One data.

What layout effects say about reading and parsing There is a temptation to use models of regression-control as though regression-control were sensitive to linguistic manipulation only. The data show that regression-control is sensitive to layout. The risk of failing to include layout effects in models of regression-control when we know that they exert an influence on regressions is that we ascribe an observed phenomenon to the wrong generative process. This becomes serious when we use the observable phenomena to distinguish between implemented accounts of eye movements in reading.

Layout effects matter because the best approach to reading is to study it in the round. Modelling only the linguistic aspects of reading and not the ocular and motor aspects of reading must lead to models that underfit the data – they underestimate the complexity of the data. Adding layout markers to a model of regression control would increase the fit of statistical models of reading scan paths generators, giving a clearer view of the effects of linguistic manipulations, making the statistical more suitable for adjudicating between competing theories.

Perhaps the most promising of the currently implemented models of eye movement control in reading from this point of view is EMMA . This is because it offers an easy way to build visual properties in to the model of the cognitive process. Adding the capability to represent the visual layout of the text would be a promising way to develop EMMA.

The contrast between the performance of the dependency parser on corpus data and its performance on disambiguating words illuminates a possibility for a dual-mode model of parsing, alluded to in Mitchell et al. (2008). A dual-mode model of parsing could adopt dependency parsing for first pass analysis, and switch modes to phrase structure parsing when dependency parsing fails to produce the correct analysis. This would explain both the co-presence of effects of linguistic manipulations and of layout manipulations in the present experiment; and the disparity between the performance of the dependency parser on non-disambiguating and disambiguating word.

Chapter 8

E2: Spatial layout affects disambiguation scan paths

This chapter describes a new experiment that manipulates the layout of materials on screen. I show that layout exerts an effect on the spatiotemporal properties of regressive scan paths. I argue that these effects are covered poorly by the computational parsers. The human reading time measures did not generally reveal effects of layout, but the likelihood of making a regression, and the shape of the regressive scan paths that were made did reveal effects of layout. These layout effects were not predicted by either of the computational parsers, but the phrase structure parser was sensitive to the manipulation of ambiguity. It is argued that the results support repair models of parsing implemented over a phrase structure grammar.

8.1 Introduction

In the Mitchell et al. (2008) paper, the manipulation of line break resulted in disambiguation at the start of a line (this was compared against the same material with a line break that presented the disambiguating word at the end of the preceding line). As a consequence of the line-initial position of the disambiguation in that condition, we could not be sure whether the disinclination to regress that we observed in that condition was due to line-initial placement rather than to merely different placement. A plausible reason for the observed disinclination was that there was no material to regress to within-line, with the consequence that any regression that was made would have to cross the line-boundary between launch and target. People might have been more willing to regress in this condition if we had provided some within-line material (perhaps by padding the line). This could have increased regression rates in the control condition, and that would have diminished the observed effects in Mitchell et al. (2008), possibly even to the point of nullifying the effects. In short the findings of Mitchell et al. (2008) are subject to doubt because there was no same-line material to serve as an attractive target of regressions.

The present experiment seeks to remedy the deficiency in the Mitchell et al. (2008) materials by offering a case where there is some within-line material to regress into, but using the same linguistic materials as the Mitchell et al. (2008) experiment to facilitate comparison. In the present experiment I show that likelihood of regression is significantly modulated by the line position of the word from which the regression is potentially launched when linguistic content is held constant and when same-line material is offered. This is interpreted as evidence that control of regressive eye movements is shared between controllers from the linguistic system and from the oculomotor system, and evidence against a strong selective reanalysis linguistic-guidance-only account, as well as against a strong decoupled-linguistic-system account.

I crossed the factor *line break* with the *ambiguity* factor to bring about two ambiguous conditions with the disambiguating word in different places and two baseline conditions that pre-disambiguated the sentences with a clause boundary-marking comma. In each ambiguous case the linguistic sequence was held constant so that participants were dealing with the same parsing problem.

Under the assumption that eye movements in regressions are subject only to linguistic governance - a strong form of selective reanalysis - no effect of line break is expected because the layout of the text should exert no influence. Under the assumption that eye movements in regressions are a random walk - a strong version of the time-out hypothesis - no effect of line break is expected because the layout of the text should not exert a patterned effect. Under the assumption that control of eye movements in regressions is a constraint-based system of control that takes influences from the linguistic system and the oculomotor system, the hypothesis is that there should be main effects of ambiguity (linguistic constraint) and line break (oculomotor constraint), and that they should interact, in the later measures of parsing.

In these materials the layout is manipulated. Example 8.1 shows one of the sentences used - the materials from Experiment 1 were used again.

(8.1) While the mob watched (,) the juggler who was gifted and nimble swallowed a silver sword that was very sharp.

Figure 8.1 shows the materials as they were laid out in the experiment, with true line breaks and a monospaced font like the one used in the experiment to preserve the relative positions of the words on screen. The disambiguating word is underlined here. In the figure, the top sentence shows the ambiguous, early line break condition; below that is the ambiguous late line break condition; below that is each of the disambiguated controls. The full list of materials is given in Appendix A.

```
While the mob watched the juggler
who was gifted and nimble swallowed
a silver sword that was very sharp.

While the mob watched the juggler who was gifted
and nimble swallowed a silver sword that was very
sharp.

While the mob watched, the juggler
who was gifted and nimble swallowed
a silver sword that was very sharp.

While the mob watched, the juggler who was gifted
and nimble swallowed a silver sword that was very
sharp.
```

Figure 8.1: Illustration of how the spatial layout manipulation was implemented in experiment 2. The conditions were as follows, in the same order as the illustration: *ambiguous, late line break*; *ambiguous, early line break*; *pre-disambiguated late line break*; *pre-disambiguated early line break*

8.2 Method

The method section gives details of participants, apparatus, and procedure.

Participants In total 21 participants were tested. Of these, 16 were retained and entered into analysis. 2 were compromised by a programming error, leaving 19. Of these, 3 were discarded so that balance could be maintained with respect to the Latin square design. Participants were native speakers of British English who were students of Psychology at the University of Exeter, were given partial course credit to participate in the experiment. All had normal or corrected to normal vision, were naive to the purpose of the experiment, and were aged between eighteen and thirty-four.

Apparatus An SR Research Eyelink II head-mounted eyetracker was used to record participants' eye movements with a sampling rate of 500 Hz. Participants read sentences displayed on a 19- inch Iiyama Vision Master Pro video monitor at 1024 x 768 resolution at a refresh rate of 60 Hz. Viewing was binocular but only the right eye was recorded. Participants sat in a dimly lit room in front of the computer at a viewing distance of approximately 75 cm. The average viewing distance was approximately 75 cm. At this viewing distance, and assuming that 1 character had 2 mm width on screen, a single character subtended 0.153° of visual angle, and approximately 6.5 characters subtended 1° of visual angle. The font used was Courier New 12 point. All sentences in this experiment were displayed on a single line with a maximum length of 100 characters. A 9 point calibration procedure was used, on which participants were required to achieve a score of 'good'. Each trial started with a drift correction routine where the participant was required to fixate a target that appeared in the same location as the first character of the sentence would subsequently occupy, and then required to press a button on the gamepad while fixating this point to start the trial.

Procedure Participants were instructed to read silently for comprehension at a comfortable speed. The practice trials and experimental trials were implemented as separate consecutive blocks. The experimental trials were randomised each time the experiment was run, i.e., in a different order for each participant, with the constraint that a maximum of two trials of a given type could appear in a continuous sequence. There were four practice sentences, followed by a drift correction routine preceding the experimental block containing 96 sentences, comprising 24 in experimental conditions (6 in each of 4 conditions); 24 foils and 48 fillers. Participants were rotated over one of four lists, implementing a Latin square design. 32 of the trials (including 8 of the experimental conditions) were followed immediately by a comprehension question. This was a simple question about the sentence immediately preceding that required the participant to make a yes or no response using the appropriate trigger button on the gamepad. The full list of questions asked may be found in the Appendix. The whole procedure took about

Table 8.1: Computational measures at the disambiguating word[a]

ambiguity	line break	DSURP	DTIME	TSURP	ER
ambiguous	early	2.51	219.86	13.45	1.67
ambiguous	late	2.51	219.86	13.45	1.66
disambiguated	early	2.51	219.86	13.17	0.85
disambiguated	late	2.51	219.86	13.17	0.85

[a] Please see page 14 for the abbreviations used in the table

20 to 40 minutes, depending on the participant.

8.3 Simulation data

Because the computational parsers used in the thesis have no way to represent the spatial arrangement of text, the results from the computational simulations are the same as reported in the previous chapter in section 7.3 on page 123. In this section I summarise those simulation results for convenience. Table 8.1 shows the mean values per condition.

Dependency surprisal There were no differences at disambiguation.

Dependency retrieval There were no differences at disambiguation.

Phrase structure surprisal There was a significant effect of ambiguity in line with the human data.

Phrase structure entropy reduction There was a significant effect of ambiguity in line with the human data.

Table 8.2: Human measures at the disambiguating word[a]

ambiguity	line break	FFD	FPRT	RPD	TSR	PREG
ambiguous	early	290	343	791	448	0.41
ambiguous	late	288	341	575	234	0.26
comma	early	276	300	442	142	0.22
comma	late	264	280	446	167	0.16

[a] Please see page 14 for the abbreviations used in the table

8.4 Eyetracking data

In this section I give the results from the standard eye movement measures at the disambiguating word, and for spatio-temporal analysis of scan paths. For each standard eye movement measure a LMER model was fitted with the following model specification: fixed effects for centred word frequency and centred word length; fixed effects for line break position and ambiguity and their interaction; and random effects for the main effects and interaction at each level of subject and item, i.e., a maximal model. The results are given in Table 8.2.

First fixation duration There were no significant effects in first fixation duration. The effect of word length was non-significant ($\beta = 9.21, SE = 9.23, t = .99$); the effect of frequency was non-significant ($\beta = 1.63, SE = 4.53, t = .36$). The main effect of ambiguity was non-significant ($\beta = 19.14, SE = 13.51, t = 1.42$). The main effect of line break was non-significant ($\beta = 6.43, SE = 14.20, t = 0.45$) as was the ambiguity x line break interaction effect ($\beta = -3.00, SE = 12.49, t = -0.24$).

First pass reading time Effects of word length and word frequency were non-significant ($\beta = 13.29, SE = 15.69, t = .85$); ($\beta = -3.85, SE = 7.68, t = -.50$). There was a significant main effect of ambiguity. Ambiguous conditions resulted in more first pass reading time ($\beta = 48.73, SE = 24.97, t = 1.95$). The effect of line break was non-significant ($\beta = 11.33, SE = 26.67, t = -0.43$). The ambiguity x line break interaction effect was not significant ($\beta = -12.28, SE = 23.21, t = -0.53$).

Regression path duration Regression path duration did not yield significant effects. Effects of word length and word frequency were non-significant: $(\beta = 1.28, SE = 54.38, t = .02)$; $(\beta = -14.73, SE = 26.63, t = -.55)$. There was a slight non-significant tendency for the ambiguous conditions to produce more RPD and a non-significant tendency for there to be more RPD for ambiguous early line break than for ambiguous late line break $(\beta = -239.63, SE = 131.72, t = 1.82)$. The effect of line break was not significant $(\beta = 116.33, SE = 91.32, t = 1.28)$. The ambiguity x line break interaction effect was not significant $(\beta = 115.92, SE = 79.86, t = 1.45)$.

Time spent regressing No effects were significant in time spent regressing. Effects of word length and word frequency were non-significant ($\beta = -14.33$, $SE = 56.82$, $t = -.25$; $\beta = -4.219$, $SE = 27.88$, $t = -.15$). The main effect of ambiguity was a trend towards increased time in the ambiguous conditions ($\beta = 188.88, SE = 126.76, t = 1.49$). The main effect of line break was a tendency toward increased time in the early line break condition ($\beta = 107.91, SE = 126.76, t = 1.29$). There was a non-significant interaction effect, with the ambiguity disadvantage greater in the early line break conditions than in the late line break positions ($\beta = 126.67, SE = 74.68, t = 1.70$).

Probability of a regression Effects of word length and word frequency were non-significant: $(\beta = -.19, SE = .26, z = -.72, p = .47)$; $(\beta = -.04, SE = .13, z = -.31, p = .76)$. There was a significant main effect of ambiguity ($\beta = 1.37, SE = 0.55, z = 2.50, p < .05$) and a significant main effect of line position ($\beta = 1.32, SE = 0.44, z = 2.99, p < .01$) but no interaction effect ($\beta = -0.03, SE = 0.38, z = -0.70, p = 0.95$).

Scan path analysis Here I give an overview of behaviours at disambiguation arranged by type: progress; refixate; regress; and skip.

First the data were restricted to movements launched from the disambiguating word. Trials on which readers skipped the disambiguation were removed from the analysis. This left 3 types of movement: simple progression where readers fixated the disambiguation once and then moved to take in later post-disambiguation material; refixation followed by progression, where readers fixated the disambiguation and then refixated it some number of times before moving to take in new

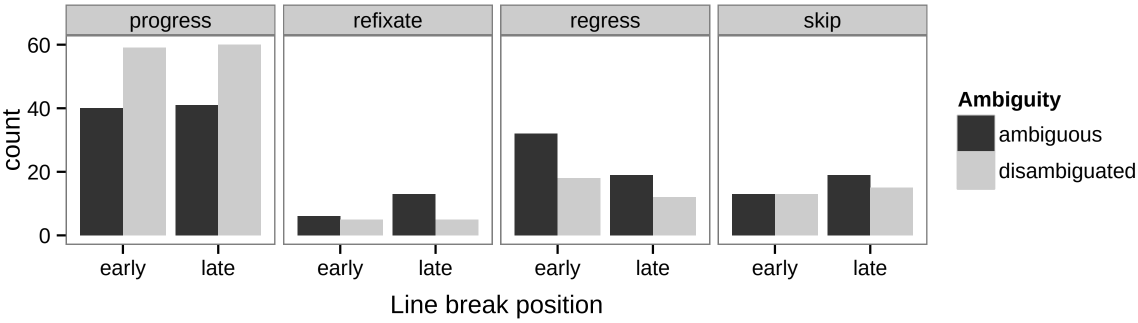


Figure 8.2: How coarse eye movement behaviours launched from disambiguation were distributed over conditions

post-disambiguation material; and regression, where people fixated the disambiguation (possibly including refixations) and then launched regressive movements into earlier pre-disambiguation material. Cases of simple progression were removed from the analysis on the grounds that simple progression provides no evidence of difficulty.

This left two classes of behaviour that were considered to indicate problems with disambiguation: refixation and regression. There were 110 trials on which these classes of behaviour were observed. 81 of these were regressions and 29 were refixations.

Scan path similarity was computed between each pair of scan paths in the set. For each scan path this yielded a vector of that scan path's dissimilarity from every other scan path. No outliers were identified using a 2.5 sd threshold.

Multi-dimensional scaling was employed to yield a lower-dimensional representation of dissimilarity space that was still faithful to the original larger space. The lower dimensional representations were computed for each of 2 to 10 dimensions. For each of these 9 dimensionalities, lower-d maps were constructed. For each of the lower-d maps, a measure of faithfulness to the original space were obtained. This measure is denoted *stress* in the function ISOMDS. Each lower-d map was analysed to derive the optimal number of clusters at that dimensionality, using the function MCLUST that carries out cluster analysis and returns the optimal number of clusters in the data for a given lower-d space, using a Bayesian Information Criterion (BIC) to select a clustering. Using this criterion penalises adding to the number of clusters which prevents over-fitting the data: without this penalty the best fit would be obtained by having one cluster per scan path.

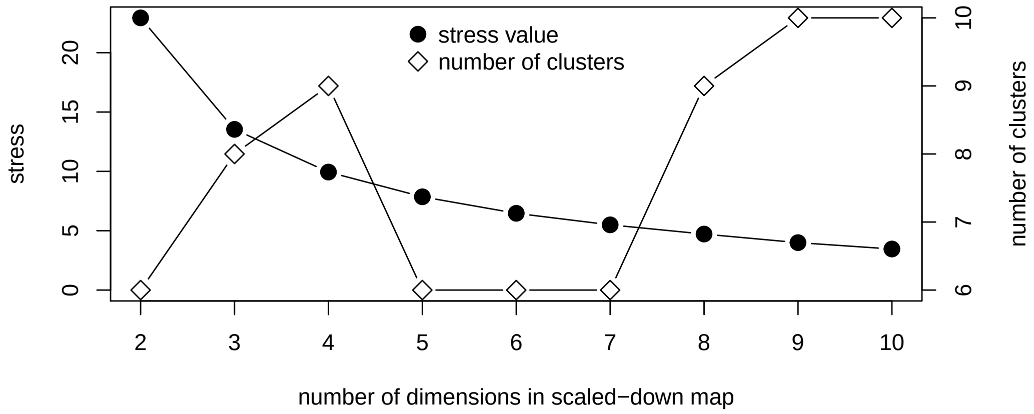


Figure 8.3: How a 7 dimension map was selected.

This offered a way to select a dimensionality with respect to the stress of the map at that dimensionality (stress values between 10 and 5 are considered to be reasonably faithful to the original space), and also with respect to the number of clusters in the lower-d space (all other things being equal one would prefer fewer clusters in the space). See Figure 8.3 for an illustration of the method for choosing a dimensionality. A map was selected that had 7 dimensions. At this dimensionality stress was reduced to nearly 5, and there was a reasonably small number of clusters (6 clusters) in the best clustering of that space.

Having identified a lower-d space to work in that was still faithful enough to the original space and which contained a reasonably small number of clusters, the clusters in the space became the focus of analysis. For each cluster, the scan path nearest the centre of the cluster was identified using the function WHICH.MEAN. Usually there is no scan path at the exact centre of the cluster, which is why the scan path nearest the centre is used to represent the central tendency of the data.

Figure 8.4 shows the shape of the scan path that sat nearest the centre of each of the 6 clusters identified in the best clustering of the data. Cluster A represented the trials on which participants refixated the disambiguation and then progressed. Clusters C, D, and E represented trials on which participants made small regressions of one or two words and then progressed to new material. These clusters differed according to how long the fixations were and whether the disambiguation was refixated before moving on. Cluster F represented a regression to the first noun phrase of the ambiguous region - there were only 3

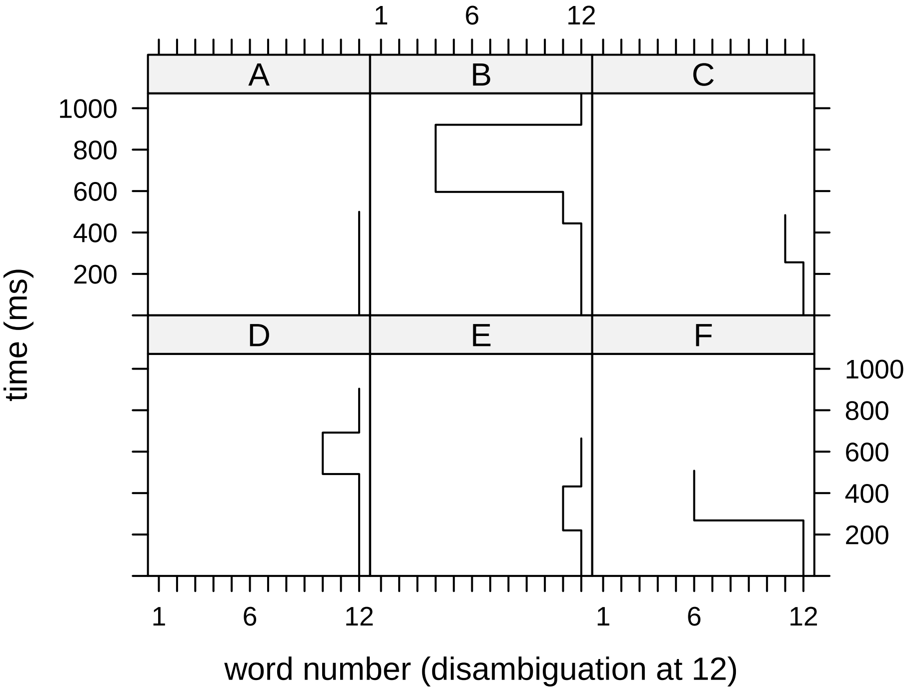


Figure 8.4: Regression strategy scan paths. Plot shows the shape of the scan path that represents the central tendency of each of the identified clusters

such regressions in the set. Cluster B represented a regression that moved back to fixate the first verb in the sentence, which participated in the ambiguity: it was the verb with the ambiguous subcategorisation frame, e.g. *watched*.

The cluster B regressions were indicative of selective reanalysis in that they mostly targeted the verb at the onset of ambiguity, and thus were selective, but which were made in materials where the onset of ambiguity was not also launch-adjacent, which could have been explained by time-out. This disambiguation strategy accounted for 46% of the 110 movements launched from disambiguation that were not simple progression cases. The fact that these movements targeted a non launch adjacent part of the sentence that instigated the syntactic ambiguity in the sentences supports the view that people made these movements in order to provide the linguistic system with disambiguation-relevant information like whether the verb licensed an alternative subcategorisation which might have required another lexicon look-up.

Having identified these regressions as theory-relevant, they were analysed in more detail. Specifically, a data set was constructed for the 110 trials on which

Table 8.3: Distribution of strategy B

	not-B	B
ambiguous end of line	15	23
ambiguous middle of line	19	13
comma end of line	14	9
comma middle of line	11	6

non-progression was observed that linked the cluster to which that trial ended up being assigned to the subject, item, and experimental condition for that trial. This made it possible to ask whether the experimental conditions predicted the B-type regressions while also accounting for the subject and item that produced an individual data point. Remaining sensitive to the subject and item that were involved on the trial prevents the model being misled by any possible over-dispersion over subjects or items. In other words a multilevel model should prevent the analyst being misled by a situation in which particular readers or sentences account for the prevalence of a given behaviour, such that the analyst does not wrongly attribute that variance to the treatment conditions.

Table 8.3 shows that the B type regressions were most common in the ambiguous end-of-line condition, as would be expected on a selective reanalysis account but not on a time-out account. However, the multilevel model gives reason to reserve judgement about this pattern since it does not significantly attribute the variance in these simple counts to the treatment conditions.

The multilevel model had the following structure. If a trial was assigned to the B cluster, this was coded as 1, with the alternative coded 0, yielding a binomially distributed dependent variable. This dependent variable was modelled as a function of ambiguity and line position and the ambiguity x line position interaction treated as fixed effects, as well as random effects representing subject and item, and the main effects modelled at each level of subject and item. Terms representing the interaction effect over subjects and items were dropped after the maximal model failed to converge. The resulting model indicated that the main effects and interaction did not significantly predict a B-type regression once subject and item-level variance was taken into account (effect of ambiguity: $\beta = .35, SE = .41, z = .87, p = .39$; effect of line position: $\beta = .63, SE = .37, z = 1.72, p = .09$; interaction effect: $\beta = .32, SE = .31, z = 1.05, p = .30$).

So far the scan path analysis has included measures from the disambiguated conditions. When analysing reading times this is appropriate because the disambiguated conditions are predicted to cause some reading time that acts as a baseline for the ambiguous conditions. This logic does not hold so well in scan path analysis after progressive fixations are excluded. This is because unproblematic words (like the critical word in a pre-disambiguated condition) are expected to yield progressive fixations. Any regressive scan paths made in these conditions must reflect non-disambiguation events. Therefore a further model was constructed that restricted the data to the ambiguous conditions and asked whether there was variance in the scan path outcomes of the ambiguous cases as a result of line position.

This model included terms for line position, and for the effect of line position at each level of subject and item. The model showed that the effect of line position exerts a significant influence on the shape of the regressive scan paths that are made from the disambiguating word when per-subject and per-item influences are also taken into account. The effect of line position was significant: regressive scan paths that sought out the onset of ambiguity were more likely to be made from the end of the line than from the middle of the line: ($\beta = .85, SE = .34, z = 2.4, p = .01$).

8.5 Discussion

In this discussion section I cover reading time measures, regression behaviour, and scan path measures.

Reading time measures Early reading time measures showed an ambiguity disadvantage (in FPRT). This was predicted by the phrase structure parser in both the surprisal and entropy reduction measures. The dependency parser did not predict this effect, or even its direction, either in the retrieval time measure or in the surprisal measure.

Effects of ambiguity and line position were not significant in the late reading time measures (RPD and TSR). However there were trends towards longer times for ambiguous conditions and longer times for disambiguation at the end of a line than in the middle. These trends were present in the phrase structure parser but not in the dependency parser.

Regression behaviour: probabilistic aspect There was evidence in the PREG measure that people deployed regressions differently according to whether the sentence was ambiguous. people were more likely to deploy a regression when the sentence was ambiguous. Surprisal computed over a phrase structure grammar does predict more difficulty in the ambiguous cases, although it does not follow from high surprisal values that a regression is predicted. Nevertheless the existing surprisal mechanism could be augmented fairly straightforwardly with a module that deploys a regression if the surprisal measure reaches a threshold. Surprisal computed over a dependency grammar was insensitive to the ambiguity effect.

There was evidence from PREG that people were more likely to deploy a regression from disambiguation at the end of a line than they were from disambiguation in the middle of the line. Both the dependency and phrase structure computational parsers were unable to represent this difference. Furthermore in the case of line position effects neither parser is equipped to predict line position effects at all since neither parser has a representation of the physical layout of the text. This is a serious limitation for both parsers that could not be overcome by the addition of a threshold-based module.

Regression behaviour: spatio-temporal aspect The regressions that people deployed had different spatio-temporal properties. 46% of these regressions were classified as indicative of selective reanalysis rather than being refixations on the disambiguation or short regressions that went back only one or two words. Within the regressions that were indicative of selective reanalysis, the ambiguous end of line condition yielded a greater proportion than any other condition (45%) although the ambiguity x line position interaction effect was not significant. When the disambiguated conditions were removed from the analysis there was a clear effect of line position such that the end of line position was more likely to yield selective reanalysis scan paths than the middle of line position.

Overall conclusion In this experiment the reading time measures considered together did not produce significant effects. Instead the behavioural variance was exhibited in the the frequency of producing a regressive scan path, and in the shape of the regressive scan paths that people made. The dependency parser was completely insensitive to experimental manipulation. The phrase structure parser predicted the direction of the trend for an ambiguity disadvantage that was weakly present in the human data, but was not sensitive to the significant effects of layout on the frequency and shape of regressive scan paths.

I now consider the contribution of the experiment to the focal theoretical questions of this thesis. (1) Repair or replacement; (2) The purpose of regressive eye movements; and (3) the coverage of the different grammar formalisms.

1. The human data showed that propensity to make a regressive scan path was greater in the ambiguous conditions than in the disambiguated conditions and greatest in the end of line ambiguous conditions. The patterned variance in probability of making a regression does not by itself adjudicate between the repair and replacement positions unless the target of the regression also varies. Evidence that the regressions tended to be directed at different targets according to the position of the disambiguating word on screen is taken to be support for the repair account. This is because replacement does not offer a reason for targeting a particular part of the input so far for re-inspection whereas repair offers that the onset of ambiguity is targeted for re-inspection because of the information that it contains with respect to a successful parse of the sentence. When the verb with ambiguous subcategorisation is re-inputted by a regressive scan path this allows a lexicon look-up to propose an alternative subcategorisation, and that alternative subcategorisation yields the correct parse of the sentence. The identity of the target matters for evaluating repair against replacement: the fact that the target was not the beginning of the sentence is important. If the parser was targeting the beginning of the sentence this would indicate that the parser was merely starting again and was unable to make use of the structure so far. The fact that the regressions targeted a non-sentence-initial location suggests that replacing the whole parse was unnecessary, and that only the structurally ambiguous verb needed to be repaired.

2. This experiment supports the view that regressive scan paths are made with the purpose of re-analysing a linguistically relevant part of the sentence so far. This can be seen by comparing the results with the time-out hypothesis that predicts a random distribution of regression targets as a function of distance from the launch site. Such an account is ruled out by the finding that regressions tend to target the onset of ambiguity.

3. The different grammar formalisms differed greatly with respect to their coverage of the human data. The dependency parser was unable to distinguish

between the conditions: the disambiguating word was predicted to cause equal difficulty in each condition. The performance of the phrase structure parser was better in that it was sensitive to ambiguity, although not to line position. For this reason the phrase structure parser does not offer complete coverage of the human data, and this boils down to the inability of these parsers to accommodate the physical layout of the text that clearly plays a role in the human processing of these sentences. This is a only result of the parser's training data not including line breaks.

Chapter 9

E3: Effects of the location of misanalysed material

In this chapter I take the data from (Mitchell et al., 2008) Experiment 2 and submit it to novel analyses of eye movement measures and scan path analysis. Novel results from computational parser simulations are also reported. The location of the misanalysed area of the sentences was manipulated and expected to influence regressive eye movements and scan paths. However, there was only a non-significant tendency for regressive scan path shape to be influenced by the location of misanalysis.

9.1 Introduction

In this chapter I take the data from (Mitchell et al., 2008) Experiment 2 and submit it to novel analyses of eye movement measures and scan path analysis.

The design crossed ambiguity with the location of misanalysed material. The manipulation of the location of misanalysed material can be seen in Figure 9.1. The verb with ambiguous subcategorisation (i.e., *watched*) and the first noun following it (e.g., *the juggler*) are kept close together, but the group of words (i.e., *watched the juggler*) appears either on the left of the sentence or shifted across the sentence to the right by lengthening the start of the sentence. Whether the material appears more leftwards on screen or more rightwards on screen is referred to as a manipulation of the *location of the misattached material*.

A strict interpretation of the Selective Reanalysis hypothesis predicts that linguistic content attracts regressions. In this experiment it is the onset of ambiguity at *watched the juggler* (called the *misanalysis area* in Mitchell et al. (2008)) that is linguistically informative for correct processing of the disambiguating word *swallowed*. Since the location of this misanalysis area onscreen is manipulated across the conditions, under this proposal returns should be made to an earlier (more leftwards) part of the sentence for early misanalysis conditions, and to later (more rightwards) parts of the sentence in late misanalysis area conditions. However, the original Mitchell et al. (2008) data analysis was unable to demonstrate significant effects of this kind.

The materials used sentences like Examples 9.1 and 9.2.

(9.1) While the mob watched(,) the juggler who was gifted and nimble swallowed a silver sword that was very sharp.

(9.2) The busy guide noted that while the mob watched(,) the juggler swallowed a silver sword that was very sharp.

Figure 9.1 shows the manipulation of misanalysis area for Early misanalysis area and Late misanalysis area respectively. The figure shows the materials as they were laid out in the experiment, with true line breaks and a monospaced font like the one used in the experiment to preserve the relative positions of the words on screen. The misanalysis area is underlined here.

```
While the mob watched(,) the juggler who was gifted and nimble swallowed
a silver sword that was very sharp.

The busy guide noted that while the mob watched(,) the juggler swallowed
a silver sword that was very sharp.
```

Figure 9.1: Figure shows how the manipulation of misanalysis area was implemented

9.2 Method

The method section gives details of participants, apparatus, and procedure.

Participants This is quoted from the original paper: “32 volunteers participated. All were students at the University of Exeter and aged between 18 and 23. Nine were male and 23 were female. 21 of the participants were paid in cash.”

Apparatus The apparatus was the same as in section 7.2 on page 120.

Procedure The procedure was the same as in section 7.2 on page 121.

9.3 Simulation data

In the models of the computational parsers' predictions, the interaction term specified at each level of item had to be dropped before the models would converge, leaving only main effects specified at each level of item.

Dependency surprisal The dependency parser's surprisal predictions showed an ambiguity x location interaction ($\beta = .48, SE = .003, t = 147.92$) such that the early head position conditions showed no effect of the location of the misattached material, whereas in the late head position conditions the disambiguated condition was predicted to be more difficult than the ambiguous condition. The human data showed no sign of this pattern.

Dependency retrieval The dependency parser's predictions for retrieval time were for a small disadvantage for late head position in the ambiguous conditions turning into a larger disadvantage for late head position in the disambiguated conditions ($\beta = 39.5, SE = .47, t = 84.06$). There was no sign of this pattern in the human data.

Phrase structure surprisal The phrase structural parser's predictions for surprisal were for a disadvantage for early head position in the ambiguous conditions and a larger disadvantage for late head position in the disambiguated conditions ($\beta = .38, SE = .02, t = 21.50$). There was no sign of this pattern in the human data.

Phrase structure entropy reduction The phrase structural parser's predictions for entropy reduction showed an ambiguity effect with the ambiguous cases leading to more entropy reduction than the disambiguated conditions ($\beta = .32, SE = .05, t = 6.75$). There was also a significant tendency for a smaller disadvantage

Table 9.1: Computational measures at the disambiguating word

Ambiguity	Misanalysis Area	DSURP	DTIME	TSURP	ER
ambiguous	early	2.55	229.88	13.45	1.67
ambiguous	late	1.93	300.65	11.76	1.95
disambiguated	early	2.55	218.93	13.73	0.80
disambiguated	late	4.13	442.15	10.49	1.37

[a] Please see page 14 for the abbreviations used in the table

for late head position in the ambiguous cases to grow larger in the disambiguated conditions ($\beta = .03, SE = .0016, t = 18.083$) The entropy reduction parser made the best predictions of the human data for 'late' eye movement measures. The human measures are presented in the next section.

9.4 Eyetracking data

This section gives the results from the eyetracking data. The data are summarised in Table 9.2. For each measure, a LMER model was constructed that modelled the measure as a function of word length, word frequency, ambiguity, the location of the misanalysed material, the ambiguity x location interaction, as well as random effects of participant and item and slopes for the ambiguity, location, and ambiguity x location interaction at each level of participant and item. Where the model had to be simplified, the necessary simplification is described in the section for that measure.

First fixation duration No effects were significant in FFD. Effects of word length and word frequency were $\beta = 6.26$, $SE = 6.70$, $t = .937$ and $\beta = -5.82$, $SE = 3.42$, $t = -1.7$ respectively. There was a marginal main effect of ambiguity with ambiguous condition tending to lead to longer times than disambiguated controls: $\beta = 14.95$, $SE = 9.39$, $t = 1.59$. The location of the misattached noun phrase did not exert a significant influence on times: $\beta = -1.01$, $SE = 7.68$, $t = -.13$. The ambiguity x location interaction was non-significant: $\beta = 6.33$, $SE = 7.81$, $t = .81$.

Table 9.2: Human measures at the disambiguating word

Ambiguity	Misanalysis Area	FFD	FPRT	RPD	TSR	PREG
ambiguous	early	226	304	498	194	0.24
ambiguous	late	221	287	498	212	0.30
disambiguated	early	215	222	236	14	0.10
disambiguated	late	199	265	349	84	0.17

[a] Please see page 14 for the abbreviations used in the table

First pass reading time Effects of word length and frequency were not significant in the model: $\beta = 8.18$, $SE = 11.15$, $t = .73$ and $\beta = -9.21$, $SE = 5.70$, $t = -1.62$ respectively. There was a significant main effect of ambiguity with ambiguous conditions leading to longer times than disambiguated controls: $\beta = 28.46$, $SE = 12.39$, $t = 2.30$. The location of the misattached noun phrase did not exert a significant influence on first pass reading time as a main effect ($\beta = -8.5$, $SE = 11.56$, $t = -.74$) and the ambiguity x location interaction was non-significant ($\beta = 20.12$, $SE = 13.44$, $t = 1.50$).

Regression path duration Effects of word length and word frequency were non-significant: $\beta = 19.95$, $SE = 28.23$, $t = .71$ and $\beta = -10.16$, $SE = 14.27$, $t = -.71$ respectively. There was a significant main effect of ambiguity with ambiguous conditions leading to more regression path duration than disambiguated controls: $\beta = 95.84$, $SE = 31.25$, $t = 3.1$. The location of the misattached noun phrase did not exert a significant main effect ($\beta = -28.96$, $SE = 33.69$, $t = -.86$), and the ambiguity x location interaction was not significant ($\beta = 25.52$, $SE = 32.20$, $t = .79$).

Probability of a regression Effects of word length and word frequency were non-significant: $\beta = .17$, $SE = .17$, $z = .98$, $p = .33$ and $\beta = -.13$, $SE = .08$, $z = -1.57$, $p = .12$ respectively. There was a significant main effect of ambiguity with ambiguous conditions leading to greater probability of regression than disambiguated controls: $\beta = .73$, $SE = .32$, $z = 2.30$, $p < .05$. The main effect of the location of the misattached noun phrase did not exert a significant influence on probability of a regression, although there was a tendency for early location to reduce regression probability: $\beta = -.52$, $SE = .30$, $z = -1.72$, $p = .09$. The

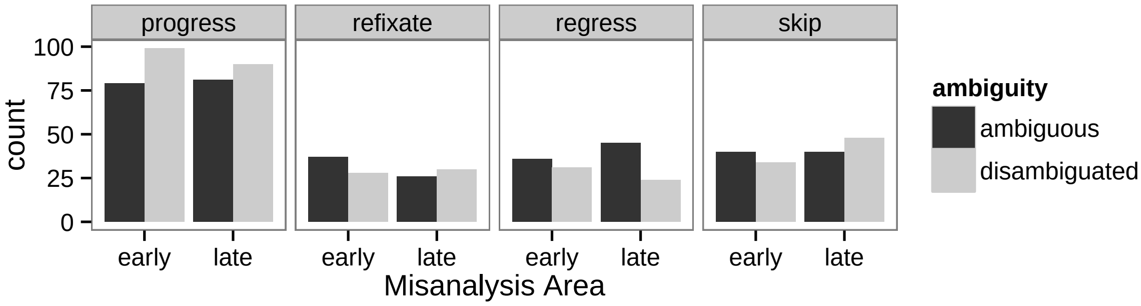


Figure 9.2: Distribution of coarse eye movement behaviours at disambiguation

ambiguity x location interaction was not significant: $\beta = .309$, $SE = .30$, $z = 1.03$, $p = .30$.

Scan path analysis First I consider the distribution of coarse eye movement behaviours, plotted in Figure 9.2. We can see that the most common behaviour was to progress, and that for the regress cases, they appear to be over-represented in the ambiguous conditions. Two sets of scan path analyses were carried out. The first included all regressions and the second discarded regressions that only went back one word. Each analysis is described separately below.

Analysis of full data set For the analysis that included all regressions, Figure 9.3 shows the ten clusters returned by scan path similarity analysis. The patterns that regressed to the preceding word (i.e, A,B,C,D,E,F,J) were grouped into a supercluster that represented a regression to the preceding word. Then a model of membership of this supercluster was constructed to see whether the strategy of regressing to the previous word was over-distributed in any particular condition. Figure 9.4 shows the mean proportion of trials that ended up as members of this supercluster for each condition. There was a tendency for people to regress to the preceding word more often in the early head position sentences than in the late head position sentences, and more in the ambiguous conditions than the disambiguated conditions. The LMER model had to be simplified before it would converge. Per-subject and per-item slopes had to be dropped. The model showed that the ambiguity effect was not significant ($\beta = .19, SE = .28, z = .68, p = .49$), that the head position effect was marginal ($\beta = .55, SE = .29, z = 1.92, p = .054$) and that the ambiguity x head position effect was non-significant ($\beta = .11, SE =$

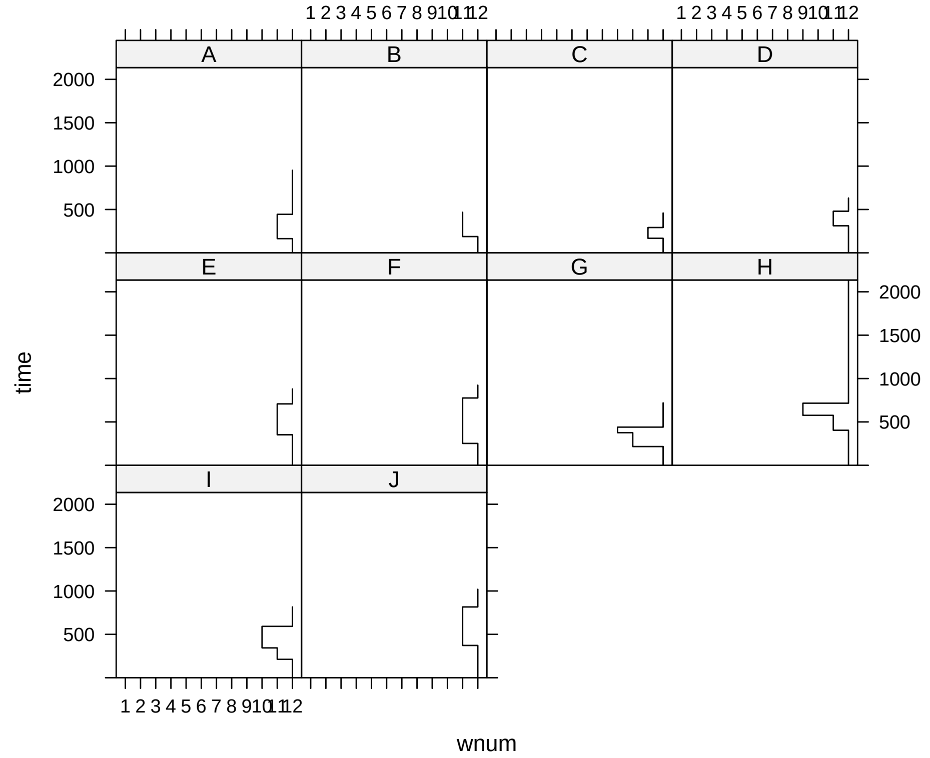


Figure 9.3: scan path strategies at disambiguation in the full analysis

$.29, z = .26, p = .71$).

Analysis of restricted data set For the analysis that discarded regressions to the preceding word, Figure 9.5 shows the most common strategies after the restriction was applied. These were then classified according to whether or not they constituted a move back to the ambiguously attached main clause verb *watched*. Figure 9.6 shows the probability that a given condition would lead to a regression of this type. On average, this type of movement was over represented in the ambiguous late head position sentences (as would be expected under a selective reanalysis account). The LMER model had to be simplified before it would converge: per-item terms for the ambiguity x head position effect had to be dropped. However the LMER model of this outcome showed that the over-representation was not statistically significant (main effect of ambiguity: $\beta = .32$, $SE = .36$, $z = .986$, $p = .37$, main effect of head position: $\beta = -.2, SE = .31, z = -.65, p = .52$, ambiguity x head position interaction: $\beta = -.29$, $SE = .30$, $z = -.974$, $p = .33$).

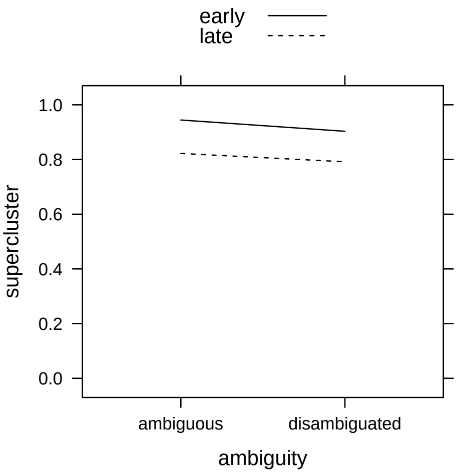


Figure 9.4: Distribution, over the conditions, of regressions to the preceding word, in the full analysis

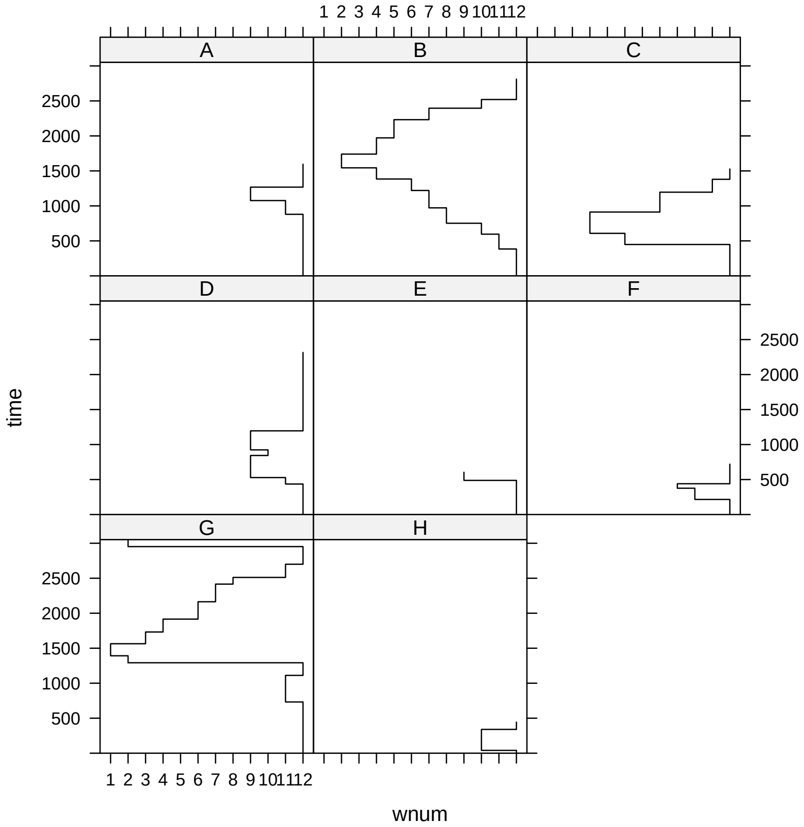


Figure 9.5: scan path strategies at disambiguation in the restricted analysis

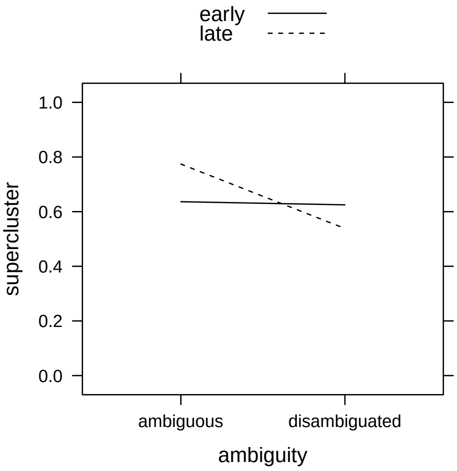


Figure 9.6: Probability of a regression to word 9: *watched* in the late head position sentences or *gifted* in the early head position conditions, for the restricted analysis

9.5 Discussion

Although there were significant main effects of ambiguity in FPRT, RPD, and PREG, the location of the misanalysed material did not exert significant influences on standard eye movement measures. This echoes the data from the Mitchell et al (2008) ANOVA analyses, except that the LMER analysis also detected a main effect of ambiguity in PREG that the ANOVA analyses missed. In the more detailed Regression Signature analyses from the original paper, there were hints of differential returns to the ambiguously attached material – when the material was in the early misanalysis position, more returns to that location were observed in the ambiguous conditions than the unambiguous conditions, and when the ambiguously attached material was in late misanalysis position, more returns were made to that location in the ambiguous than the unambiguous conditions, but this pattern was not significant on materials analysis. SCASIM analysis was unable to add to this evidence. The SCASIM analysis of all regressive movements showed that a strategy of regressing to the immediately preceding word was not over-distributed in any particular condition. When these one-word regressions were discarded from the data set and the remaining longer regressions were analysed, these too were not over distributed in any particular condition. The novel analyses were able to add only the main effect of ambiguity in PREG (that the ANOVA analysis missed) to the original Mitchell et al. (2008) conclusions.

For the matter of which computational parser best approximated the human data, the phrase structure grammar parser’s entropy reduction measure was the only computational measure that anticipated the shape of the human disambiguation difficulty.

Chapter 10

E4: Verb subcategorisation frames affect regressions

This experiment introduce verbs that have a different subcategorisation preference than those examined so far. These verbs are those like ‘noticed’ that can take a sentential complement. The Diagnosis model of Fodor and Inoue (1998) predicts that these verbs are processed differently from verbs like ’saluted’. This experiment tests for effects of different subcategorisation frames in their interaction with effects of ambiguity. Overt complementisers are used to pre-disambiguate the ’noticed’ forms. This experiment shows that the extra difficulty of disambiguating the NP/Z forms shows up in longer reading times, greater regression frequencies, and also the relative over-dispersal of scan path patterns that target the onset of ambiguity.

10.1 Introduction

The linguistic phenomenon under consideration is the difficulty people have resolving an ambiguity afforded by particular verbs with more than one subcategorisation frame. In Example 10.1 the verb *saluted* has only one subcategorisation frame for a direct object. In Example 10.2 the verb *noticed* has two subcategorisation frames, one for a direct object - e.g., *noticed the siren* and the other for a complement such as a clause - e.g., *noticed the siren had been sounded.* A noun phrase like *the captain* that follows a verb with multiple subcategorisation frames like *noticed* can be attached preferentially as its object, or alternatively as the subject of its complement. Such a noun phrase, if attached as the object of a subordinate clause, is revealed to be the subject of the main (complement) clause instead, when the main clause is indicated by the appearance of its verb at disambiguation.

Ambiguity with respect to subcategorisation can be resolved by cues in the text that indicate which subcategorisation is appropriate. For example, a clause barrier can be overtly marked by punctuation: a complement clause can be initialised overtly by a complementiser. When sentences are pre-disambiguated in this way, difficulty at the first verb can be measured when it is not carrying disambiguating information, which provides a baseline for comparison with reading difficulty at the first verb in ambiguous versions of the sentences, when the first verb has additional consequences for parsing.

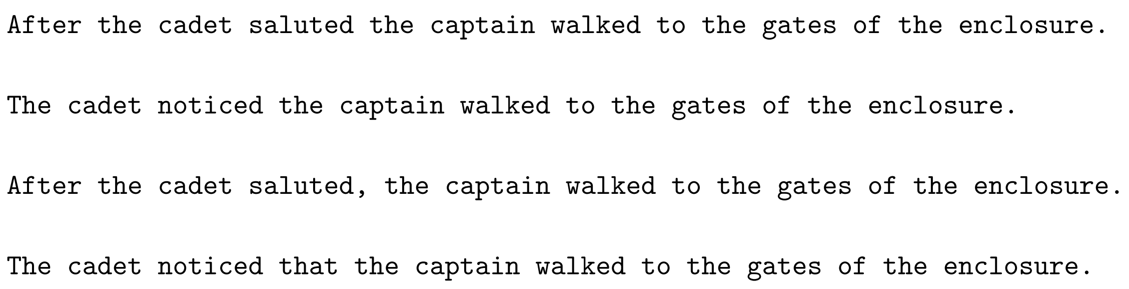
The design crossed two factors. These were ambiguity (the sentence either was or was not disambiguated by a comma immediately following the first verb) and subcategorisation frame (the verb was either NP/S or NP/Z).

With reference to the descriptions of capture and theft given in section 3.4, we can say that the diagnosis model predicts an interaction between ambiguity and sentence type such that a theft sentence should have a larger ambiguity cost than an equivalent capture sentence. The larger ambiguity cost is due to the parser's inability to attach the initially wrongly attached material as a complement of the verb to which it is wrongly attached in the theft condition - in the capture condition the parser is able to attach this material easily as a complement.

Twenty-four experimental sentences were generated in each of the four conditions. Examples of each condition follow. Example 10.1 represents the ambiguous and disambiguated versions of the theft sentences (disambiguating comma in parentheses), and Example 10.2 represents the ambiguous and disambiguated versions of the capture sentences (disambiguating overt complementiser in parentheses). A phrasemarker showing the correct analysis of the theft sentences may be found in Figure 10.2 and a phrasemarker representing the correct analysis of the capture sentences may be found in Figure 10.3. Please note that one item (item 18) had to be removed from the analysis because in one condition, a programming error introduced an additional word into the text display. The additional word was an adjective that would not be expected to unduly complicate the parsing for that sentence, but which rendered the sentence inconsistent in terms of word numbering and word order with other sentences in the set. The full set of sentences presented may be found in Appendix A.

(10.1) After the cadet saluted(,) the captain walked to the gates of the enclosure. *(Theft)*

(10.2) The cadet noticed (that) the captain walked to the gates of the enclosure. *(Capture)*



```
After the cadet saluted the captain walked to the gates of the enclosure.

The cadet noticed the captain walked to the gates of the enclosure.

After the cadet saluted, the captain walked to the gates of the enclosure.

The cadet noticed that the captain walked to the gates of the enclosure.
```

Figure 10.1: Sentence layout showing how the materials were arranged on screen, in the following order: ambiguous NP/Z; ambiguous NP/S; control NP/Z; control NP/S

10.2 Method

The method section gives details of participants, apparatus, and procedure.

Participants Participants were forty native speakers of British English who were students of Psychology at the University of Exeter, were given partial course credit to participate in the experiment. All had normal or corrected to normal vision, were naive to the purpose of the experiment, and were aged between eighteen and thirty-four.

Apparatus An SR Research Eyelink II head-mounted eyetracker was used to record participants' eye movements with a sampling rate of 500 Hz. Participants read sentences displayed on a 19 inch Iiyama Vision Master Pro video monitor at 1024 x 768 resolution at a refresh rate of 60 Hz. Viewing was binocular but only the right eye was recorded. Participants sat in a dimly lit room in front of the computer at a viewing distance of approximately 75 cm the average viewing distance

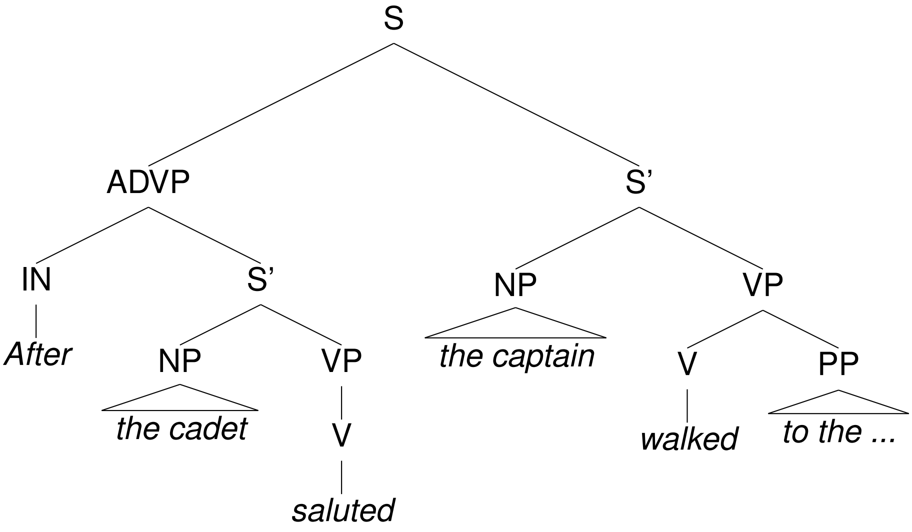


Figure 10.2: Phrasemarker for theft sentences

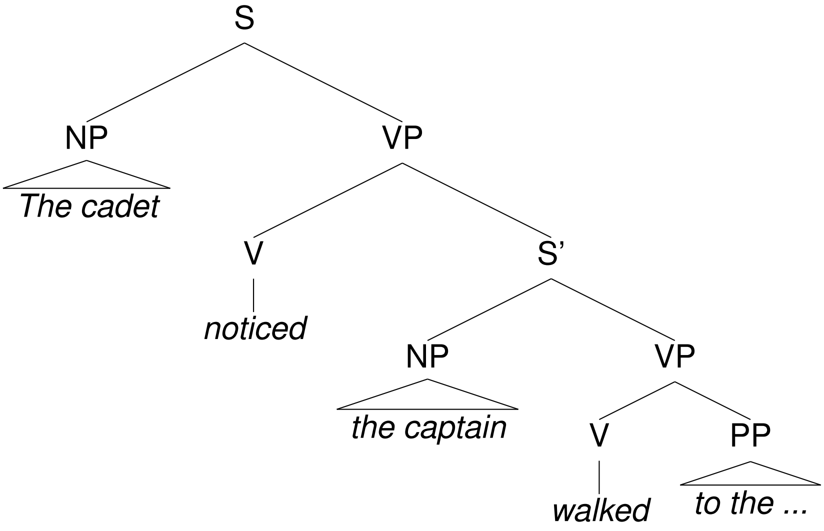


Figure 10.3: Phrasemarker for capture sentences

was approximately 75 cm. At this viewing distance, and assuming that 1 character had 2 mm width on screen, a single character subtended 0.153° of visual angle, and approximately 6.5 characters subtended 1° of visual angle. The font used was Courier New 12 point. All sentences in this experiment were displayed on a single line with a maximum length of 100 characters. A 9 point calibration procedure was used, on which participants were required to achieve a score of 'good'. Each trial started with a drift correction routine where the participant was required to fixate a target that appeared in the same location as the first character of the sentence would subsequently occupy, and then required to press a button on the gamepad while fixating this point to start the trial.

Procedure Participants were instructed to read silently for comprehension at a comfortable speed. The practice trials and experimental trials were implemented as separate consecutive blocks. The experimental trials were randomised by Experiment Builder each time the experiment was run, i.e., in a different order for each participant, with the constraint that a maximum of two trials of a given type could appear in a continuous sequence. There were four practice sentences, followed by a drift correction routine preceding the experimental block containing 96 sentences, comprising 24 in experimental conditions (6 in each of 4 conditions); 24 foils and 48 fillers. Participants were rotated over one of four lists, implementing a Latin square design. 32 of the trials (including 8 of the experimental conditions) were followed immediately by a comprehension question. This was a simple question about the sentence immediately preceding that required the participant to make a yes or no response using the appropriate trigger button on the gamepad. The full list of questions asked may be found in the Appendix. The whole procedure took about 20 to 40 minutes, depending on the participant.

10.3 Simulation data

The following predictions of processing difficulty at the disambiguating word were derived: surprisal from the dependency parser; surprisal from the phrase structure parser; and entropy reduction from the phrase structure parser.

The predictions are given in Table 10.1. For each predicted measure a multilevel model was constructed. Each model included terms for centred log word frequency and centred word length; terms for the fixed effects ambiguity, sentence

Table 10.1: Computational measures at the disambiguating word

Ambiguity	sentence type	DSURP	DTIME	TSURP	ER
ambiguous	NP/S	1.26	209	9.36	2.37
ambiguous	NP/Z	1.26	209	10.28	2.64
unambiguous	NP/S	1.08	280	8.76	2.07
unambiguous	NP/Z	1.43	209	8.29	1.67

[a] Please see page 14 for the abbreviations used in the table

type and the ambiguity x sentence type interaction; and terms for the random effects of ambiguity, sentence type and the ambiguity x sentence type interaction at each level of item. These models did not include terms for subjects because the computational measures did not vary over subjects. The results of each model are given in the relevant section below.

Dependency surprisal The mean values of dependency surprisal at the disambiguating word show that ambiguous NP/S and ambiguous NP/Z are predicted to be equal. For the unambiguous cases NP/Z is predicted to be more difficult than NP/S. The term for the interaction effect over items had to be dropped in order to achieve model convergence, but individual terms for ambiguity and sentence type could be computed at each level of item. The model indicated that word length and word frequency did not exert significant effects on the dependency surprisal measure ($\beta = -0.002$, $SE = 0.008$, $t = -0.29$ and $\beta = -0.008$, $SE = 0.007$, $t = -1.12$ respectively). Ambiguity did not exert a significant effect on dependency surprisal ($\beta = 0.0002$, $SE = 0.01$, $t = 0.01$). The effect of sentence type was significant, with NP/Z causing more dependency surprisal ($\beta = -0.09$, $SE = 0.01$, $t = -6.26$). The ambiguity x sentence type interaction was very significant in the model ($\beta = 0.09$, $SE = 0.002$, $t = 39.67$). For the NP/S sentences the control sentence was easier – there was an ambiguity disadvantage. For the NP/Z sentences, the control condition was more difficult – there was an ambiguity advantage.

Dependency retrieval The mean values for retrieval predicted that ambiguous NP/S, ambiguous NP/Z and unambiguous NP/Z should be equally difficult, with unambiguous NP/Z predicted to cause the most difficulty The term for the interaction effect over items had to be dropped in order to achieve model convergence, but individual terms for ambiguity and sentence type could be computed at each level of item. Word length and word frequency did not influence the dependency retrieval measure ($\beta = -0.01$, $SE = 0.20$, $t = -0.06$ and $\beta = -0.004$, $SE = 0.17$, $t = -0.02$ respectively). Main effects of ambiguity and sentence type were significant in the model ($\beta = -17.7$, $SE = 0.60$, $t = -29.72$ and $\beta = 17.7$, $SE = 0.6$, $t = 29.72$ respectively). There was a significant ambiguity x sentence type interaction ($\beta = -17.7$, $SE = 0.09$, $t = -191.25$).

Phrase structure surprisal Phrase structure surprisal predicted that the ambiguous cases would be harder then the unambiguous cases; and that the NP/Z disadvantage in the ambiguous cases would turn around into a NP/S disadvantage in the unambiguous conditions. The term for the interaction effect over items had to be dropped in order to achieve model convergence, but individual terms for ambiguity and sentence type could be computed at each level of item. Word length was not important in the phrase structure parser's surprisal measure ($\beta = -0.18$, $SE = 0.18$, $t = -1.03$). Word frequency did exert a significant effect ($\beta = -0.97$, $SE = 0.15$, $t = -6.41$). Effects of ambiguity, sentence type and the ambiguity x sentence type interaction were all significant in the model ($\beta = 0.65$, $SE = 0.05$, $t = 12.32$, $\beta = -0.11$, $SE = 0.03$, $t = -3.25$, and $\beta = -0.35$, $SE = 0.01$, $t = -62.35$ respectively).

Phrase structure entropy reduction The directions of the entropy reduction hypothesis predictions were the same as for phrase structure surprisal, although there was a relatively greater difficulty with the NP/S cases versus surprisal. The term for the interaction effect over items had to be dropped in order to achieve model convergence, but individual terms for ambiguity and sentence type could be computed at each level of item. Word length and word frequency were both unimportant in the model ($\beta = .05$, $SE = .11$, $t = .44$ and $\beta = .02$, $SE = .09$, $t = .17$ respectively). Effects of ambiguity, sentence type and the ambiguity x sentence type interaction were all significant in the model ($\beta = 0.32$, $SE = 0.02$, $t = 14.04$, $\beta = -0.03$, $SE = 0.02$, $t = -2.05$, and $\beta = -0.17$, $SE = 0.002$, $t = -55.79$ respectively).

10.4 Eyetracking data

For each dependent variable a model was constructed that contained terms for word length, word frequency, ambiguity, sentence type, the ambiguity x sentence type interaction, and terms for the random effects of subject and item, as well as terms for the main effects of ambiguity and sentence type and the ambiguity x sentence type interaction at each level of subject and item.

First fixation duration First fixation duration detected a main effect of ambiguity, and a marginal effect of sentence type but no interaction effect. The effect of word length was not significant ($\beta = 4.74, SE = 3.59, t = 1.32$) and neither was the effect of word frequency ($\beta = -4.01, SE = 2.93, t = -1.37$). The ambiguous conditions resulted significantly in approximately 30 ms more FFD than the disambiguated controls ($\beta = 29.13$, $SE = 7.90$, $t = 3.69$). The main effect of sentence type was a weak trend towards approximately 14 ms more FFD in the NP/Z conditions than in the NP/S conditions ($\beta = -14.23$, $SE = 8.11$, $t = -1.76$). The NP/Z disadvantage tended to be larger in the ambiguous cases than in the disambiguated cases, but the interaction effect was very weak ($\beta = -2.82$, $SE = 6.93$,$t = -.41$).

First pass reading time Word length exerted a non-significant influence ($\beta = 10.22, SE = 5.58, t = 1.83$). More frequent words resulted in less FPRT and the effect was significant ($\beta = -9.66, SE = 4.63, t = -2.09$). There was a significant ambiguity disadvantage of 31 ms ($\beta = 29.28$, $SE = 12.36$, $t = 2.37$). There was a significant effect of sentence type with NP/S reducing FPRT ($\beta = -19.78$, $SE = 10.36$, $t = -1.91$), but hardly any hint of an interaction effect ($\beta = 1.33$, $SE = 11.32$, $t = .12$).

Regression path duration Word length and word frequency both exerted non-significant influences ($\beta = -0.82$, $SE = 14.31, t = -0.06$ and $\beta = -16.32, SE = 11.72, t = -1.40$ respectively). There was a significant effect of ambiguity with the ambiguous conditions leading to 146 ms more RPD than the disambiguated conditions ($\beta = 135.15$, $SE = 37.60$, $t = 3.56$). There was a significant NP/Z disadvantage of 79 ms as a main effect ($\beta = -68.59$, $SE = 30.66$, $t = -2.27$); and a significant interaction effect with the NP/Z disadvantage increasing in the ambiguous conditions ($\beta = -64.28$, $SE = 31.33$, $t = -2.05$).

Table 10.2: Human measures at the disambiguating word[a]

Ambiguity	sentence type	FFD	FPRT	RPD	TSR	PREG
ambiguous	NP/S	257	295	382	87	0.16
ambiguous	NP/Z	275	317	534	217	0.24
unambiguous	NP/S	231	265	309	43	0.09
unambiguous	NP/Z	241	284	314	29	0.07

[a] Please see page 14 for the abbreviations used in the table

Time spent regressing Word length and word frequency exerted non-significant effects ($\beta = -8.92$, $SE = 11.81$, $t = -0.80$ and $\beta = -10.50$, $SE = 9.74$, $t = -1.08$ respectively). Ambiguity significantly increased TSR by 116 ms on average ($\beta = 106.30$, $SE = 35.55$, $t = 2.99$). There was a non-significant main effect of sentence type with the NP/Z conditions tending to attract more TSR ($\beta = -49.97$, $SE = 30.47$, $t = -1.64$). The ambiguity x sentence type interaction approached significance, with the ambiguity disadvantage tending to be much greater in the NP/Z conditions than in the NP/S conditions ($\beta = -63.86, SE = 34.01, t = -1.88$).

Probability of a regression Probability of regression was conditionalised on there being a valid first pass fixation in the word: i.e., if the word was skipped, a missing value was recorded. This model was implemented with a binomial link function appropriate to the binary-valued data representing whether on the trial a first pass regression was launched from the disambiguating word. This was effected by adding the argument $family = "binomial"$ to the call to the LMER function. This explains why the statistic reported for this measure is z rather than t. Word length did not significantly affect PREG ($\beta = -.20, SE = .11, z = -1.70, p = .09$). Word frequency exerted a significant influence ($\beta = -.28$, $SE = .09$, $z = -3.02$, $p < .01$). The effect of ambiguity was significant, taking the proportion of trials from $.08$ to $.20$ ($\beta = 1.57, SE = .32, z = 4.88, p < .001$). The effect of sentence type was non-significant ($\beta = .30, SE = .34, z = .89, p = .37$). There was a marginal ambiguity x sentence type interaction ($\beta = -.60, SE = .32, z = -1.88, p = .059$).

scan path analysis In this section I present analyses of the spatio-temporal dynamics of scan paths that were launched from the disambiguating word and

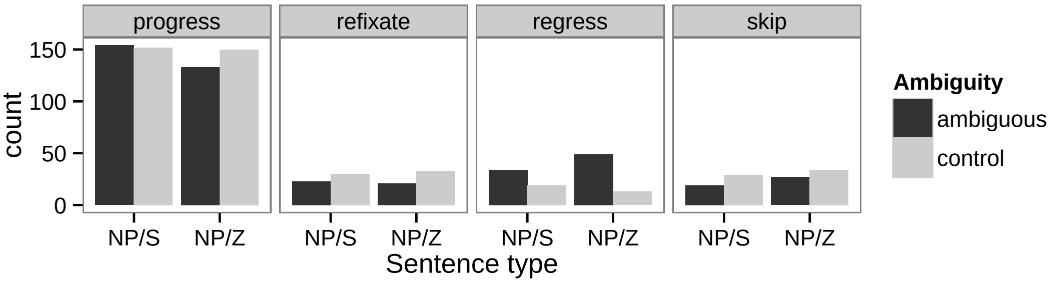


Figure 10.4: Distribution of coarse behaviours at disambiguation

that ended just before a fixation in new material.

Distribution of disambiguation behaviours In this section I consider a coarse-grained analysis of behaviour at disambiguation using the classification set out in the section above. 920 trials remained from 960, after excluding all trials from item 18 as described above. The most common pattern was *progress*, with 589 trials (64%). The next most frequent pattern was *regress* with 115 trials (13%). *Skip* accounted for 109 trials (12%). 107 trials resulted in *refixate* (11%). One can see that the critical ambiguous NP/Z condition was the least frequent cause of *progress* behaviour, and the most frequent cause of *regress* behaviour.

Distribution of regression strategies In this section I will limit the scope of enquiry to the spatio-temporal *regress* cases. I will retain each of the 115 trials (representing 13% of all trials) that resulted in this behaviour and discard the others. This leaves a set of scan paths that started with a legitimate first pass fixation in the disambiguating word, and then possibly refixated it some number of times, and then took in upstream words, and then possibly refixated the disambiguating word some number of times, and then ended by including the last fixation before new material was taken in. One wishes to know whether within this set there were any patterns across the treatment conditions. I approach the problem by clustering the scan paths into self-similar groups, and asking for each group of scan paths whether the likelihood that a given scan path is a member of the group varies as a function of the fixed factors *ambiguity* and *sentence type* and their interaction *ambiguity x sentence type*.

This approach requires a measure of self-similarity. The measure should ideally tell us for each scan path in a pair of scan paths how self-similar each is to the other. Ideally the measure should be sensitive to spatial as well as temporal differences between the scan paths. The measure I use is SCASIM (*sca*npath *sim*ilarity), provided in the R library SCAN PATH (von der Malsburg, 2010), as used in von der Malsburg and Vasishth (2011) to investigate syntactic reanalysis.

This initial scasim run is done in order to check for outlier trials that could skew the later parts of the analysis. I applied a threshold at 2.5 s.d. either side of the mean which excluded one trial.

At this stage the matrix of pairwise dissimilarity values is very large ($114*114 = 12996$ dimensions). This needs to be scaled down to a more manageable unit. From the tools available in R I use the function ISOMDS because it considers various dimensionalities and assigns to each dimensionality a stress value that serves as a measure of goodness of fit and an optimal number of clusters. This allows one to choose a dimensionality with reference to how much goodness of fit one is willing to sacrifice in order to get a reasonable number of clusters. Consider briefly that one could in the limit have a model with a cluster for each scan path with no loss of goodness of fit at all, but which would be useless as a model. At the other limit one could derive the single most typical scan path at the cost of nearly all goodness of fit: again useless for present purposes. Figure 10.5 shows that stress reduces smoothly with increased number of dimensions, and that a 5 dimensional model with 8 clusters is available at a stress of about 10%. The optimal number of clusters at a given dimensionality is identified by the function Mclust using a mixture of gaussians approach and the information criterion BIC to prefer the model that minimises BIC.

For each cluster, I obtained the scan path closest to its centroid. This is a sort of measure of central tendency, similar to the mean, but for cases like these where the mean scan path is not a coherent notion – instead one chooses the scan path that was actually made that sits closest to the centroid of the cluster. Note that I use the term centroid rather than center because the clusters are imposed on the data, and there may be no data point at the theoretical centroid of a given cluster. These scan paths are plotted in Figure 10.6.

The next step is to consider how these 8 patterns might constitute eye movement strategies for disambiguation. In so doing one may legitimately group some of these patterns together if they identify the same eye movement strategy (von

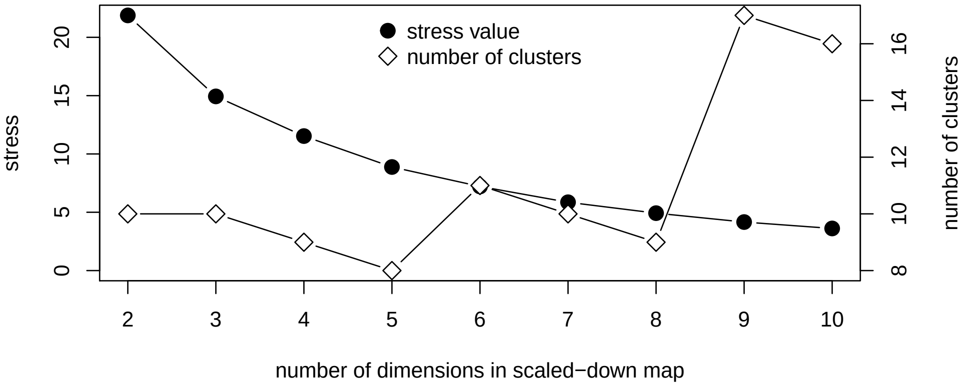


Figure 10.5: Illustration of method for choosing a lower dimensional map of scan path similarity

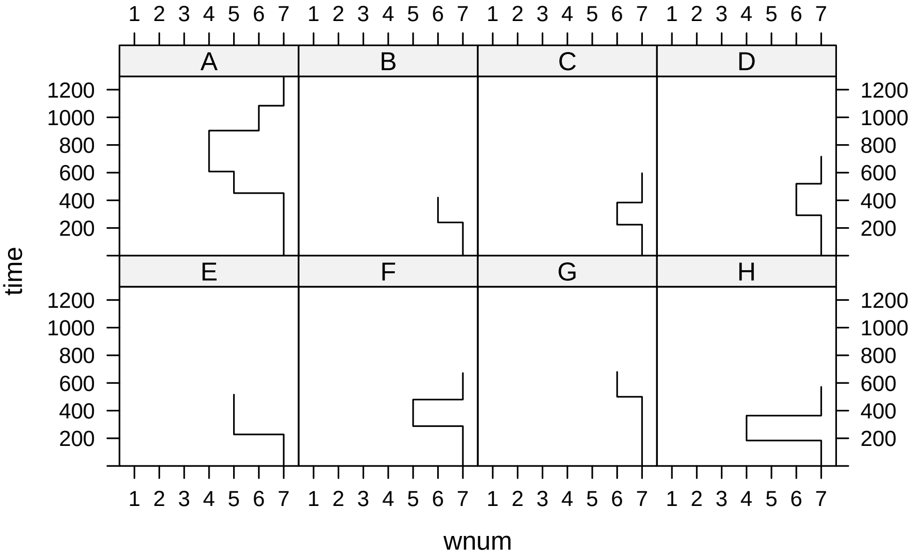


Figure 10.6: scan path behaviours at disambiguation

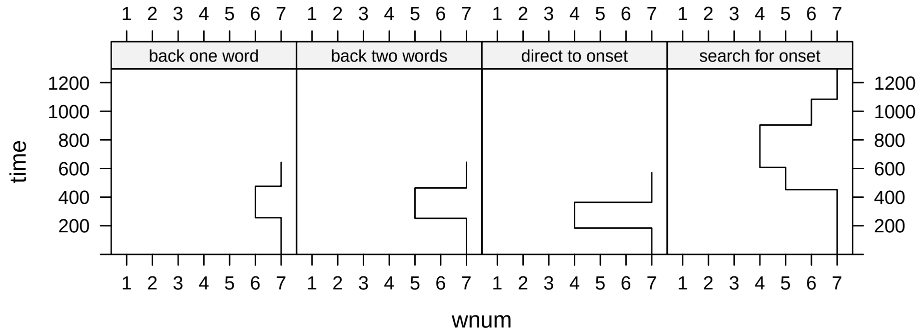


Figure 10.7: Regression strategies at disambiguation

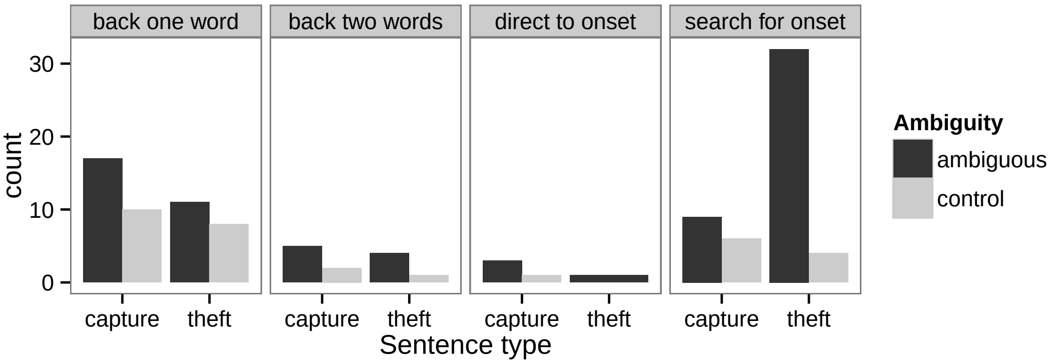


Figure 10.8: Distribution of regression strategies

der Malsburg & Vasishth, 2011). In these data the 8 patterns form 4 reading strategies that are illustrated in Figure 10.7. These are *checking back one word*; *checking back two words*; *direct to onset* (i.e., direct to onset of ambiguity) and *search for the onset of ambiguity*.

Figure 10.8 shows that the *search for onset* supercluster appears to be distributed unevenly in the manner predicted by the Diagnosis model, with the ambiguous NP/Z condition generating more than other conditions.

I tested this formally with a multilevel model describing membership of this supercluster as a function of condition. Trials that did not lead to a regression of this type were coded 0 [1] and the outcome was treated as binomially distributed. The maximal model failed to converge, so I dropped the by-participant and by-item random slopes from the model but retained their intercepts. Results are

[1] the other option was to encode non-regressions as missing data, but I feel that every trial had the opportunity to be a member and that therefore failing to become a member constitutes non-membership (0) rather than exclusion from the possibility of membership (NA).

reported with Bonferroni-corrected p values. The model showed that the ambiguous conditions resulted in greater likelihood of the regression type *search for onset* ($\beta = 1.38$, $SE = .42$, $z = 3.27$, Bonferrroni-adjusted $p < .01$): mean probability for ambiguous was $.09$; mean for control was $.02$. There was no significant main effect of sentence type. There was a significant ambiguity x sentence type interaction ($\beta = 1.08$, $SE = .42$, $z = -2.54$, Bonferroni-adjusted $p = 0.04$). This took the form of no sentence type effect in the control conditions, but a large and significant NP/Z disadvantage in the ambiguous conditions.

Model evaluation results In this section the predictions from the computational parsers are evaluated against the human behavioural results. RPD correlated positively with ER and TSURP (Pearson's $r(809) = .16, p < .001$ and $r(809) = .24, p < .001$ respectively). Dependency surprisal bore no relationship to RPD ($r(809) = -.03, p = .42$) while dependency retrieval time was *negatively* correlated with RPD ($r(809) = -0.12, p < .001$).

10.5 Discussion

First I consider the results from analysis of the standard em measures. First I will recapitulate the results. In every measure, the ambiguous cases significantly caused much more difficulty than the disambiguated versions. This shows both that disambiguation was effective on the whole, and that people found the ambiguous cases hard to process at disambiguation: there was a parser load associated with carrying out disambiguation processes. The fact that FFD captured this effect suggests that it exerts its influence immediately. The fact that the effect manifested in the later measures too indicates that there was a load imposed by ambiguity on higher-order parsing processes such as integration. Considered as a main effect, sentence type influenced only RPD, where temporally longer regressions were found as a result of NP/Z compared with NP/S. However, sentence type exerted an influence on the ambiguity disadvantage in its interaction effect that was captured in the RPD measure as well as the PREG measure, where the fairly large disadvantage associated with ambiguous NP/Z (versus ambiguous NP/S) was significantly greater than the disadvantage associated with disambiguated NP/Z (which was often negligible or even negligibly reversed versus disambiguated NP/S).

Now I consider the theoretical implications of the results. The prediction that the Diagnosis model makes for these sentences, of a NP/Z disadvantage in integration processes and repair processes, holds up fairly well under this scrutiny. The evidence shows that early measures were relatively unaffected by sentence type, consistent with the prediction, and that late measures were subject to the predicted inflation, but only in the ambiguous cases. This latter caveat is important because it shows that the disadvantage associated with NP/Z obtains in late measures at disambiguation, but not when the same linguistic material is pre-disambiguated. Therefore it is safe to conclude that the ambiguous NP/Z condition is throwing the parser into difficulties, and that these difficulties take longer to resolve than the equivalent difficulties in ambiguous NP/S.

However, these analyses of standard measures have not yet licensed the claim that the repairs are in fact different in nature across the two conditions, only that they take longer for NP/Z (measured in RPD). The demonstration that repairs are qualitatively different is found in the spatio-temporal dynamics of the scan paths launched from the disambiguating word. The pattern denoted “search for onset" was significantly overrepresented in the ambiguous NP/Z condition. This shows that the regressive landing points are not simply the result of a random walk over the preceding words. Instead it is clear that certain sentence types lead to strategic patterns of eye movements. A natural interpretation of these strategic patterns is that when parsing breaks down at disambiguation in the ambiguous NP/Z case, readers are seeking out the onset of ambiguity.

The surprisal measures present different coverage of the human data depending on whether the surprisal was computed over a dependency parser or a phrase structure grammar. The phrase structure grammar tracked human difficulty in late parsing measures well, in the aggregate measure and also in both of its lexical and syntactic subcomponents. In contrast, the dependency parser's surprisal was sensitive to the ambiguity in the NP/S cases in the same way as humans but made the wrong prediction for the effect of ambiguity in the NP/Z cases.

One explanation for the wrong prediction of the dependency parser for the NP/Z cases is that the ambiguity on the NP/Z cases was effected by leaving out a piece of clause-marking punctuation. While the phrase structure parser processed the disambiguation following punctuation in line with the human data, the dependency parser appears to predict that processing the comma makes disambiguation harder not easier. This may be because the punctuation is considered as a terminal in its own right in the dependency grammar, and dependency arcs are assigned to the relations between the punctuation terminals and the lexical terminals. The phrase structure parser also treats the punctuation as a terminal in its own right, but makes much more out of the information provided by the punctuation. Since the phrase structure parser explicitly assigns hierarchical structure to the input where the dependency parser does not, it is better placed than the dependency grammar parser to benefit from the clause boundary information that is inherent in the punctuation but speaks to a level higher than the terminal level. This clause-marking information makes parsing at disambiguation easier for both the human parser and the phrase structure parser but not for the dependency parser.

In contrast the dependency parser is indeed capable of responding in a human-like way to the overt complementiser in the NP/S conditions. This complementiser carries both sequential terminal information and higher-level clause marking information. The phrase structure parser benefits from this information, but so does the dependency parser. One explanation for the dependency parser's capitalisation on the complementiser's presence, but not on the punctuation's presence is the suggestion that the complementiser might have better signal to noise ratio than the comma. The reasoning here is that commas demarcate, for example, lists or prosodic boundaries just as commonly as they do clauses, whereas the only irrelevant use of the string *that* is as a demonstrative pronoun, so that the clause-marking information conveyed by an overt complementiser to a dependency parser might be more useful than the clause-marking information carried by a comma to a dependency parser.

Chapter 11

E5: Head position interaction effects: part 1 short phrases

In this experiment, an ambiguously attached NP is manipulated. The NP is extended by adding a qualifier. For example, a noun phrase like the captain *is extended by a qualifier like* of the squadron *to yield the extended NP* the captain of the squadron*. A further manipulation is the location of the head noun of the extended NP within the extended NP. When the qualifier comes before the head noun, the head noun is said to be in late head position, and is closer to the disambiguating word. When the qualifier comes after the head noun, the result is that the head noun is in early head position, increasing the distance between the head noun and the disambiguation. Early head position is said to be harder than late head position because the misattached head is further away from disambiguation.*

11.1 Introduction

This section builds upon the findings of the previous chapter and provides a test of the Diagnosis model. We will see that the results provide some support for the Diagnosis model's prediction in terms of how likely it is that a regression will be made in the service of disambiguation: and what spatio-temporal shape such a regression will take.

The linguistic phenomenon under consideration is the difficulty people have resolving an ambiguity afforded by particular verbs with more than one subcategorisation frame. A noun phrase that follows such a verb can be attached preferentially as its object, or alternatively as the subject of its complement. Such a noun phrase, if attached as the object of a subordinate clause, is revealed to be the subject of the main (complement) clause instead, when the main clause is indicated by the appearance of its verb at disambiguation.

In the sentences used in this pair of experiments, I manipulate whether the verb preceding the critical noun phrase does, or does not, permit a sentential complement. This affects the ease with which the critical phrase can be moved. Within the critical noun phrase, the head of the relative clause is either the initial or the final constituent of the clause. When at the start of the relative clause, there is a long distance between the head and the disambiguating verb: when at the end of the clause, a short distance. This distance is also affected by the length of the relative clause itself, which was varied across the two experiments.

This experiment used sentences with a short ambiguous phrase. The manipulation of head position amounted to putting the syntactic head of the phrase in a phrase-initial or a phrase-final position in the surface form. Fixed effects were *head position*, *sentence type*, and their interaction effect *head position* x *sentence type*. *Head position* referred to whether the syntactic head of the ambiguous noun phrase appeared early in the clause or late in the clause and had two levels: *early*, and *late.* sentence type referred to whether the required repairs were licensed or not by the GDP (section 3.4) and had two levels: *NP/S* for licensed cases, and *NP/Z* for unlicensed cases. Random effects were *participant* and *item*.

The head position effect is predicted by the Decay model and others that share its commitment to the decay of activation with increasing content intervening between the onset of activation and its later measurement at disambiguation. The prediction is that the early head position sentences should cause more difficulty than the late head position sentences because of more intervening material between the onset and resolution of syntactic ambiguity. There is no prediction for a difference between the sentence types used here, so no head position by sentence type interaction is predicted either. The Decay model is compatible with the interaction but does not predict it, and cannot account for it.

The Diagnosis model predicts that sentence type will modulate the head position effect, with NP/S yielding benefits only while the activation has not yet fallen below a threshold for detection. The prediction is that NP/Z should be harder than NP/S and that the extra difficulty due to NP/Z should be greater only for the late head position sentences. For the early head position sentences the Diagnosis model predicts that insufficient error signal remains at disambiguation for the NP/S sentence type to bring about a benefit versus NP/Z.

In these materials, the syntactic head of the ambiguous noun phrase appears either early in phrase-initial position or late in phrase-final position. The ambiguous noun phrase has two forms. The early form is achieved by adding words after the noun whereas the late form is achieved by adding words before the noun. The words appended in the early condition constituted a prepositional phrase like ’of the squadron’. In the late condition, the noun was preceded by an adjective to give an adjectival noun phrase like ‘the distressed patient’, or a noun to give a compound noun phrase like ‘the squadron captain’. Examples of the sentences used in this experiment follow. The sentences were presented in a monospaced font and with all the words on a single line.

(11.1) The cadet noticed the captain of the squadron walked to the gates of the enclosure. *(Early NP/S)*

(11.2) The cadet noticed the squadron captain walked to the gates of the enclosure. *(Late NP/S)*

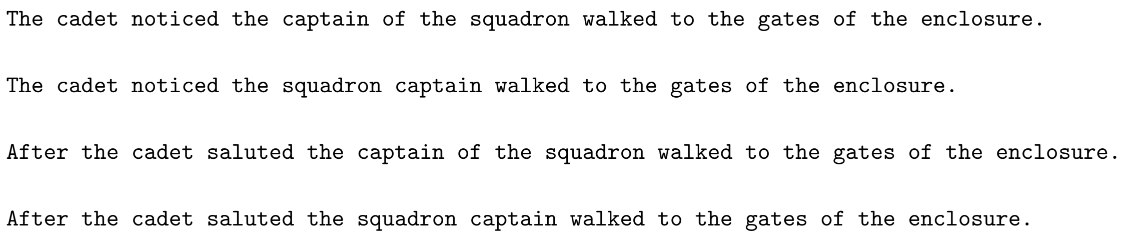
(11.3) After the cadet saluted the captain of the squadron walked to the gates of the enclosure. *(Early NP/Z)*

(11.4) After the cadet saluted the squadron captain walked to the gates of the enclosure. *(Late NP/Z)*

Materials like these that exhibit head-position effects have been used in previous work by, e.g., Ferreira and Henderson (1991b); Sturt et al. (1999); Tabor and Hutchins (2004).

11.2 Method

The method section gives details of participants, apparatus, and procedure.



```
The cadet noticed the captain of the squadron walked to the gates of the enclosure.

The cadet noticed the squadron captain walked to the gates of the enclosure.

After the cadet saluted the captain of the squadron walked to the gates of the enclosure.

After the cadet saluted the squadron captain walked to the gates of the enclosure.
```

Figure 11.1: Figure shows how the sentences were arranged on screen

Participants Participants were twenty-four native speakers of British English who were students of Psychology at the University of Exeter, were given partial course credit to participate in the experiment. All had normal or corrected to normal vision, were naive to the purpose of the experiment, and were aged between eighteen and thirty-four.

Apparatus The apparatus used was the same as in the previous chapter. The description is given again here for convenience. An SR Research Eyelink II head-mounted eyetracker was used to record participants' eye movements with a sampling rate of 500 Hz. Participants read sentences displayed on a 19 inch Iiyama Vision Master Pro video monitor at 1024 x 768 resolution at a refresh rate of 60 Hz. Viewing was binocular but only the right eye was recorded. Participants sat in a dimly lit room in front of the computer at a viewing distance of approximately 75 cm the average viewing distance was approximately 75 cm. At this viewing distance, and assuming that 1 character had 2 mm width on screen, a single character subtended 0.153° of visual angle, and approximately 6.5 characters subtended 1° of visual angle. The font used was Courier New 12 point. All sentences in this experiment were displayed on a single line with a maximum length of 100 characters. A 9 point calibration procedure was used, on which participants were required to achieve a score of 'good'. Each trial started with a drift correction routine where the participant was required to fixate a target that appeared in the same location as the first character of the sentence would subsequently occupy, and then required to press a button on the gamepad while fixating this point to start the trial.

Procedure The procedure was the same as the procedure in the previous experiment. The description is given again here for convenience. Participants were instructed to read silently for comprehension at a comfortable speed. The practice trials and experimental trials were implemented as separate consecutive blocks. The experimental trials were randomised by Experiment Builder each time the experiment was run, i.e., in a different order for each participant, with the constraint that a maximum of two trials of a given type could appear in a continuous sequence. There were four practice sentences, followed by a drift correction routine preceding the experimental block containing 96 sentences, comprising 24 in experimental conditions (6 in each of 4 conditions); 24 foils and 48 fillers. Participants were rotated over one of four lists, implementing a Latin square design. 32 of the trials (including 8 of the experimental conditions) were followed immediately by a comprehension question. This was a simple question about the sentence immediately preceding that required the participant to make a yes or no response using the appropriate trigger button on the gamepad. The full list of questions asked may be found in Appendix A. The whole procedure took about 20 to 40 minutes, depending on the participant.

11.3 Simulation data

The following measures of processing difficulty at the disambiguating word were computed: surprisal from the dependency parser; retrieval time from the dependency parser; surprisal from the phrase structure parser; entropy reduction from the phrase structure parser. The predictions are given in Table 11.1. For each predicted measure a multilevel model was constructed. Each model included terms for the following: centred log word frequency; centred word length; the fixed effects of head position and sentence type and their interaction; and terms for the fixed effects and their interaction at each level of item. These models did not include terms for subject because the measures did not vary over subjects. The results of each prediction model are given below in the relevant section.

Dependency surprisal The interaction effect over items had to be dropped to achieve model convergence but individual terms for head position and sentence type could be computed at each level of item. The model indicated that effects of word length and word frequency were not significant ($\beta = .01$, $SE = .04$, $t = .38$

Table 11.1: Computational measures at the disambiguating verb

Head Position	sentence type	DSURP	DTIME	TSURP	ER
early	NP/S	1.21	356	11.66	0.79
early	NP/Z	1.65	351	12.85	0.98
late	NP/S	1.30	228	9.74	1.14
late	NP/Z	1.37	220	10.66	1.22

and $\beta = .010$, $SE = .03$, $t = .31$ respectively). There was no main effect of head position ($\beta = .05$, $SE = .07$, $t = .68$). There was only a marginal main effect of sentence type with NP/Z tending to be harder than NP/S ($\beta = -.13$, $SE = .07$, $t = -1.69$). The sentence type x head position interaction effect was significant ($\beta = -.09$, $SE = .01$, $t = -8.27$): the NP/Z disadvantage was greater in the early head position sentences than in the late head position sentences.

Dependency retrieval The interaction effect over items had to be dropped to achieve model convergence but individual terms for head position and sentence type could be computed at each level of item. Word length and word frequency had significant effects in the model ($\beta = 57.8$, $SE = 21.99$, $t = 2.63$ and $\beta = 106$, $SE = 5.4$, $t = 19.64$ respectively). The head position effect was significant, with early head position more difficult than late head position ($\beta = 66.72$, $SE = 6.21$, $t = 10.75$). There was no main effect of sentence type ($\beta = 1.1$, $SE = 2.4$, $t = .45$). The sentence type x head position interaction was just significant, with a smaller NP/S disadvantage in early than late head position ($\beta = .99$, $SE = .53$, $t = 1.893$).

Phrase structure surprisal The interaction effect over items had to be dropped to achieve model convergence but individual terms for head position and sentence type could be computed at each level of item. Word length and word frequency had significant effects in the model ($\beta = -.33$, $SE = .17$, $t = -1.93$ and $\beta = -1.22$, $SE = .11$, $t = -10.72$ respectively). The main effect of head position was significant with early harder than late ($\beta = 1.00$, $SE = .14$, $t = 7.16$), as was the main effect of sentence type with NP/Z harder than NP/S ($\beta = -.51$, $SE = .10$, $t = -4.97$), and the head position x sentence type interaction ($\beta = -.09$, $SE = .02$, $t = -5.22$) with the sentence type effect bigger for early than late head position.

Table 11.2: Human parser measures at the disambiguating verb

Head Position	sentence type	FFD	FPRT	RPD	TSR	PREG
early	NP/S	277	331	453	122	0.16
early	NP/Z	273	338	563	225	0.25
late	NP/S	255	306	374	68	0.11
late	NP/Z	267	302	432	130	0.17

Phrase structure entropy reduction The interaction effect over items had to be dropped to achieve model convergence but individual terms for head position and sentence type could be computed at each level of item. Word length and word frequency had significant effects in the model ($\beta = -.27$, $SE = .13$, $t = -2.1$ and $\beta = -.35$, $SE = .04$, $t = -8.56$ respectively). There was no main effect of head position ($\beta = -.15$, $SE = .17$, $t = -.89$) but there was a significant main effect of sentence type ($\beta = -.06$, $SE = .02$, $t = -3.16$) and a significant head position x sentence type interaction ($\beta = -.03$, $SE = .004$, $t = -7.64$).

11.4 Eyetracking data

This section gives the results from analyses of the human data. For each human measure, a LMER model was fitted, with model specification given here: fixed effects for head position and sentence type and their interaction; covariates for centred word length and centred logged word frequency; random intercepts for participants and items, and random slopes for the head position x sentence type interaction effect at each level of participants and items, i.e. a maximal model.

First fixation duration There were no significant effects in first fixation duration. Effects of word length and word frequency were both non-significant ($\beta = 2.70, SE = 4.22, t = 0.64$ and $\beta = -1.31, SE = 3.42, t = -0.38$ respectively). The main effect of head position was a non-significant disadvantage for early head position ($\beta = 14.208, SE = 10.45, t = 1.35$); the main effect of sentence type was a small non-significant NP/Z disadvantage ($\beta = -3.75, SE = 11.46, t = -.33$). There was a non-significant trend towards an interaction ($\beta = -8.02, SE = 9.74, t = .82$).

First pass reading time Effects of word length and word frequency were both non-significant ($\beta = 9.03, SE = 6.20, t = 1.48$ and $\beta = -.26, SE = 5.09, t = -.05$ respectively). There was a robust disadvantage for early head position in first pass reading time as a main effect ($\beta = 30.67, SE = 12.49, t = 2.45$). Neither the main effect of sentence type nor the head position x sentence type interaction was significant in FPRT ($\beta = -3.85, SE = 19.59, t = -0.19$ and $\beta = -3.82, SE = 13.22, t = -.29$ respectively).

Regression path duration Effects of word length and word frequency were both non-significant ($\beta = 9.11, SE = 18.25, t = -.50$ and $\beta = -11.04, SE = 14.91, t = -.74$ respectively). There was a significant main effect of head position with a disadvantage for early head position ($\beta = 105.79, SE = 48.98, t = 2.16$). There was a trend towards a disadvantage of NP/Z as a main effect that did not reach significance ($\beta = -87.49, SE = 46.45, t = -1.88$). The interaction effect was non-significant ($\beta = -27.66, SE = 42.90, t = -.65$).

Time spent regressing Effects of word length and word frequency were both non-significant ($\beta = -15.41, SE = 16.08, t = -.96$ and $\beta = -9.53, SE = 13.14, t = -.73$ respectively). A trend towards a disadvantage for early head position approached significance ($\beta = 77.85, SE = 47.20, t = 1.65$). There was a significant disadvantage for NP/Z ($\beta = -86.07, SE = 41.71, t = -2.06$). The interaction was not significant ($\beta = -23.97, SE = 42.26, t = -0.57$).

Probability of a regression Effects of word length and word frequency were both non-significant ($\beta = -.10, SE = .12, t = -.84, p = .40$ and $\beta = -.12, SE = .09, t = -.13, p = .19$ respectively). The trend towards a disadvantage of early head position approached reliability ($\beta = .66, SE = .35, z = 1.86, p = .06$). There was a significant disadvantage for NP/Z ($\beta = -.85, SE = .30, z = -2.76, p < .01$). The interaction was not significant ($\beta = .12, SE = .30, z = .41, p = .68$).

Scan path analysis In this section the focus is on the sequence and duration of regressive fixations launched from disambiguation. The distribution of disambiguation behaviours is treated in the next part, and the distribution of regression strategies over the conditions in the following part).

Distribution of disambiguation behaviours Here I give an overview of behaviours at disambiguation arranged by type. Types were: (1) skip; (2) progress; (3) refixate; (4) regress. The distribution of these behaviours over the treatment conditions is plotted in Figure (11.2). We can see that the progress behaviour is the most common, and that skipping was very uncommon. The regress cases appear to be unevenly distributed over the conditions.

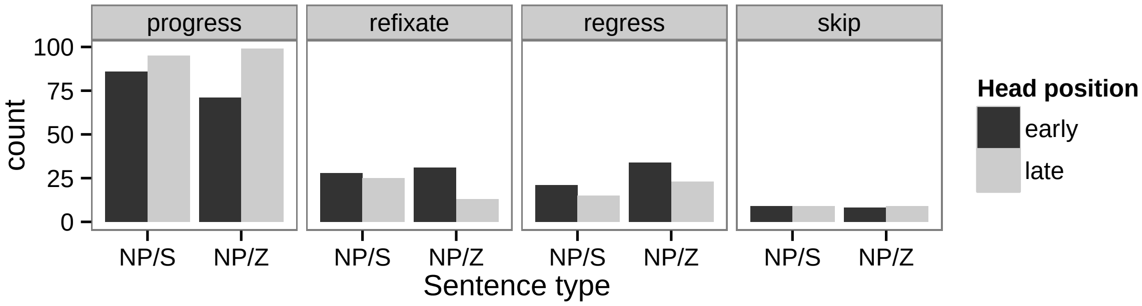


Figure 11.2: How coarse eye movement behaviours launched from disambiguation were distributed over conditions

Distribution of regression strategies Here the scope narrows to include only regressions (from disambiguation). An initial scasim run was done to look for outliers. The initial scasim run did not yield any outliers using a 2.5 s.d. threshold so the full data set numbering 93 regressions was carried forward and subjected to multidimensional scaling and cluster analysis. Figure (11.3) shows the method for selecting a number of dimensions for the scaled down map. A 3 dimensional map provided stress of 13.57 and 3 clusters. in a model described as 'spherical, varying volume (VII) with 3 components'. These components are represented in Figure (11.4). Cluster A contained long regressions into early parts of the sentence.

Table 11.3: Coefficients for fixed effects from a multilevel model of membership of distribution of cluster A

	Estimate	Std. Error	z value	Pr($>$ |z|)
Head Position	0.13	0.22	0.60	0.55
Sentence Type	-0.25	0.23	-1.10	0.27
Head Position x Sentence Type	0.58	0.22	2.58	0.01

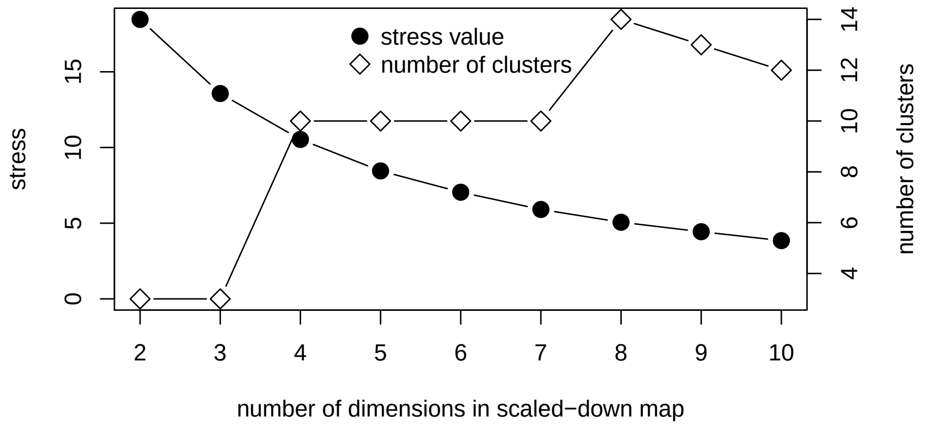


Figure 11.3: Plot illustrates how a lower dimensional representation of scan path similarity space was chosen

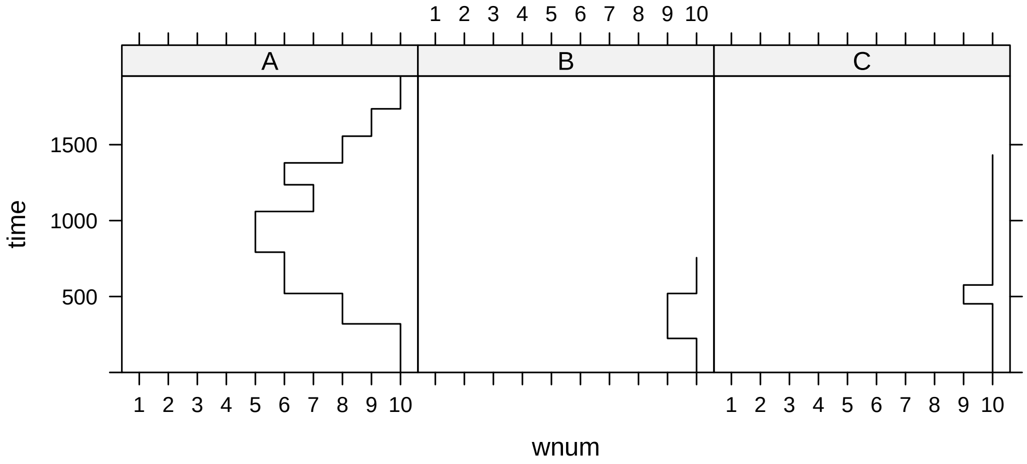


Figure 11.4: Plot shows the scan path that sits nearest the centroid of each cluster. scan paths from different conditions had different numbers of words before disambiguation. Sentences were aligned with disambiguation at word 10 by padding the start of sentences with fewer words using null words. This means that the word number 10 is the disambiguation in every condition, but leftmost extent of the regression pattern typified as A is a fixation that falls in different words across conditions. The reason for pursuing cluster A is not that it targeted a particular word, but that it moved furthest away from disambiguation.

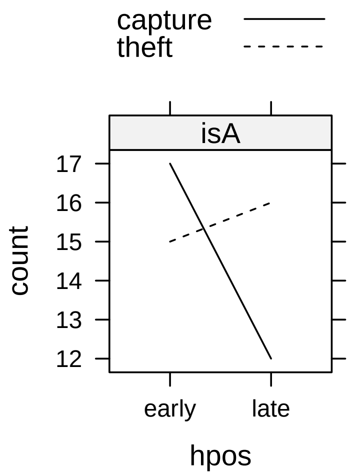


Figure 11.5: Distribution of category A scan paths

The focus is now on the question whether cluster A regressions were evenly distributed over the treatment conditions. Every trial was treated as either belonging or not belonging to cluster A. A multilevel model was constructed that treated head position and their interaction as fixed effects with corresponding random effects in a maximal model. This model showed that although neither main effect was significant there was a significant interaction. The effect of head position manifested differently according to sentence type, with a bigger sentence type difference in late than in early head position.

Model evaluation results In this section the predictions from the computational parsers are evaluated against the human behavioural results. Entropy reduction correctly predicts a NP/Z disadvantage but wrongly predicts a disadvantage for late head position. Entropy reduction was negatively correlated with RPD ($r(539) = -.10$, $p = .02$). Phrase structure surprisal correctly predicts a NP/Z disadvantage and also correctly predicts a disadvantage for early head position. Phrase structure surprisal was positively correlated with RPD ($r(539) = .16$, $p < .001$). Dependency surprisal correctly predicts the NP/Z disadvantage, but only correctly predicts the direction of the early head position disadvantage in the NP/Z cases. In the NP/S cases it wrongly predicts an early head position advantage. Dependency surprisal was weakly positively correlated with RPD ($r(539) = .06$, $p = .18$). Dependency retrieval time correctly predicts an early head position disadvantage but wrongly predicts a disadvantage for NP/S. Dependency retrieval was positively correlated with RPD ($r(539) = .15$, $p < .001$).

11.5 Discussion

First pass reading time and regression path duration showed the expected main effect that early head position induced more reading time than late head position. This disadvantage for early head position was also marginal in frequency of regression.

The crucial question for this experiment was whether the early head disadvantage would manifest differently in the NP/Z (NP, *Z*) cases than in the NP/S NP/S conditions. It was expected that early head position would induce more difficulty than late head position and that NP/Z would be harder than NP/S, and that the disadvantage of NP/Z would be greater in the early head positions than it would be in the late head position sentences.

There was a very significant interaction in the distribution of scan paths that went back further than the ambiguously attached NP, but this interaction effect was not the one predicted. When these scan paths were made from early head position sentences, they were more common in early NP/S than in early NP/Z. When these scan paths were made from late head position sentences this turned around, with more made from late NP/Z than from late NP/S.

Phrase structure surprisal performed best out of the computational predictions.

Chapter 12

E6: Head position interaction effects: part 2 long phrases

This experiment implements a head position manipulation in sentences with longer ambiguously-attached noun phrases. The longer phrases showed significant effects of head position in regression path duration and time spent regressing, as well as interaction effects between head position and sentence type in regression path duration.

12.1 Introduction

This section provides a further test of the Diagnosis model, especially the prediction that sentence type modulates the head position effect. I report the effects of extending the material in the ambiguously attached noun phrase further than was done in the previous experiment. The phrase is extended with an embedded relative clause.

The linguistic phenomenon under consideration is the difficulty associated with moving syntactic units that have some non-negligible internal structural complexity themselves. The question whether this difficulty is influenced by sentence type is addressed.

Fixed effects were *head position*, *sentence type*, and their interaction effect *head position* x *sentence type*. *Head position* referred to whether the syntactic head of the ambiguous noun phrase appeared early in the clause or late in the clause and had two levels: *early*, and *late.* sentence type referred to whether the required repairs were licensed or not by the GDP and had two levels: *NP/S* for licensed cases, and *NP/Z* for unlicensed cases. Random effects were *participant* and *item*.

The head position effect is predicted by the Decay model and others that share its commitment to the decay of activation with increasing content intervening between the onset of activation and its later measurement at disambiguation. The prediction is that the early head position sentences should cause more difficulty than the late head position sentences because of more intervening material between the onset and resolution of syntactic ambiguity. There is no prediction for a difference between the sentence types used here, so no head position by sentence type interaction is predicted either. The Decay model is compatible with the interaction but does not predict it, and cannot account for it.

The Diagnosis model predicts that sentence type will modulate the head position effect, with NP/S yielding benefits only while the activation has not yet fallen below a threshold for detection. The prediction is that NP/Z should be harder than NP/S and that the extra difficulty due to NP/Z should be greater only for the late head position sentences. For the early head position sentences the Diagnosis model predicts that insufficient error signal remains at disambiguation for the NP/S sentence type to bring about a benefit versus NP/Z.

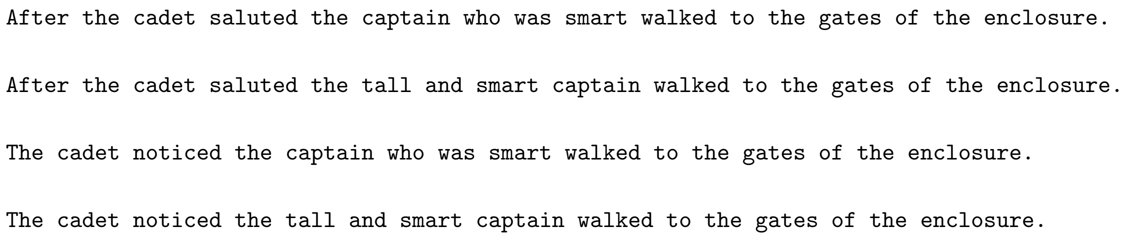
In these materials, the syntactic head of the ambiguous noun phrase appears either early in phrase-initial position or late in phrase-final position. For the early conditions, the ambiguous noun phrase is achieved by adding a relative clause like ‘who was smart’. For the late conditions, the ambiguous noun phrase was achieved by adding an adjectival phrase like ‘the tall and smart’ before the noun. Examples of the sentences used in the experiment follow. The sentences were presented in a monospaced font and with all the words on a single line.

(12.1) After the cadet saluted the captain who was smart walked to the gates of the enclosure. *(Early NP/Z*

(12.2) After the cadet saluted the tall and smart captain walked to the gates of the enclosure. *(Late NP/Z)*

(12.3) The cadet noticed the captain who was smart walked to the gates of the enclosure. *(Early NP/S)*

(12.4) The cadet noticed the tall and smart captain walked to the gates of the enclosure. *(Late NP/S)*



```
After the cadet saluted the captain who was smart walked to the gates of the enclosure.

After the cadet saluted the tall and smart captain walked to the gates of the enclosure.

The cadet noticed the captain who was smart walked to the gates of the enclosure.

The cadet noticed the tall and smart captain walked to the gates of the enclosure.
```

Figure 12.1: Figure shows how the materials were arranged on screen

12.2 Method

The method section gives details of participants, apparatus, and procedure.

Participants Participants were forty native speakers of British English who were students of Psychology at the University of Exeter, were given partial course credit to participate in the experiment. All had normal or corrected to normal vision, were naive to the purpose of the experiment, and were aged between eighteen and thirty-four.

Apparatus The apparatus used was the same as in the previous chapter. The description is given again here for convenience. An SR Research Eyelink II head-mounted eyetracker was used to record participants' eye movements with a sampling rate of 500 Hz. Participants read sentences displayed on a 19 inch Iiyama Vision Master Pro video monitor at 1024 x 768 resolution at a refresh rate of 60 Hz. Viewing was binocular but only the right eye was recorded. Participants sat in a dimly lit room in front of the computer at a viewing distance of approximately 75 cm the average viewing distance was approximately 75 cm. At this viewing distance, and assuming that 1 character had 2 mm width on screen, a single character subtended 0.153° of visual angle, and approximately 6.5 characters subtended 1° of visual angle. The font used was Courier New 12 point.

All sentences in this experiment were displayed on a single line with a maximum length of 100 characters. A 9 point calibration procedure was used, on which participants were required to achieve a score of ‘good’. Each trial started with a drift correction routine where the participant was required to fixate a target that appeared in the same location as the first character of the sentence would subsequently occupy, and then required to press a button on the gamepad while fixating this point to start the trial.

Procedure Participants were instructed to read silently for comprehension at a comfortable speed. The practice trials and experimental trials were implemented as separate consecutive blocks. The experimental trials were randomised by Experiment Builder each time the experiment was run, i.e., in a different order for each participant, with the constraint that a maximum of two trials of a given type could appear in a continuous sequence. There were four practice sentences, followed by a drift correction routine preceding the experimental block containing 96 sentences, comprising 24 in experimental conditions (6 in each of 4 conditions); 24 foils and 48 fillers. Participants were rotated over one of four lists, implementing a Latin square design. 32 of the trials (including 8 of the experimental conditions) were followed immediately by a comprehension question. This was a simple question about the sentence immediately preceding that required the participant to make a yes or no response using the appropriate trigger button on the gamepad. The full list of questions asked may be found in the Appendix. The whole procedure took about 20 to 40 minutes, depending on the participant.

12.3 Simulation data

Both types of parser reproduced the head position effect in their surprisal measures, but only phrase structure surprisal also accounted for the interaction with sentence type. Dependency retrieval time produced wrong predictions. Entropy reduction predicted the wrong head position effect but was sensitive to the sentence type effect.

Dependency surprisal The model had to be greatly simplified before it would converge, retaining only a random intercept for items. Effects of word length and word frequency were both non-significant ($\beta < .01$, $SE < .01, t = .6$ and $\beta < .01$,

Table 12.1: Computational measures at the disambiguating word

Head Position	sentence type	DSURP	DTIME	TSURP	ER
early	NP/S	2.47	199.41	13.04	1.25
early	NP/Z	2.47	199.41	13.72	1.40
late	NP/S	1.21	211.49	9.93	2.05
late	NP/Z	1.21	211.49	10.87	2.16

$SE < .01$, $t = 1.1$ respectively). The head position effect was significant ($\beta < .01$, $SE < .01$, $t = 383.3$). Effects of sentence type and the sentence type x head position interaction were both effectively incalculable ($\beta < .01$, $SE < .01$, $t = 0.0$ and $\beta < .01$, $SE < .01$, $t = 0.0$ respectively).

Dependency retrieval Dependency retrieval did not capture any of the human effects. It was insensitive to sentence type and predicted the wrong direction for the head position effect. Effects of word length and word frequency were both non-significant ($\beta < -.01$, $SE < .01$, $t = -.48$ and $\beta < -.01$, $SE < .01$, $t = -.50$ respectively). The head position effect was significant, but in the wrong direction ($\beta = -6.0$, $SE < .01$, $t = -23.44$). Effects of sentence type and the sentence type x head position interaction were both effectively incalculable ($\beta < -.01$, $SE < .01$, $t = 0.0$ and $\beta < .01, SE < .01, t = 0.0$ respectively).

Phrase structure surprisal The model had to be simplified before it would converge. The interaction effect over items had to be dropped but slopes were retained for each of the main effects over items. The effect of word length was not significant ($\beta = -.22, SE = .14, t = -1.5$). The effect of word frequency was significant ($\beta = -1.06, SE = .12, t = -8.93$). Phrase structure surprisal has the right slope for the head position effect, which was significant ($\beta = 1.49, SE = .14, t = 10.71$). The effect of sentence type was significant and in the same direction as the human measures ($\beta = -.40, SE = .11, t = -3.71$). It is also sensitive to the interaction effect with sentence type, with the effect in the same direction as in the human measures ($\beta = .07, SE = .01, t = 6.09$).

Phrase structure entropy reduction The model had to be simplified before it would converge. The interaction effect over items had to be dropped but slopes were retained for each of the main effects over items. Word length and word frequency were not significant in the model ($\beta = .06, SE = .08, t = .758$ and $\beta = .03, SE = . - 7, t = .502$ respectively). The main effect of head position was significant, showing a disadvantage of late head position in the opposite direction to phrase structure surprisal ($\beta = -.39, SE = .10, t = -4.09$). The main effect of sentence type was a significant NP/Z disadvantage ($\beta = -.07, SE = .02, t = -2.67$). There was a significant head position x sentence type interaction ($\beta = -.007, SE = .003, t = -2.47$).

12.4 Eyetracking data

In this section I give the results from analysis of the standard em measures. For each measure, a LMER model was fitted, with model specification given here: fixed effects for head position and sentence type and their interaction; covariates for centred word length and centred logged word frequency; random intercepts for participants and items, and random slopes for the head position x sentence type interaction effect at each level of participants and items, i.e. a maximal model.

First fixation duration No effects were significant in FFD. Effects of word length and word frequency were $\beta = 3.01, SE = 2.8, t = 1.06$ and $\beta = -1.468, SE = 2.2, t = -0.66$ respectively. There was a slight tendency towards a disadvantage of early head position, but it was non-significant ($\beta = 7.30, SE = 6.59, t = 1.11$). There was no effect of sentence type ($\beta = -1.24, SE = 8.19, t = -.15$). The head position x sentence type interaction was non-significant ($\beta = -3.78, SE = 6.34, t = -.60$).

First pass reading time The only significant effect in FPRT was of word length ($\beta = 9.8, SE = 4.05, t = 2.44$). The effect of frequency was non-significant ($\beta = -2.17, SE = 3.25, t = -0.67$). The main effect of head position was not significant ($\beta = -.43, SE = 8.85, t = -.05$). The main effect of sentence type was not significant ($\beta = -14.80, SE = 10.01, t = -1.48$). The interaction effect was not significant ($\beta = 9.08, SE = 9.13, t = .99$).

Table 12.2: Human measures at the disambiguating word

Head Position	sentence type	FFD	FPRT	RPD	TSR	PREG
early	NP/S	248	285	438	153	0.21
early	NP/Z	253	291	508	218	0.29
late	NP/S	246	276	385	108	0.17
late	NP/Z	242	300	382	82	0.11

Regression path duration Effects of word length and word frequency were non-significant ($\beta = 16.72, SE = 16.43, t = 1.02$ and $\beta = -8.76, SE = 13.09, t = -0.67$ respectively). RPD yielded a significant disadvantage of early head position ($\beta = 87.38, SE = 30.31, t = 2.88$). The main effect of sentence type was non-significant ($\beta = -32.29, SE = 28.17, t = -1.15$). There was a trend for the early head position disadvantage to be larger for NP/Z than for NP/S but this was non-significant. ($\beta = -34.63, SE = 28.71, t = -1.21$)

Time spent regressing Effects of word length and word frequency were non-significant ($\beta = 3.82, SE = 16.09, t = 0.24$ and $\beta = -9.78, SE = 12.71, t = -0.77$ respectively). The results from analysis of TSR were the same as for RPD, with a disadvantage of early head position ($\beta = 86.93, SE = 30.87, t = 2.82$) but no main effect of sentence type ($\beta = -17.16, SE = 28.44, t = -.60$), and no significant interaction ($\beta = -43.10, SE = 28.88, t = -1.49$) although there was again a trend for the early head position disadvantage to be greater in the NP/Z conditions than in the NP/S conditions.

Probability of a regression The model of PREG had to be simplified before it would converge. The least simple model that converged successfully contained no random slopes but only random intercepts. Effects of word length and word frequency were non-significant ($\beta = 0.09, SE = 0.09, z = 0.95, p = 0.34$ and $\beta = -0.06, SE = 0.07, z = -0.80, p = 0.42$ respectively). PREG yielded a disadvantage of early head position ($\beta = .40, SE = .09, z = 4.32, p < .001.$ There was no main effect of sentence type ($\beta = .03, SE = .09, z = .29, p = .77$). The head position x sentence type interaction effect was significant, with NP/Z tending to be harder than NP/S in the early head position cases, but NP/S harder than NP/Z in late head position ($\beta = -.27, SE = .09, z = -2.83, p < .05$).

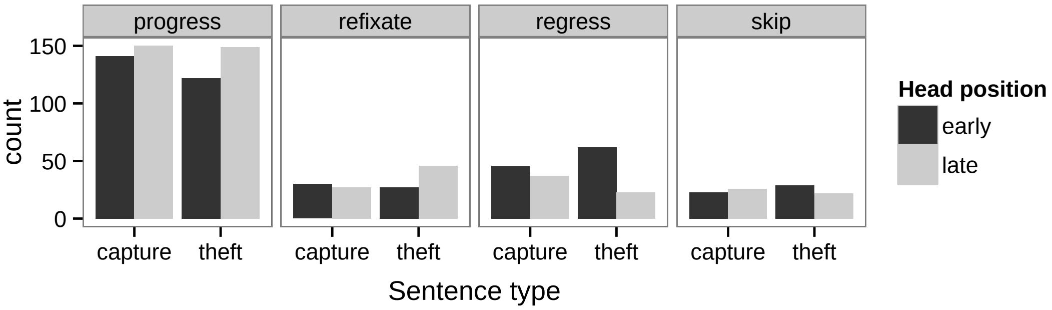


Figure 12.2: How coarse behaviours were distributed over conditions

Scan path analysis I divided scan paths into 4 types for a coarse-grained view of behaviours at disambiguation. These types were (1) skip; (2) progress; (3) refixate; (4) regress. Next I focused on the regress cases and asked whether within this class of behaviour there were subtypes that were over-distributed in the early NP/Z condition purported to be most difficult for the human parser.

Distribution of disambiguation behaviours A description of the distribution of behaviours at disambiguation follows. Progress 562, 58%; regress 168, 18%; refixate 130, 14%; skip 100, 10%. The way in which counts varied over conditions within this classification is plotted in Figure 12.2.

Distribution of regression strategies In this section I limit the scope to regressions from disambiguation and consider their shape. An initial scasim run was computed to check for outliers using a 2.5 sd threshold. The initial scasim run did not identify any outliers so the full data set numbering 168 regressions was submitted to multidimensional scaling and cluster analysis. Figure (12.3) shows the method for choosing a lower-dimensional representation of the scan path similarity space. A 4 dimensional map provided stress 11.0 and 12 clusters in a model described “VEI (diagonal, equal shape) model with 12 components”. These 12 components are represented in Figure 12.4 which plots the scan path that sits nearest the centroid of each cluster or component. Cluster E was indicative of long regressions back to early parts of the sentence.

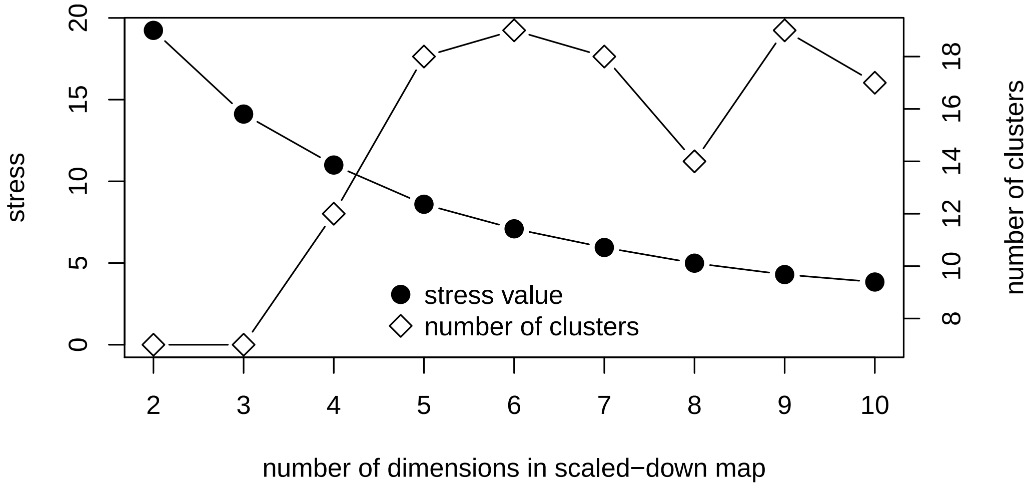


Figure 12.3: Plot illustrates how a lower dimensional representation of scan path similarity space was chosen

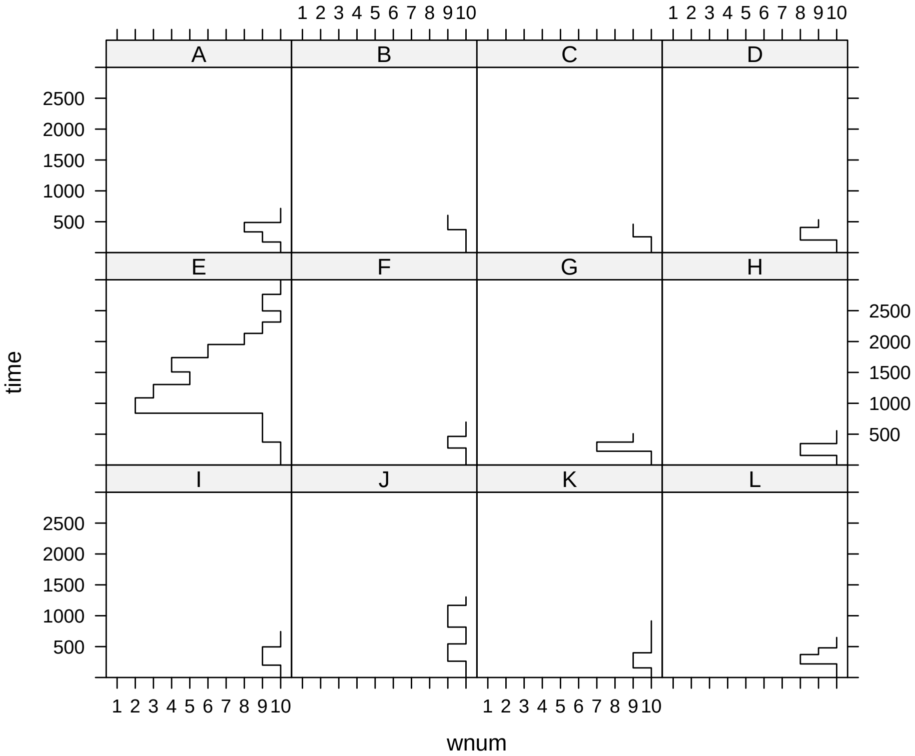


Figure 12.4: Plot shows the scan path that sits nearest the centroid of each cluster

Distribution of cluster E regressions In this section I focus on the distribution of Cluster E regressions over the treatment conditions. Diagnosis predicts that this pattern should be more common in NP/Z than in capture. This is because Diagnosis predicts that capture should be resolvable by consultation with the grammar. There should be no need for such movements seeking out terminals in capture. In the early head position cases Diagnosis predicts that this extra difficulty for NP/Z will be exacerbated versus the late head position cases.

Every trial was treated as either leading to membership of cluster E or not leading to it, as a binary dependent variable. A multilevel model of membership was constructed that treated head position and sentence type and their interaction as fixed effects, with corresponding random effects in a maximal model. This model failed to converge. A simpler model was constructed in response, that left out only the term representing the interaction effect over items. This model, only slightly simpler, converged. No effects were significant in the model, indicating that the strategy of interest was approximately equally distributed over the treatment conditions (head position effect, $\beta = 5.82, SE = 4.63, z = 1.26, p = 0.21$; sentence type effect, $(\beta = -5.89, SE = 4.65, z = -1.27, p = 0.21)$; head position x sentence type interaction, $(\beta - 3.56, SE = 4.07, z = -0.88, p = 0.38)$.

Model evaluation results In this section the predictions from the computational parsers are evaluated against the human behavioural results. Entropy reduction is sensitive to the NP/Z disadvantage but wrongly predicts that the late head position conditions were harder than the early head position conditions. Entropy reduction was very weakly negatively correlated with regression path duration $(r(858) = -.005, p = .87)$. Phrase structure surprisal is sensitive to the NP/Z disadvantage and the early head position disadvantage. However it predicts that the NP/Z disadvantage is greater in late head position sentences whereas the human data show that the NP/Z disadvantage diminishes or turns around into a NP/S disadvantage in late head position. Phrase structure surprisal was significantly positively correlated with regression path duration at disambiguation $(r(858) = .11, p = .002)$. Dependency surprisal is sensitive to the early head position disadvantage but predicts no effect of sentence type and no interaction effect. Dependency surprisal was significantly positively correlated with regression path duration at disambiguation $(r(858) = .13, p < .001)$. Dependency retrieval time wrongly predicts a disadvantage for late head position. It is not sensitive to the effect of sentence type and neither does it predict the interaction effect. There was a weak negative correlation with regression path duration at disambiguation ($r(858) = -.02, p = .49$).

12.5 Discussion

There were no significant effects on FFD or FPRT. The late temporal measures detected significant head position effects, with early head position causing longer RPD and TSR, making a difference of about 86 ms. The percent regressions measure indicated that participants were significantly more likely to make a regression in the early head conditions, and that this likelihood was significantly modulated by sentence type.

The shape of the interaction indicates that while there was only a small disadvantageous effect of early head position for the NP/S sentence type, there was a relatively large early head position cost for the theft sentence type. Although the 'late' human measures indicate trends towards longer regression times in theft than capture, these trends do not emerge significantly. The pattern of human measures can be interpreted in line with the Diagnosis model. Because the GDP licenses the repair operations in capture by means of a subcategorisation frame look-up, there is less need to revisit earlier material in capture than theft. In contrast, because there are no GDP links in the theft sentence type, repair operations for theft sentence types must be conducted at the terminal level. The eyes are employed to seek out a promising terminal.

Turning now to the scan path level analysis, the mere presence of a disambiguation strategy like E provides some support for the argument that repair is necessary. Participants did make these movements in response to the materials, and the fact that the distribution over the treatment conditions is approximately equal may indicate that this strategy was deployed on a subset of each of the sentence types. Such an account claims that some subset of participants resolves ambiguity by repair, but that participants are not more likely to deploy repair in any one of the treatment conditions.

Chapter 13

Conclusions

This chapter summarises the conclusions that can be drawn from the experimental work in the thesis, and presents some suggestions for future work.

13.1 A note on sparseness of data

Because the regressive eye movements of interest in this thesis happen quite rarely this can lead to sparse data when examining the distribution of those events over another variable (e.g., which word a regression saccade landed in). This sparseness can prevent lmer models converging. It can also lead to type II errors. It can also distort the calculations in scasim. It can also weaken claims about the representativeness of patterns observed over these variables. The experiments in this thesis used sample sizes that became standard when aggregated ANOVA was the main statistical framework - the results from the analyses in the thesis suggest that larger sample sizes should become standard for analysing rare events like regressions in reading now that modern methods like lmer that can gain better purchase with very rich data have supplanted ANOVA. However, analyses with small sample sizes like those used in the thesis can still reach statistical significance, and we should not disregard altogether such significant effects as we observe, even when they are computed over small sample sizes - rather we should take the sample size into account when interpreting the effects.

13.2 Regressions: spatial and linguistic properties

In the three experiments that manipulated a spatial factor (experiments one to three) there is evidence that regressive eye movements are initially driven by linguistic factors and then modulated by spatial factors. In these experiments, the manipulated linguistic factor was ambiguity and the manipulated spatial factor was the layout of the materials on screen. In experiments one and two the position of the disambiguation was manipulated. In experiment three the position of the ambiguously attached material was manipulated. Overall these experiments suggest that regressive eye movements are initially driven by linguistic factors and then modulated by spatial factors.

In experiment one, the early measure FPRT was sensitive to ambiguity (see 7.4 on page 129). In experiment two the early measure FPRT was sensitive to ambiguity (see 8.4 on page 142). In experiment three the early measure FPRT was sensitive to ambiguity (see 9.4 on page 155).

Layout affected the later measures in the experiments that manipulated the spatial position of the launch site, and tended to affect the later measures in the experiment that manipulated the spatial location of the ambiguously attached material. In experiment one, effects of layout were present in TSR, PREG, and scan path shape (see 7.4 on page 129; 7.4 on page 129; and 7.4 on page 133). In experiment two, effects of layout were present in PREG and scan path shape (see 8.4 on page 143 and 8.4 on 148). When the location of the misanalysed ambiguous material was manipulated, the early measure FPRT was sensitive to ambiguity (see 9.4 on page 155). There were trends towards effects of location of misanalyses material in the later measure scan path shape (see 9.4 on page 158).

13.3 Evidence for human repair parsing

Models of parsing that do not execute repair have only limited coverage of the human data. The extent of their coverage is that the model's dependent variable increases when human reading times increase and when regression frequency increases. The models lack coverage of variance in regressive scan path characteristics. Human difficulty produces responses of qualitatively different kinds, different reading time patterns on the one hand and different regression strategies on the other hand, but the models produce responses of only one kind - they can index the reading time effects but not the distribution of regression strategies.

Most accounts of sentence parsing assume that the reader accomplishes word recognition successfully as a precursor to the syntactic process of integrating a newly identified word into the partial syntactic structure of the sentence. Bicknell and Levy (2010a, 2010b, 2011); Levy et al. (2009) consider the consequences of an alternative assumption about word recognition. They start from the observation that perceptual input is inherently noisy, and note that this introduces doubt into the word recognition process, particularly in case where the word in question has orthographic *near neighbours*, e.g., *flour* which has the near-neighbour *floor*. Given the assumption of noisy perception, *floor* must sometimes be mis-read as *flour* and vice versa. Levy et al propose that readers maintain "probability distributions [. . .] extended over the *content* of the sentence as well as over its analysis" (Levy et al., 2009, p. 21087, my emphasis).

This contrasts with the general assumption that readers maintain a veridical representation of prior content. If readers do maintain probabilistic beliefs about the prior contents of a sentence, then when they encounter a word that is impossible to integrate with the existing partial representation of the sentence's meaning, this could cause a change in those beliefs: perhaps the the failure to integrate the new word is due to having mis-perceived an earlier word of the sentence, which would lead to an incorrect representation of the sentence's meaning so far. Levy et al suggest that such changes of belief about prior contents "would be likely to trigger regressive eye movements to earlier parts of a text and the locus of uncertainty." (p. 21087). This account offers an explanation for the fact of regression, and it also offers an explanation of where regressions are directed. However, the regressions indicated by this model would only target words with orthographic near neighbours. In its current form the model fails to offer an explanation for regressions that arise, not as a result of having mis-identified an earlier word, but as a result of having made an incorrect choice from among the various structural roles that a correctly-identified word can play in a representation of a sentence's structure. As a consequence, the model in its current form does not offer explanatory purchase on the cases examined in this thesis. However, if the model was extended so that words with alternative syntactic realisations had their syntactic realisation represented probabilistically rather than all-or-none, this would offer a good explanation of regressions that target words that have ambiguous syntactic assignments. This would explain regressions that target the onset of ambiguity in this thesis as regressions that target words with alternative syntactic realisations. However this does not obviate a repair mechanism, because once the parser has regressed to a word with alternative syntactic realisations, it must have a way to follow through the syntactic consequences of alternative realisations, and this latter process is equivalent with the notion of repair that is used in the thesis.

In order for the replacement models to offer an accurate account of human parsing while retaining their central commitments they would have to show that there is a threshold in the value of their response variable, such that when the threshold is exceeded this triggers a particular type of regression strategy. Currently the models can predict the fact of a regression but they cannot predict its shape. To the extent that regressions have shapes that are linked to the linguistic manipulations, the replacement parsers fall short of covering the data.

What is the best case scenario for replacement theories with respect to eye-tracking evidence? The nature of their coverage is that they predict the time that the parser takes to settle on a replacement parse. This settling time could be linked to regressions such that small settling times lead to short regressions that extend backwards only a short way, and large settling times lead to long regressions that extend backwards a long way. If this was sufficient to account for the observed evidence then we would be unable to reject the replacement theories. My claim is that accounts that link settling time to regression shapes using this simple index are not sufficient to account for the data. In what follows I take each experiment in turn and show that a simple index linking settling time to regression length (spatial length) is insufficient to account for the data from that experiment.

In experiment one, an account that links settling time to regression length would predict that there should be regressions of length one word, two words, three words and so on up to the maximum possible regression length of twelve words. However in that experiment the best clustering of the regression shape data indicates (see Figure 7.7 on page 132) that there are clusters of regressions of length 0 (regressions that comprise only refixations); two types of regressions back one word (one type that refixates the disambiguation and another type that does not); regressions back two words followed by a refixation on the disambiguation; regressions back six words; and regressions back eight words. There are no clusters going back three words; no clusters going back four words; no clusters going back five words; no clusters going back seven words; and no clusters going back for numbers of words between nine and twelve words. Although the absence of these clusters could be due to sparse data, the unevenness of the distribution of regressions over the possible targets suggests that there is no smooth function linking settling times to regression lengths. Instead it seems that there is a space near the disambiguation that attracts regressive fixations as a function of distance from disambiguation, and then particular words that have linguistic importance and that serve as attractors for regressive fixations over longer distances.

In experiment two these words that serve as attractors are the head of the misattached noun phrase (attracting the cluster of regressions that goes back six words in Figure 8.4 on page 146) and the verb to which this noun phrase is misattached (attracting the cluster of regressions that goes back eight words). There were too few regressions going back six words for analysis: however, there were enough regressions going back eight words for an analysis that shows that these regressions were unevenly distributed over the conditions (see 8.4 on page 148). For a replacement account to predict these accurately, it would have to propose that there are different indices linking settling time to regression length depending on how the materials are laid out on screen. Since no replacement account does propose such a thing, replacement accounts would have to be endowed with a representation of spatial layout of materials before they could start to offer a principled account of regressions that have different spatio-temporal properties as a function of text layout. Also replacement accounts do not explicitly have a notion of ambiguity, except to the extent that ambiguity earlier in the sentence affects settling time on the disambiguation. Repair accounts offer a more principled treatment of these phenomena. The way in which they account for regression shapes in this experiment is that they grant privileged status to linguistically relevant word-targets, and cast the process of regressing as a search for these word-targets. Repair accounts only have to be augmented with a notion that regressions to the line above the launch site are inhibited for oculomotor reasons before they can account for the uneven distribution (over conditions of text layout) of regressions that go back eight words. To account for the uneven distribution of these regressions over different conditions of ambiguity, repair accounts can appeal to linguistic relevance - words that take part in the structural ambiguity serve as more attractive targets of regressions that seek information that can help to resolve ambiguity by repair.

In experiment three scan path analysis did not reveal any statistically significant patterning of regression destination and linguistic manipulation (see the analyses in 9.4 starting on page 158).

In experiment four, there is more evidence for linguistically targeted regressions. Regressions that targeted the onset of ambiguity were significantly over represented in the ambiguous theft condition that readers found most difficult (see Figure 10.8 on 174). Replacement accounts struggle to account for this targeting - they could claim that increasing difficulty as measured by the settling metric leads to longer regressions, but they cannot account for the over-representation of linguistic targeting in a given condition.

In experiment five, long regressions were made unevenly over the conditions (see 11.4). These long regressions targeted different words in each condition but always went back further than the misattached noun phrase. Replacement accounts could cope with this if it were not also for the uneven distribution of these movements. There was a significant interaction between head position and sentence type in this experiment for scan path shape, showing that non-launch-adjacent regressions are distributed differently over the conditions. Replacement accounts struggle to account for this distribution.

In experiment six the long and complex regressions were approximately equally distributed over the conditions (see 12.4). This does not offer any additional reasons to prefer repair accounts.

Selectivity of regressions The evidence from this thesis and from other investigations (e.g., Inhoff & Weger, 2005; Weger & Inhoff, 2007) into the selectivity of regressions made in reading suggest that regressive eye movements are at least partially selective.

Purpose of regressions R. Booth and Weger (2012, Experiment 3) changed the target words in normal sentences after reading. They found that when the eyes later regressed to these words, “participants generally remained unaware of the change, and their answers to comprehension questions indicated that the new meaning of the changed word was what determined their sentence representations”. The authors claim that these results suggest that readers use regressions “to reread words and not to cue their memory for previously read words”.

Should replacement accounts be discarded in favour of repair accounts? The best-performing replacement account (a phrase structure parser generating an Entropy Reduction measure) is capable of indexing the human reading time and regression probability data for most of the experiments in the thesis. On this basis, there is an argument that since this parser does not implement repair, and yet manages to index some kinds of human data, it is not necessary to assume a repair mechanism in the human parser. However, the scan path analyses in the thesis show that there is another dimension to human behaviour at disambiguation, a spatio-temporal aspect to the human regressions data, and that none of the replacement parsers is endowed in principle with the ability to cover this kind of data. If the scan path analyses did not add anything to the findings from standard temporal and probabilistic measures, then the proper conclusion would be that the performance of the entropy reduction parser shows that it is not necessary to endow the human parser with a repair strategy. So the question whether replacement models of the human parser can be disregarded amounts to the question whether the results of the scan path analyses show anything additional to the standard analyses. The thesis provides evidence that regression spatio-temporal strategies are deployed differently according to the linguistic demands of the particular disambiguation required, in some situations. I argue that this amounts to a demonstration that full coverage of the human parser at disambiguation requires machinery for repair over and above the machinery used in normal parsing.

In normal first pass parsing, the problem can be characterised as one of assigning new words to an incrementally built representation of the sentence's structure. In repair parsing, the problem can be characterised as finding, in response to an error signal, a representation of the sentence's structure that does away with the error signal. These seem to me to be fundamentally different problems, both of which the human parser is good at solving. There seems to be good evidence that normal first pass parsing is well modelled by a replacement parser that operates over a sufficiently rich grammar formalism. There also seems to be good evidence that human behaviours at disambiguation include targeted, linguistically-guided sequences of regressive movements which a straightforward entropy reduction parser is under-equipped to model. The human parser may be a hybrid of the two strategies – on the one hand a set of probabilistic, exposure-sensitive first pass parsing routines that can maintain several candidate analyses ranked in memory and replace one with another in response only to the properties of the incoming word and the existing partial representation; and on the other hand a set of problem-solving strategies that can take the existing currently-preferred candidate analysis and carry out repair operations to yield a different representation that is both compatible with what has been read so far, consistent with the grammar, and also capable of offering a place for the new word to fit in.

Such a hybrid parser might seem to lack mathematical elegance when compared with the entropy reduction parser considered in the thesis, and is certainly less parsimonious. However, an appeal on the basis of parsimony should not be allowed to succeed if the most parsimonious model is unable to cover such a great part of the range of human behavioural responses to disambiguation, just as the replacement parsers are unable to account for the spatio-temporal aspect of the human regression behaviour in the thesis. On this basis the existing data from the thesis and from previous work showing patterned linguistically-guided search behaviour in regressions from disambiguation support a hybrid model of the human parser: one that carries out first-pass parsing in the style of the TD-PARSE parser, but is capable of repairing a currently-favoured analysis to create a new analysis that is not part of the existing set of top-ranked alternatives and that can support the integration of the new word. The justification for recommending a parser that does have machinery for repair is to have a model that achieves broad coverage of the human behavioural phenomena at the expense of parsimony of mechanism where these principles conflict.

What kind of repair does the best evidence support? Given that the evidence supports the inclusion of machinery for repair into a model of the human parser, and that there are several proposals for constraints on the repair machinery, it is natural to ask whether the data favour any particular repair proposal over the others.

The principal claim of the Diagnosis model (Fodor & Inoue, 1998) for the sentences examined in this thesis is that repairs are more difficult for NP/Z sentences than for NP/S sentences, and that head position should interact with this effect of sentence type. Experiment Four deals with the prediction for extra disambiguation difficulty for NP/Z sentences versus NP/S sentences; and Experiments Five and Six deal with the prediction for an interaction with head position.The data from Experiment Four show that disambiguation (and not just normal parsing) is harder for theft (NP/Z) than for capture (NP/S) and that the difficulty is manifest in RPD and marginally in PREG. Also scan path analysis showed that a regression strategy that can be characterised as *search for the onset of ambiguity* was over-distributed in the ambiguous theft cases. This shows that the disambiguation response in theft (but not capture) requires the onset of ambiguity to be re-examined, possibly to retrieve alternative subcategorisation frames for the ambiguous verb and its temporarily attached NP. In capture cases the parser appears to be able to carry out disambiguation without re-inspecting earlier material. The Diagnosis claim that repairing capture sentences can be done with the aid of the grammar under the GDP, but that theft sentences can only be resolved by re-inspecting the string of terminal words, is entirely compatible with the data from this experiment. In Experiment Five, with a short ambiguous phrase that either had the head at the beginning or at the end of the short phrase, Diagnosis predicted an interaction between head position and sentence type, with the capture cases both being soluble with reference to the grammar and the theft cases requiring a terminal search that must be longer for early head position than for late head position because of the relative numbers of intervening words. While head position affected FPRT, RPD, and marginally PREG as a main effect, head position did not affect sentence type in these measures. The standard measures do not support the Diagnosis prediction for the sentence type x head position interaction. The scan path analyses do not offer any support for the Diagnosis prediction either. In Experiment Six, with longer ambiguous phrases, the number of terminals that the parser must revisit in the early head position increases, and this seems to offer a better chance for the theft disadvantage to be modulated by head position, since in capture the length in terminals is irrelevant due to capture using the grammar to perform disambiguation. In this case early head position exerted a main effect on RPD that was not modulated by sentence type, and so does not support the Diagnosis prediction. Early head position exerted an effect as part of a head position x sentence type interaction in PREG in the direction predicted by the Diagnosis model. However this only speaks to how likely a regression was and the Diagnosis claim is cast in terms of the duration of a search through terminals, so this PREG interaction offers limited support for the Diagnosis claim.

The Decay model (Ferreira & Henderson, 1991b, 1998) claims only that early head position is harder then late head position because roles get assigned earlier in early head position and have therefore decayed more in early head position cases by the time disambiguation is reached. Experiments Five and Six manipulated head position. In Experiment Five the main effect of head position manifested in FPRT, RPD, and marginally in PREG - all of these effects can be taken as support for the Decay model. The fact that head position did not interact with sentence type in these measures does not diminish the support for Decay in the same way that it does for Diagnosis. In Experiment Six, the main effect of head position manifested in RPD, and PREG. An interaction was also found with sentence type in PREG. The main effects support the Decay model, and the interaction does not diminish the support for Decay. So the Decay model finds some support in the evidence from the thesis. However it is a model that makes only limited claims relative to the Diagnosis claims, so it is perhaps not surprising that the Decay claims are borne out by the evidence.

Deciding on a recommendation for the type of repair that the human parser carries out requires a comparison between a model that makes strong claims (Diagnosis) with moderate support and one that makes relatively weak claims (Decay) but with strong support. The head position main effects do not adjudicate between Diagnosis and Decay – both predict these main effects. The head position x sentence type interaction effects do not offer additional support for Diagnosis – they are predicted but not detected in the human data. The ambiguity x sentence type Experiment Four showed that when ambiguous NP/Z sentences are compared with ambiguous NP/S sentences and their unambiguous counterparts, the extra difficulty due to sentence type was thrown into sharp relief, and it was possible to see how regression shapes differed in this experiment according to sentence type. The evidence from this experiment tips the balance in favour of the Diagnosis model as a candidate for explaining how repair is carried out.

The Diagnosis model belongs in a space of models that distinguish between destructive and non-destructive repair in the explanation of differential repair difficulty. For the Diagnosis model, the case where destructive repair is mandated is labeled *theft*, and the case where non-destructive repair is possible is called *capture*. As such the Diagnosis model is a particular example of a class of monotonic repair model in the sense of Sturt (1998) and Sturt and Crocker (1996, 1997, 1998).

13.4 Importance of grammar formalism

Grammar formalism matters when it comes to modelling human reading data from the laboratory with temporarily syntactically ambiguous sentences, even if not for modelling eye movement corpus data for normal sentences. In the laboratory, parsers that operate over a phrase structure grammar outperform parsers that

Table 13.1: Summary table showing the coverage of the computational measures[a]

Parser Measure	E1	E2	E3	E4	E5	E6	overall
Dependency Surprisal	×	×	×	✓	×	(✓)	1.5
Dependency Retrieval	×	×	×	×	(✓)	×	0.5
Phrase Structure Surprisal	(✓)	(✓)	×	✓	✓	✓	4.0
Phrase Structure Entropy Reduction	(✓)	(✓)	✓	✓	(✓)	(✓)	4.0

[a] Key. ✓ represents good coverage of the human data; (✓) indicates partial coverage; × indicates poor coverage of the human data. Assigning one point for good coverage, half a point for partial coverage, and no point for bad coverage yields the overall score in the final column.

operate over a dependency grammar, when performance is measured as the extent to which the model's response variable predicts the shape of reading time effects (see Table 13.1). However, neither parsers operating over phrase structure grammars nor parsers operating over dependency grammars are equipped to predict particular regressive scan path strategies.

In experiment one, both dependency grammar measures were insensitive to effects of ambiguity and line position by the disambiguating word (sections 7.3 and 7.3). Both phrase structure grammar measures for the disambiguating word detected the effect of ambiguity that was present in RPD and PREG, and so are marked in the table as constituting partial coverage of the human data, but neither was sensitive to the layout manipulation (sections 7.3 and 7.3).

In experiment two, the eyetracking effect of ambiguity was present in PREG (section 8.4). Neither of the dependency parser measures was sensitive to this (sections 8.3 and 8.3), but both of the phrase structure parser's measures were sensitive to the effect of ambiguity (sections 8.3 and 8.3). The eyetracking data showed an effect of line position too in PREG, but none of the parser measures detected this. This is represented in the table as partial coverage for the phrase structure grammar measures.

In experiment three, the human data at the disambiguating word showed a significant main effect of ambiguity in RPD and PREG (sections 9.4 and 9.4). There was also a non-significant main effect of layout in the eyetracking data. The dependency parser made wrong predictions for these sentences, in both measures (sections 9.3 and 9.3). For the phrase structure parser the picture was mixed: the surprisal measure made the wrong predictions, but the entropy reduction measure made predictions in line with the eyetracking results (sections 9.3 and 9.3).

In experiment four, the eyetracking data (regression path duration) showed main effects of ambiguity and sentence type and their interaction (section 10.4). Dependency surprisal made the wrong predictions (section 10.3). Dependency retrieval made the wrong predictions (section 10.3). Both of the phrase structure grammar parser's predictions were in line with the human data (sections 10.3 and 10.3).

In experiment five the eyetracking data showed a significant disadvantage for theft in PREG (section 11.4) and a significant disadvantage for early head position in RPD (section 11.4). Dependency surprisal made the wrong predictions (section 11.3). Dependency retrieval predicted the head position effect (section 11.3). Phrase structure surprisal predicted both effects (section 11.3). Phrase structure entropy reduction predicted the theft disadvantage but also predicted the wrong direction for the head position effect (section 11.3).

In experiment six the human data showed an interaction between ambiguity and head position with a theft disadvantage in the early head position cases and a capture disadvantage in the late head position cases. Overall there was disadvantage for early head position (sections 12.4 and 12.4). The dependency surprisal measure had the right slope for the head position effect (section 12.3). Dependency retrieval predicted the wrong slope for the head position effect (section 12.3). Phrase structure surprisal was sensitive to the head position main effect as well as to the interaction with sentence type, with directions of effects lining up well with the human measures (section 12.3). Phrase structure entropy reduction had the wrong slope for the head position effect, but the right slope for the sentence type effect (section 12.3).

In experiment seven, across a series of comparisons with the human data, the phrase structure parser emerged better than the dependency parser, but neither managed to track the human data particularly well (section 12.5).

Parsers operating over phrase structure grammars have greater potential to predict particular regression strategies, although they lack this ability in current implementations. This is because dependency parsers are limited to probabilistically weighted dependency tuples for their prediction of difficulty in a response variable. These tuples do not distinguish between dependency tuples that have the same relation but vary in the difficulty exhibited by the human parser when one of the members of the dependency is involved in resolving syntactic ambiguity.

Because a phrase structure parser can appeal to already-built structure, it can in principle represent the difference between attaching a complement minimally or maximally, whereas dependency grammar only allows the representation of complement attachment itself. The limited representational capacity of dependency grammar suffices for coverage of eye movement corpora but not for experimental manipulations of the resolution of syntactic ambiguity. By distinguishing between minimal and maximal attachments of a given complement, a phrase structure grammar parser could in principle be adapted so that it could represent the complexity of attaching a word as well as the differential difficulty of attaching the word in different places in the hierarchy of the partially built phrase marker.

One caveat that applies to the conclusions is that the influence of the number of parses maintained in memory (the influence of the value of k in a k-best parser) was not explored. At higher values of k, the parsers might have been better equipped to cover the human data. However, allowing this value to vary as a free parameter has a disadvantage. It allows the computational parser to maintain in memory numbers of alternatives that seem implausible given constraints on human working memory. This carries the danger that models of parsing employ methods that are not available to the human parser in order to better fit the human data.

13.5 Suggestions for future work

Models of parsing need to output more than one dependent variable if they are to successfully cover the human data on syntactic ambiguity resolution. The existing dependent variables that they output only cover reading time and regression frequency at best. It would be possible for the models to be extended to output another response variable that described the type of difficulty, and the type of regression employed, for particular disambiguation problems.

Producing such a response variable would be more difficult for parsers operating over simple dependency grammar formalisms than for parsers operating over hierarchical phrase structure grammar formalisms. The parsers that would result from the requirement to cover both the amount and the nature of the difficulty induced by integrating a new word into a sentence would be better tools for investigating the details of the human sentence processing mechanism.

Currently most work is directed at parallel parsers that use probabilistic grammars to process input. These models now cover first pass parsing quite well. However none as currently implemented can hope to offer complete coverage of human parsing because of the regression behaviours that the human parser deploys and which they do not model. The most promising avenue for research in parsing seems to be to take the existing models that have good coverage of normal first pass passing, and to explore how they might be augmented with routines for carrying out repair. Adding repair to existing first pass routines would require an interface between the two to be well specified. How, and when, should control of the parser be ceded from first pass routines to repair routines and then back again? Should there be an executive part of the parser that handles the transitions of control between first pass and repair routines? Could repair functions be written into extended versions of the existing first pass routines? These would be promising questions for future research.

Appendix A

This appendix lists the sentences and questions for each experiment.

Experiment One materials The early line break conditions have the line break at \. The late line break conditions have the line break at \\. The ambiguous versions omitted the comma in parentheses, the disambiguated conditions included the comma in parentheses.

While the mob watched(,) the juggler who was gifted and nimble\ swallowed\\ a silver sword that was very sharp. While those men hunted(,) the moose that was sturdy and nimble\ hurried\\ into the woods and took cover. Though both lads phoned(,) the coach who was furious and bitter\ refused\\ to permit them to join the team. While the baby watched(,) her mother who was tired and fragile\ prepared\\ a new bottle of powdered milk. After the vet visited(,) the farmer who was shifty and evasive\ admitted\\ that some of his animals were ill. Though the dog sniffed(,) his trainer who was peeved and grumpy\ avoided\\ all further attempts to teach him tricks. While the fox stalked(,) the geese that were plump and healthy\ continued\\ to peck at grain on the ground. After the nun helped(,) the refugee who was sickly and afraid\ recovered\\ slowly in the camp near the river. While the maid dressed(,) the queen who was grouchy and aloof\ dismissed\\ all the other ladies in waiting. After the girl awoke(,) her father who was drunken and drowsy\ exploded\\ in anger about being disturbed so early. After the cadet saluted(,) the major who was brusque and remote\ ordered\\ the sergeant to punish the whole company. After the diva married(,) her agent who was dynamic and astute\ secured\\ her a lucrative contract with the theatre. While the team trained(,) the striker who was injured and unfit\ wondered\\ whether the damage would take long to heal. After the crowd heckled(,) the comic who was nervous and scared\ appeared\\ to cut his act short in humiliation. After the reps lobbied(,) the union that was divided and weary\ directed\\ its committee to approve the proposal. While the crew filmed(,) the actress who was fuming and cursing\ stormed\\ off the set of the film in a tantrum. After the boxer fought(,) the medic who was anxious and worried\ carried\\ a stretcher to the side of the ring. Though the horse kicked(,) the trainer who was quick and agile\ remained\\ calm and managed to avoid getting hurt. After the woman taught(,) the pupils who were bright and smart\ realised\\ that they could now solve the equations. After the boss ordered(,) the waiter who was ancient and doddery\ mumbled\\ the details to the chef incorrectly. While the woman bathed(,) her husband who was muddy and bruised\ announced\\ that he wanted to have a shower. After the army attacked(,) the rebels who were quiet and swift\ launched\\ a counter attack and inflicted huge losses. After the fire burned(,) the workman who was careful and dutiful\ laboured\\ to make sure the area was secure. While the temp assisted(,) the tycoon who was pompous and aloof\ committed\\ a series of white collar financial crimes.

Experiment Two materials Short line sentences are given first followed by long line sentences. The comma in parentheses is removed for the ambiguous versions.

While the mob watched(,) the juggler\ who was gifted and nimble swallowed\ a silver sword that was very sharp. While the mob watched(,) the juggler who was gifted\ and nimble swallowed a silver sword that was very\ sharp. While those men hunted(,) the moose\ that was sturdy and nimble hurried\ into the woods and took cover. While those men hunted(,) the moose that was sturdy\ and nimble hurried into the woods and took cover. Though both lads phoned(,) the coach\ who was furious and bitter refused\ to permit them to join the team. Though both lads phoned(,) the coach who was furious\ and bitter refused to permit them to join the team. While the baby watched(,) her mother\ who was tired and fragile prepared\ a new bottle of powdered milk. While the baby watched(,) her mother who was tired\ and fragile prepared a new bottle of powdered milk. After the vet visited(,) the farmer \ who was shifty and evasive admitted\ that some of his animals were ill. After the vet visited(,) the farmer who was shifty\ and evasive admitted that some of his animals were\ ill. Though the dog sniffed(,) his trainer\ who was peeved and grumpy avoided all\ further attempts to teach him tricks. Though the dog sniffed(,) his trainer who was peeved\ and grumpy avoided all further attempts to teach him\ tricks. While the fox stalked(,) the geese\ that were plump and healthy continued\ to peck at grain on the ground. While the fox stalked(,) the geese that were plump\ and healthy continued to peck at grain on the ground. After the nun helped(,) the refugee\ who was sickly and afraid recovered\ slowly in the camp near the river. After the nun helped(,) the refugee who was sickly\ and afraid recovered slowly in the camp near the\ river. While the maid dressed(,) the queen\ who was grouchy and aloof dismissed\ all the other ladies in waiting. While the maid dressed(,) the queen who was grouchy\ and aloof dismissed all the other ladies in waiting. After the girl awoke(,) her father\ who was drunken and drowsy exploded\ in anger about being disturbed so\ early. After the girl awoke(,) her father who was drunken\ and drowsy exploded in anger about being disturbed\ so early. After the cadet saluted(,) the major\ who was brusque and remote ordered\ the sergeant to punish the whole\ company. After the cadet saluted(,) the major who was brusque\ and remote ordered the sergeant to punish the whole\ company. After the diva married(,) her agent\ who was dynamic and astute secured\ her a lucrative contract with the\ theatre. After the diva married(,) her agent who was dynamic\ and astute secured her a lucrative contract with\ the theatre. While the team trained(,) the striker\ who was injured and unfit wondered \ whether the damage would take long\ to heal. While the team trained(,) the striker who was injured\ and unfit wondered whether the damage would take long\ to heal. After the crowd heckled(,) the comic\ who was nervous and scared appeared\ to cut his act short in humiliation. After the crowd heckled(,) the comic who was nervous\ and scared appeared to cut his act short in humiliation. After the reps lobbied(,) the union\ that was divided and weary directed\ its committee to approve the proposal. After the reps lobbied(,) the union that was divided\ and weary directed its committee to approve the proposal. While the crew filmed(,) the actress\ who was fuming and cursing stormed\ off the set of the film in a tantrum. While the crew filmed(,) the actress who was fuming\ and cursing stormed off the set of the film in a\ tantrum. After the boxer fought(,) the medic\ who was anxious and worried carried\ a stretcher to the side of the ring. After the boxer fought(,) the medic who was anxious\ and worried carried a stretcher to the side of the\ ring. Though the horse kicked(,) the trainer\ who was quick and agile remained calm\ and managed to avoid getting hurt. Though the horse kicked(,) the trainer who was quick\ and agile remained calm and managed to avoid getting\ hurt. After the woman taught(,) the pupils\ who were bright and smart realised\ that they could now solve the equations. After the woman taught(,) the pupils who were bright\ and smart realised that they could now solve the equations. After the boss ordered(,) the waiter\ who was ancient and doddery mumbled\ the details to the chef incorrectly. After the boss ordered(,) the waiter who was ancient\ and doddery mumbled the details to the chef incorrectly. While the woman bathed(,) her husband\ who was muddy and bruised announced\ that he wanted to have a shower. While the woman bathed(,) her husband who was muddy\ and bruised announced that he wanted to have a\ shower. After the army attacked(,) the rebels\ who were quiet and swift launched\ a counter attack and inflicted huge\ losses. After the army attacked(,) the rebels who were quiet\ and swift launched a counter attack and inflicted\ huge losses. After the fire burned(,) the workman\ who was careful and dutiful laboured\ to make sure the area was secure. After the fire burned(,) the workman who was careful\ and dutiful laboured to make sure the area was secure. While the temp assisted(,) the tycoon\ who was pompous and aloof committed\ a series of white collar financial\ crimes. While the temp assisted(,) the tycoon who was pompous\ and aloof committed a series of white collar financial\ crimes. **questions** Did the mother prepare food as well as milk? Were the geese plump? Was the girl's father happy to be woken up? Was the injured player a striker? Did the actress walk off quietly? Was the teacher a woman? Did the rebels attack first?

Experiment Three materials The early misanalysis area conditions are given first, followed by the late misanalysis area conditions. For the disambiguated versions, include the comma in parentheses: for the ambiguous conditions, omit the comma. The line break was always at \\.

While the mob watched(,) the juggler who was gifted and nimble swallowed\\

a silver sword that was very sharp. The busy guide noted that while the mob watched(,) the juggler swallowed\\ a silver sword that was very sharp. While those men hunted(,) the moose that was sturdy and nimble hurried\\ into the woods and took cover. One sole hiker spotted that while those men hunted(,) the moose hurried\\ into the woods and took cover. Though both lads phoned(,) the coach who was furious and bitter refused\\ to permit them to join the team. Their sad tutor moaned that though both lads phoned(,) the coach refused\\to permit them to join the team. While the baby watched(,) her mother who was tired and fragile prepared\\ a new bottle of powdered milk. The idle boy noticed that while the baby watched(,) her mother prepared\\ a new bottle of powdered milk. After the vet visited(,) the farmer who was shifty and evasive admitted\\ that some of his animals were ill. The local man stated that after the vet visited(,) the farmer admitted\\ that some of his animals were ill. Though the dog sniffed(,) his trainer who was peeved and grumpy avoided\\ all further attempts to teach him tricks. The alert judge mused that though the dog sniffed(,) his trainer avoided\\ all further attempts to teach him tricks. While the fox stalked(,) the geese that were plump and healthy continued\\ to peck at grain on the ground. The farm hand believed that while the fox stalked(,) the geese continued\\ to peck at grain on the ground. After the nun helped(,) the refugee who was sickly and afraid recovered\\ slowly in the camp near the river. The calm aide stressed that after the nun helped(,) the refugee recovered\\ slowly in the camp near the river. While the maid dressed(,) the queen who was grouchy and aloof dismissed\\ all the other ladies in waiting. The high lord implied that while the maid dressed(,) the queen dismissed\\ all the other ladies in waiting. After the girl awoke(,) her father who was drunken and drowsy exploded\\ in anger about being disturbed so early. The wary guest related that after the girl awoke(,) her father exploded\\ in anger about being disturbed so early. After the cadet saluted(,) the major who was brusque and remote ordered\\ the sergeant to prepare the ammunition. The new NCO recorded that after the cadet saluted(,) the major ordered\\ the sergeant to prepare the ammunition. After the diva married(,) her agent who was dynamic and astute secured\\ her a lucrative contract with the theatre. The nosy hack revealed that after the diva married(,) her agent secured\\ her a lucrative contract with the theatre. While the team trained(,) the striker who was injured and unfit wondered\\ whether the damage would take long to heal. The news show stated that while the team trained(,) the striker wondered\\ whether the damage would take long to heal. After the crowd heckled(,) the comic who was nervous and scared appeared\\ to cut his act short in humiliation. The daily rag claimed that after the crowd heckled(,) the comic appeared\\ to cut his act short in humiliation. After the reps lobbied(,) the union that was divided and weary directed\\ its committee to approve the proposal. The miners all gloated that after the reps lobbied(,) the union directed\\ its committee to approve the proposal. While the crew filmed(,) the actress who was fuming and cursing stormed\\ off the set of the film in a tantrum. The wise agent heard that while the crew filmed(,) the actress stormed\\ off the set of the film in a tantrum. After the boxer fought(,) the medic who was anxious and worried carried\\ a stretcher to the side of the ring. The keen fan lamented that after the boxer fought(,) the medic carried\\ a stretcher to the side of the ring. Though the horse kicked(,) the trainer who was quick and agile remained\\ calm and managed to avoid getting hurt. The old groom showed that though the horse kicked(,) the trainer remained\\ calm and managed to avoid getting hurt. After the woman taught(,) the pupils who were bright and smart realised\\ that they could now solve the equations. The new head observed that while the woman taught(,) the pupils realised\\ that they could now solve the equations. After the boss ordered(,) the waiter who was ancient and doddery mumbled\\ the details to the chef incorrectly. The grill cook remarked that after the boss ordered(,) the waiter mumbled\\ the details to the chef incorrectly. While the woman bathed(,) her husband who was muddy and bruised announced\\ that he wanted to have a shower. The hotel maid joked that while the woman bathed(,) her husband announced\\ that he wanted to have a shower. After the army attacked(,) the rebels who were quiet and swift launched\\ a counter attack and inflicted huge losses. The war diary argued that after the army attacked(,) the rebels launched\\ a counter attack and inflicted huge losses. After the fire burned(,) the workman who was careful and dutiful laboured\\ to make sure the area was secure. The male nurse spotted that after the fire burned(,) the workman laboured\\ to make sure the area was secure. While the temp assisted(,) the tycoon who was pompous and aloof committed\\ a series of white collar financial crimes. The desk clerk swore that while the temp assisted(,) the tycoon committed\\ a series of white collar financial crimes.

Experiment Four materials Capture sentences are given first followed by Theft sentences. For the ambiguous version, include the material in parentheses.

The cadet noticed (that) the captain walked to the gates of the enclosure. After the cadet saluted(,) the captain walked to the gates of the enclosure. The trustees regretted (that) the proposal attracted attention from the press. After the trustees debated(,) the proposal attracted attention in the press. The class believed (that) the model cooked a meal in the kitchen of the flat. After the class painted(,) the model cooked a meal in the kitchen of the flat. The thief reported (that) the guard called out to his colleague to bring tea. After the thief awoke(,) the guard called out to his colleague to bring tea. The nurse remembered (that) the patient asked for a bowl of fruit in her room. After the nurse visited(,) the patient asked for a bowl of fruit in her room. The manager announced (that) the player resigned from the football team. After the manager phoned(,) the player resigned from the football team. The lecturer recommended (that) the students tried to understand the theory. After the lecturer taught(,) the students tried to understand the theory. The youths observed (that) the lion died from the wounds they inflicted. After the youths hunted(,) the lion died from the wounds they inflicted. The crew knew (that) the actress retired to her room on the set. After the crew filmed(,) the actress retired to her room on the set. The horse saw (that) the jockey rushed out of the enclosure in a panic. After the horse kicked(,) the jockey rushed out of the enclosure in a panic. The crowd noted (that) the comedian decided to cut short his act. After the crowd heckled(,) the comedian decided to cut short his act. The jury decided (that) the matter took only a short time to resolve. After the jury settled(,) the matter took only a short time to resolve. The temp forgot (that) the tycoon bought the oil company from a competitor. After the temp assisted(,) the tycoon bought the oil company from a competitor. The pilot found (that) the trainee became an excellent officer. After the pilot helped(,) the trainee became an excellent officer. The critics understood (that) the film received an award in the ceremony. After the critics watched(,) the film received an award in the ceremony. The general advised (that) the rebels changed their plans quickly. After the general attacked(,) the rebels changed their plans quickly. The teacher denied (that) the pupils cheated in their final examinations. After the teacher trained(,) the pupils cheated in their final examinations. The chauffeur recognised (that) the car crashed into the wall by the gate. After the chauffeur parked(,) the car crashed into the wall by the gate. The maids accepted (that) the queen made a grand entrance into the room. After the maids dressed(,) the queen made a grand entrance into the room. The visitor felt (that) the dog moved very slowly into the small garden. After the visitor washed(,) the dog moved very slowly into the small garden. The mother promised (that) the kids behaved very well sometimes. After the mother bathed(,) the kids behaved very well sometimes. The princess admitted (that) the prince became rather quiet and withdrawn. After the princess married(,) the prince became rather quiet and withdrawn. The couple respected (that) the neighbours wanted to thank them. After the couple entertained(,) the neighbours wanted to thank them. The officials checked (that) the committee agreed to change its policy. After the officials lobbied(,) the committee agreed to change its policy. **questions** Did the captain drive to the gates? Did the tycoon sell the oil company? Did the patient ask for fruit? Did the comedian make his act longer? Did the driver win his race? Did the proposal attract attention from the press? Did the player resign from the team? Did the lion die from its wounds? Did the trainee become a useless officer?

Experiment Five materials Capture sentences are given first followed by Theft sentences. For early head position versions use the first argument in the brace, for the late head position versions use the second argument in the brace.

The cadet noticed {the captain of the squadron, the squadron captain} walked to the gates of the enclosure. After the cadet saluted {the captain of the squadron, the squadron captain} walked to the gates of the enclosure. The trustees regretted {the proposal of merger, the merger proposal} attracted attention in the press. After the trustees debated {the proposal of merger, the merger proposal} attracted attention in the press. The class believed {the model of swimwear, the swimwear model} cooked a meal in the basement of the flat. After the class painted {the model of swimwear, the swimwear model} cooked a meal in the basement. The thief reported {the guard of the prison, the prison guard} called out to his colleague to bring tea. After the thief awoke {the guard of the prison, the prison guard} called out to his colleague. The nurse remembered {the patient in distress, the distressed patient} asked for a bowl of fruit in her room. After the nurse visited {the patient in distress, the distressed patient} asked for a bowl of fruit in the room. The manager announced {the player of dignity, the dignified player} resigned from his place on the team. After the manager phoned {the player of dignity, the dignified player} resigned from his place on the team. The lecturer recommended {the students of politics, the politics students} tried to understand the theory. After the lecturer taught {the students of politics, the politics students} tried to understand the theory. The youths observed {the lion of beauty, the beautiful lion} died from the injuries it had sustained. After the youths hunted {the lion of beauty, the beautiful lion} died from the wounds they inflicted. The crew knew {the actress of elegance, the elegant actress} retired to her room on the film set. After the crew filmed {the actress of elegance, the elegant actress} retired to her room on the film set. The horse saw {the jockey in trouble, the troubled jockey} rushed out of the enclosure in a panic. After the horse kicked {the jockey in trouble, the troubled jockey} rushed out of the enclosure in a panic. The crowd noted {the comedian of renown, the renowned comedian} decided to cut short his act. After the crowd heckled {the comedian of renown, the renowned comedian} decided to cut short his act. The jury decided {the matter of culpability, the culpability matter} took only a short time to resolve. After the jury settled {the matter of culpability, the culpability matter} took only a short time to resolve. The temp forgot {the tycoon of retail, the retail tycoon} bought the company from a competitor. After the temp assisted {the tycoon of retail, the retail tycoon} bought the company from a competitor. The pilot found {the trainee with insight, the insightful trainee} became an excellent officer in the army. After the pilot helped {the trainee with insight, the insightful trainee} became an excellent officer in the army. The critics understood {the film of documentary, the documentary film} received an award in the ceremony. After the critics watched {the film of documentary, the documentary film} received an award in the ceremony. The general advised {the rebels of courage, the courageous rebels} changed their plans quickly. After the general attacked {the rebels of courage, the courageous rebels} changed their plans quickly. The teacher denied {the pupils of intelligence, the intelligent pupils} cheated in their final examinations. After the teacher trained {the pupils of intelligence, the intelligent pupils} cheated in their final examinations. The driver recognised {the car of aluminium, the aluminium car} crashed into the wall by the gate. After the driver parked {the car of aluminium, the aluminium car} crashed into the wall by the gate. The maids accepted {the queen of the colony, the colonial queen} made a grand entrance into the room. After the maids dressed {the queen of the colony, the colonial queen} made a grand entrance into the room. The visitor felt {the dog of the family, the family dog} moved very slowly into the small garden. After the visitor washed {the dog of the family, the family dog} moved very slowly into the small garden. The mother promised {the children with autism, the autistic children} behaved very well sometimes. After the mother bathed {the children with autism, the autistic children} behaved very well sometimes. The princess admitted {the prince of the region, the regional prince} became rather quiet and withdrawn. After the princess married {the prince of the region, the regional prince} became rather quiet and withdrawn. The couple respected {the family of importance, the important family} wanted to thank them formally. After the couple entertained {the family of importance, the important family} wanted to thank them formally. The officials checked {the committee for ethics, the ethics committee} agreed to change its policy. After the officials lobbied {the committee for ethics, the ethics committee} agreed to change its policy. **questions** Did the queen make a grand entrance? Did the tycoon buy the company? Did the prince become loud and outgoing? Did the driver win the race? Was the jockey completely calm? Did the captain walk to the gates? Were the students politics students? Did the rebels stick to the same plan?

Experiment Six materials Capture sentences are given first followed by Theft sentences. For early head position versions use the first argument in the brace, for the late head position versions use the second argument in the brace.

The cadet noticed {the captain who was smart, the tall and smart captain} walked to the gates of the enclosure. After the cadet saluted {the captain who was smart, the tall and smart captain} walked to the gates of the enclosure. The trustees regretted {the proposal that was harsh, the cold and harsh proposal} attracted attention in the press. After the trustees debated {the proposal that was harsh, the cold and harsh proposal} attracted attention in the press. The class believed {the model who was pretty, the young and pretty model} cooked a meal in the basement of the flat. After the class painted {the model who was pretty, the young and pretty model} cooked a meal in the basement of the flat. The thief reported {the guard who was tired, the old and tired guard} called out to his colleague to bring tea. After the thief awoke {the guard who was tired, the old and tired guard} called out to his colleague to bring tea. The nurse remembered {the patient who was hungry, the hungry and thirsty patient} asked for a bowl of fruit in her room. After the nurse visited {the patient who was hungry, the hungry and thirsty patient} asked for a bowl of fruit in her room. The manager announced {the player who was famous, the rich and famous player} resigned from his place on the team. After the manager phoned {the player who was famous, the rich and famous player} resigned from his place on the team. The lecturer recommended {the students who were clever, the keen and clever students} tried to understand the theory. After the lecturer taught the {the students who were clever, the keen and clever students} tried to understand the theory. The youths observed {the lion that was fierce, the fierce and proud lion} died from the wounds they inflicted. After the youths hunted {the lion that was fierce, the fierce and proud lion} died from the wounds they inflicted. The crew knew {the actress who was quiet, the quiet and serious actress} retired to her room on the set. After the crew filmed {the actress who was quiet, the quiet and serious actress} retired to her room on the set. The horse saw {the jockey who was scared, the scared and worried jockey} rushed out of the enclosure in a panic. After the horse kicked {the jockey who was scared, the scared and worried jockey} rushed out of the enclosure in a panic. The crowd noted {the comedian who was nervous, the nervous and timid comedian} decided to cut short his act. After the crowd heckled {the comedian who was nervous, the nervous and timid comedian} decided to cut short his act. The jury decided {the matter that was serious, the serious and weighty matter} took only a short time to resolve. After the jury settled {the matter that was serious, the serious and weighty matter} took only a short time to resolve. The temp forgot {the tycoon who was ruthless, the hard and ruthless tycoon} bought the oil company from a competitor. After the temp assisted {the tycoon who was ruthless, the hard and ruthless tycoon} bought the oil company from a competitor. The pilot found {the trainee who was careful, the careful and quiet trainee} became an excellent officer in the army. After the pilot helped {the trainee who was careful, the careful and quiet trainee} became an excellent officer in the army. The critics understood {the film that was boring, the long and boring film} received an award in the ceremony. After the critics watched {the film that was boring, the long and boring film} received an award in the ceremony. The general advised {the rebels who were strong, the tough and strong rebels} changed their plans quickly. After the general attacked {the rebels who were strong, the tough and strong rebels} changed their plans quickly. The teacher denied {the pupils who were bright, the bright and smart pupils} cheated in their final examinations. After the teacher trained {the pupils who were bright, the bright and smart pupils} cheated in their final examinations. The driver recognised {the car that was rusty, the black and rusty car} crashed into the wall by the gate. After the driver parked {the car that was rusty, the black and rusty car} crashed into the wall by the gate. The maids accepted {the queen who was noble, the quiet and noble queen} made a grand entrance into the room. After the maids dressed {the queen who was noble, the quiet and noble queen} made a grand entrance into the room. The visitor felt {the dog that was brown, the brown and white dog} moved very slowly into the small garden. After the visitor washed {the dog that was brown, the brown and white dog} moved very slowly into the small garden. The mother promised {the kids who were unruly, the wild and unruly kids} behaved very well sometimes. After the mother bathed {the kids who were unruly, the wild and unruly kids} behaved very well sometimes. The princess admitted {the prince who was wealthy, the young and wealthy prince} became rather quiet and withdrawn. After the princess married {the prince who was wealthy, the young and wealthy prince} became rather quiet and withdrawn. The couple respected {the family who were pleasant, the kind and pleasant family} wanted to thank them. After the couple entertained {the family who were pleasant, the kind and pleasant family} wanted to thank them. The officials checked {the committee that was weak, the weak and greedy committee} agreed to change its policy. After the officials lobbied {the committee that was weak, the weak and greedy committee} agreed to change its policy. **questions** Did the captain drive to the gates? Did the proposal attract attention from the press? Did the trainee become a useless officer? Did the player resign from the team? Did the driver win his race? Did the tycoon sell the oil company? Did the comedian make his act longer? Did the patient ask for fruit? Did the lion die from its wounds?

Experiment Seven materials The disambiguated versions of the materials included the comma in parentheses below; the ambiguous versions did not contain the comma.

After the cadet saluted(,) the captain the men drove to the port ate a hearty lunch. After the cadet saluted the captain(,) the men driven to the port ate a hearty lunch. After the trustees debated(,) the change the committee took to be useful was started that afternoon. After the trustees debated the change(,) the committee taken to be useful was started that afternoon. After the class painted(,) the model the artist drew for the museum was heard to disapprove. After the class painted the model(,) the artist drawn for the museum was heard to disapprove. After the thief awoke(,) the guard the official outran to the boat cried out in frustration. After the thief awoke the guard(,) the official outrun to the boat cried out in frustration. After the nurse visited(,) the thief the solicitor knew to be vindictive made an official complaint. After the nurse visited the thief(,) the solicitor known to be vindictive made an official complaint. After the fire burned(,) the workman the lawyer proved to have lied lost the court case. After the fire burned the workman(,) the lawyer proven to have lied lost the court case. After the manager phoned(,) the player the captain chose for the team gave up his place. After the manager phoned the player(,) the captain chosen for the team gave up his place. After the lecturer taught(,) the students the staff outdrank in the pub tried to explain the theory. After the lecturer taught the students(,) the staff outdrunk in the pub tried to explain the theory. After the youths hunted(,) the lion the leopard bit at the waterhole died from its injuries. After the youths hunted the lion(,) the leopard bitten at the waterhole died from its injuries. After the crew filmed(,) the actress the press forgot in the hallway waved to attract attention. After the crew filmed the actress(,) the press forgotten in the hallway waved to attract attention. After the horse kicked(,) the jockey the owner outrode in the derby retired the same year. After the horse kicked the jockey(,) the owner outridden in the derby retired the same year. After the crowd heckled(,) the magician the assistant sawed in two left the stage feeling unhappy. After the crowd heckled the magician(,) the assistant sawn in two left the stage feeling unhappy. After the temp assisted(,) the tycoon the executive outdid in the exercise bought a round of drinks. After the temp assisted the tycoon(,) the executive outdone in the exercise bought a round of drinks. After the stuntman helped(,) the trainee the pilot outflew in the display resolved to win next year. After the stuntman helped the trainee(,) the pilot outflown in the display resolved to win next year. After the audience watched(,) the film the studio gave the prize was ridiculed in all the reviews. After the audience watched the film(,) the studio given the prize was ridiculed in all the reviews. After the general attacked(,) the rebels the soldiers beat in the desert changed their strategy. After the general attacked the rebels(,) the soldiers beaten in the desert changed their strategy. After the teacher trained(,) the girls the boys showed up in the exam decided to revise even harder. After the teacher trained the girls(,) the boys shown up in the exam decided to revise even harder. After the chauffeur parked(,) the saloon the van overtook in the car park was taken to be cleaned. After the chauffeur parked the saloon(,) the van overtaken in the car park was taken to be cleaned. After the maids dressed(,) the queen the women hid in the kitchen made a grand entrance. After the maids dressed the queen(,) the women hidden in the kitchen made a grand entrance. After the visitor washed(,) the dog the boy spoke to by the sofa got up and ran round the garden. After the visitor washed the dog(,) the boy spoken to by the sofa got up and ran round the garden. After the mother bathed(,) the kids the guests awoke at daybreak complained that it was too early. After the mother bathed the kids(,) the guests awoken at daybreak complained that it was too early. After the princess married(,) the prince the duke forbade to attend felt very angry about it. After the princess married the prince(,) the duke forbidden to attend felt very angry about it. After the couple entertained(,) the neighbours the men saw earlier smiled as they walked past. After the couple entertained the neighbours(,) the men seen earlier smiled as they walked past. After the officials lobbied(,) the board the bank underwrote in the recession had to pay its debts. After the officials lobbied the board(,) the bank underwritten in the recession had to pay its debts. **questions** Was it the captain who made the phone call? Was it the captain who saluted? Was it the maids who hid in the kitchen? Did someone wave to attract attention? Was it the princess who got married? Did someone get beaten in the desert? Was it the tycoon who gave assistance? Did they run to a boat?

References

Adams, B., Clifton Jr, C., & Mitchell, D. (1998). Lexical guidance in sentence processing? *Psychonomic Bulletin & Review*, *5*(2), 265–270.
26

Anderson, J. (2005). Human symbol manipulation within an integrated cognitive architecture. *Cognitive science*, *29*(3), 313–341.
72

Anderson, J., & Lebiere, C. (1998). *The atomic components of thought*. Lawrence Erlbaum.
72

Angele, B., & Rayner, K. (2012). Processing the in the parafovea: Are articles skipped automatically? *Journal of Experimental Psychology: Learning, Memory, and Cognition*.
113

Baayen, R., Davidson, D., & Bates, D. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of memory and language*, *59*(4), 390–412.
109

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278.
108, 111, 112, 113, 128

Bates, D. M., Maechler, M., & Bolker, B. (2011). lme4: Linear mixed-effects models using S4 classes [Computer software manual]. (R package version 0.999375-39)
109

Bever, T. (1970). The Cognitive Basis for Linguistic Structures. *Cognition and the Development of Language*, *279*(362), 1–61.
47, 86

Bicknell, K., & Levy, R. (2010a). Rational eye movements in reading combining uncertainty about previous words with contextual probability. In *Proceedings of the 32nd annual conference of the cognitive science society.*
96, 203

Bicknell, K., & Levy, R. (2010b). A rational model of eye movement control in reading. In *Proceedings of the 48th annual meeting of the association for computational linguistics.*
96, 203

Bicknell, K., & Levy, R. (2011). Why readers regress to previous words: A statistical analysis. In *Proceedings of cogsci 2011.*
89, 96, 203

Binder, K., Duffy, S., & Rayner, K. (2001). The effects of thematic fit and discourse context on syntactic ambiguity resolution. *Journal of Memory and Language*, *44*(2), 297–324.
89

Booth, R., & Weger, U. (2012). The function of regressions in reading: Backward eye movements allow rereading. *Memory & Cognition*, 1-16.
206

Booth, T. L. (1969). Probabilistic representation of formal languages. In *Switching and automata theory, 1969., ieee conference record of 10th annual symposium on* (pp. 74–81).
35

Boston, M. (2012). *A computational model of cognitive constraints in syntactic locality* (Unpublished doctoral dissertation). Cornell University.
38, 43

Boston, M. (2013). *Humdep3.0. An incremental dependency parser developed for human sentence processing modeling.* http://conf.ling.cornell.edu/Marisa/#papers.
39, 83, 119, 123

Boston, M., & Hale, J. (2007). Garden-pathing in a statistical dependency parser. In *Proceedings of the midwest computational linguistics colloquium.*
41

Boston, M., Hale, J., Kliegl, R., Patil, U., & Vasishth, S. (2008). Parsing costs as predictors of reading difficulty: An evaluation using the Potsdam Sentence Corpus. *Journal of Eye Movement Research*, *2*, 1–12.
83, 85, 113

Boston, M., Hale, J. T., Vasishth, S., & Kliegl, R. (2011). Parallel processing

and sentence comprehension difficulty. *Language and Cognitive Processes*, *26*(3), 301-349. doi: 10.1080/01690965.2010.492228
84, 85

Brandt, S., & Stark, L. (1997). Spontaneous eye movements during visual imagery reflect the content of the visual scene. *Journal of Cognitive Neuroscience*, *9*(1), 27–38.
92, 93

Brysbaert, M., Drieghe, D., & Vitu, F. (2003). Word skipping: implications for eye movement control in reading. In G. Underwood (Ed.), *Cognitive processes in eye guidance.* Oxford University Press.
18

Buswell, G. T. (1922). *Fundamental reading habits: a study of their development.* Univ. of Chicago, Suppl. Educ. Monog.: Chicago.
19, 89

Campbell, F., & Wurtz, R. (1978). Saccadic amission: Why we do not see a grey-out during a saccadic eye movement. *Vison Research*, *18*, 1297–1303.
17

Cardona, G. (1998). *Pānini: A survey of research.* Motilal Banarsidass Publ.
84

Charniak, E. (2000). *BLLIP 1987-89 WSJ Corpus Release 1.* Linguistic Data Consortium, Philadelphia.
36

Chomsky, N. (1956). *The logical structure of linguistic theory.* University of Chicago Press.
22

Christianson, K., Hollingworth, A., Halliwell, J., & Ferreira, F. (2001). Thematic roles assigned along the garden path linger. *Cognitive Psychology*, *42*(4), 368–407.
26

Clark, H. (1973). The Language-as-Fixed-Effect Fallacy: A Critique of Language Statistics in Psychological Research. *Journal of Verbal Learning and Verbal Behavior*, *12*(4), 335–359.
106, 107, 108, 111

Clifton Jr, C. (1993). Thematic roles in sentence parsing. *Canadian Journal of Experimental Psychology*, *47*(2), 222–46.
26

Clifton Jr, C., Staub, A., & Rayner, K. (2007). Eye movements in reading words

and sentences. In *Eye movement research: A window on mind and brain* (pp. 341–372). Oxford: Elsevier.
19

Coco, M. (2011). *Coordination of Vision and Language in Cross-Modal Referential Processing* (Unpublished doctoral dissertation). School Of Informatics, Institute of Language, Cognition and Computation, University of Edinburgh.
108

Coleman, E. (1964). Generalizing to a language population. *Psychological Reports*, *16*, 219–226.
107

Cristino, F., Mathôt, S., Theeuwes, J., & Gilchrist, I. (2010). ScanMatch: A novel method for comparing fixation sequences. *Behavior research methods*, *42*(3), 692–700.
92, 93

Daniel, P., & Whitteridge, D. (1961). The representation of the visual field on the cerebral cortex in monkeys. *The Journal of Physiology*, *159*(2), 203.
93

Demberg, V., & Keller, F. (2008). Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. *Cognition*, *109*(2), 193–210.
84

DenBuurman, R., Boersma, T., & Gerrisen, J. (1981). Eye movements and the perceptual span in reading. *Reading Research Quarterly*, *16*, 227–235.
18

Durbin, R. (1998). *Biological sequence analysis: Probabilistic models of proteins and nucleic acids*. Cambridge, UK: Cambridge University Press.
91

Earley, J. (1970). An efficient context-free parsing algorithm. *Commun. ACM*, *13*(2), 94–102.
81

Engbert, R., Nuthmann, A., Richter, E., & Kliegl, R. (2005). SWIFT: a dynamical model of saccade generation during reading. *Psychological Review*, *112*(4), 777.
89

Engelmann, F., Vasishth, S., Engbert, R., & Kliegl, R. (2013). A framework for modeling the interaction of syntactic processing and eye movement control. *Topics in cognitive science*, *5*(3), 452–474.
89, 96

Ferreira, F., & Henderson, J. (1991a). How is verb information used during syntactic parsing? In G. Simpson (Ed.), *Understanding Word and Sentence* (pp. 303–330). Amsterdam: North-Holland: Elsevier Science Productions.
52, 53

Ferreira, F., & Henderson, J. (1991b). Recovery from misanalyses of garden-path sentences. *Journal of Memory and Language*, *30*(6), 725–745.
26, 51, 53, 54, 101, 180, 209

Ferreira, F., & Henderson, J. (1993). Reading processes during syntactic analysis and reanalysis. *Canadian Journal of Experimental Psychology*, *47*, 247–247.
101

Ferreira, F., & Henderson, J. (1998). Syntactic reanalysis, thematic processing, and sentence comprehension. In J. D. Fodor & F. Ferreira (Eds.), *Reanalysis in Sentence Processing.* The Netherlands: Kluwer Academic Publishers.
26, 51, 209

Ferreira, F., & Henderson, J. M. (1990). Use of verb information in syntactic parsing: evidence from eye movements and word-by-word self-paced reading. *J Exp Psychol Learn Mem Cogn*, *16*(4), 555-68.
42

Ferreira, F., & Patson, N. (2007). The 'good enough' approach to language comprehension. *Language and Linguistics Compass*, *1*(1-2), 71–83.
85

Ferretti, T., & McRae, K. (1999). Modeling the role of plausibility and verb-bias in the direct object/sentence complement ambiguity. In *Proceedings of the 21st annual conference of the cognitive science society* (pp. 161–66).
89

Fillmore, C. J. (1968). The case for case. In *Universals in linguistic theory* (pp. 1–88). New York: Holt, Rinehart & Winston.
23

Fisher, R. (1918). The correlation between relatives on the supposition of Mendelian inheritance. *Transactions of the Royal Society of Edinburgh*, *52*, 399–433.
107

Fisher, R. (1921). On the 'probable error' of a coefficient of correlation deduced from a small sample. *Metron*, *1*, 3–32.
107

Fisher, R. (1925). *Statistical methods for research workers*. Edinburgh.

107

Fodor, J., & Inoue, A. (1994). The Diagnosis and Cure of Garden Paths. *Journal of Psycholinguistic Research*, *23*, 407–434.
61, 101

Fodor, J., & Inoue, A. (1998). Attach anyway. In J. Fodor & F. Ferreira (Eds.), *Reanalysis in sentence processing* (pp. 101–141). Dordrecht, The Netherlands: Kluwer Academic Publishers.
29, 61, 62, 63, 66, 67, 101, 162, 208

Fodor, J., & Inoue, A. (2000). Garden path reanalysis: Attach (anyway) and revision as last resort. In M. DiVincenzi & V. Lombardo (Eds.), *Cross-linguistic perspectives on language processing* (pp. 21–61). Dordrecht, The Netherlands: Kluwer.
61

Forster, K. I., & Masson, M. E. (2008). Special Issue: Emerging Data Analysis. *Journal of Memory and Language*, *59*(4).
106, 108, 109, 111

Fraley, C., & Raftery, A. (2002). Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association*, *97*(458), 611–631.
95

Frazier, L. (1978). *On Comprehending Sentences: Syntactic Parsing Strategies* (Doctoral dissertation, University of Conneticut). available from IU Linguistics Club, 310 Lindley Hall, University of Indiana, Bloomington, Indiana.
49, 86

Frazier, L., & Clifton Jr, C. (1995). *Construal*. MIT Press.
86

Frazier, L., & Fodor, J. (1978). The sausage machine: A new two-stage parsing model. *Cognition*, *6*, 291–295.
48, 49

Frazier, L., & Rayner, K. (1982). Making and Correcting Errors During Sentence Comprehension: Eye Movements in the Analysis of Structurally Ambiguous Sentences. *Cognitive Psychology*, *14*(2), 178–210.
21, 26, 50, 51, 88, 89, 90, 96, 97, 102, 117, 118

Frazier, L., & Rayner, K. (1987). Resolution of syntactic category ambiguities: Eye movements in parsing lexically ambiguous sentences. *Journal of Memory and Language*, *26*(5), 505–526.
87, 121

Garnsey, S. M., Pearlmutter, N. J., Myers, E., & Lotocky, M. A. (1997). The contributions of verb bias and plausibility to the comprehension of temporarily ambiguous sentences. *Journal of Memory and Language*, *37*(1), 58–93.
54

Gelman, A., & Hill, J. (2007). Data Analysis Using Regression and Multilevel/Hierarchical Models. In R. M. Alvarez, N. L. Beck, & L. W. Lawrence (Eds.), *Data Analysis Using Regression and Multilevel/Hierarchical Models.* Cambridge, UK: Cambridge University Press.
113

Gibson, E. (1991). *A computational theory of human linguistic processing: Memory limitations and processing breakdown.* Carnegie Mellon University, Pittsburgh, PA. (Unpublished doctoral dissertation)
49

Gorrell, P. (1995). *Syntax and Parsing*. Cambridge, UK: Cambridge University Press.
56

Green, M., & Mitchell, D. (2006). Absence of Real Evidence against Competition during Syntactic Ambiguity Resolution. *Journal of Memory and Language*, *55*(1), 1–17.
70

Grill-Spector, K., & Malach, R. (2004). The Human Visual Cortex. *Annual Review of Neuroscience*, *27*(1), 649-677.
93

Grodner, D., Gibson, E., Argaman, V., & Babyonyshev, M. (2003). Against repair-based reanalysis in sentence comprehension. *Journal of psycholinguistic research*, *32*(2), 141–166.
86

Gruber, J. (1965). *Studies in lexical relations.* (Unpublished doctoral dissertation). Massachusetts Institute of Technology.
23

Hale, J. (2001). A Probabilistic Earley Parser as a Psycholinguistic Model. In *Proceedings Of The Second Meeting Of The North American Chapter Of The Association For Computational Linguistics* (pp. 1–8). Morristown, NJ, USA: Association for Computational Linguistics.
81

Hale, J. (2004). The information-processing difficulty of incremental parsing. In F. Keller, S. Clark, M. Crocker, & M. Steedman (Eds.), *Acl workshop incre-*

mental parsing: Bringing engineering and cognition together (pp. 58–65). Association for Computational Linguistics.
84

Hale, J. (2006). Uncertainty about the rest of the sentence. *Cognitive Science*, *30*(4), 643–672.
46, 84

Hamming, R. (1950). Error detecting and error correcting codes. *Bell System Technical Journal*, *29*(2), 147–160.
92

Hays, D. G. (1964). Dependency theory: A formalism and some observations. *Language*, *40*(4), 511–525.
84, 85

Holmes, V., Kennedy, A., & Murray, W. (1987). Syntactic structure and the garden path. *The Quarterly Journal of Experimental Psychology Section A: Human Experimental Psychology*, *39*(2), 2 – 277.
26, 27, 28

Howell, D. (2009). *Statistical methods for psychology*. Wadsworth Pub Co.
107

Inhoff, A., & Weger, U. (2005). Memory for word location during reading: Eye movements to previously read words are spatially selective but not precise. *Memory & Cognition*, *33*(3), 447–461.
100, 206

Jaeger, T. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language*, *59*(4), 434–446.
21, 108, 112

Johansson, R., & Nugues, P. (2007). Extended constituent-to-dependency conversion for english. In (pp. 105–112). University of Tartu, Estonia.
39

Josephson, S., & Holmes, M. (2002a). Attention to repeated images on the World-Wide Web: Another look at scanpath theory. *Behavior Research Methods, Instruments, & Computers*, *34*(4), 539.
92, 93

Josephson, S., & Holmes, M. (2002b). Visual attention to repeated internet images: testing the scanpath theory on the world wide web. In *Proceedings of the 2002 symposium on Eye tracking research & applications* (p. 49).
92, 93

Jurafsky, D. (1996). A probabilistic model of lexical and syntactic access and disambiguation. *Cognitive Science*, *20*(2), 137-194. (cited By (since 1996) 129)
69

Jurafsky, D., & Martin, J. (2009). *Speech and language processing* (Second edition ed.). Prentice Hall.
23

Just, M., & Carpenter, P. (1980). A theory of reading: From eye fixations to comprehension. *Psychological Review*, *87*(4), 329–354.
19, 97

Just, M., & Carpenter, P. (1992). A capacity theory of comprehension: Individual differences in working memory. *Psychological review*, *99*, 122–149.
89

Just, M., Carpenter, P., & Woolley, J. (1982). Paradigms and processes in reading comprehension. *Journal of Experimental Psychology: General*, *111*(2), 228–238.
28

Kennedy, A., & Murray, W. (1987). Spatial coding and reading: Comments on Monk (1985). *Quarterly Journal of Experimental Psychology*, 649–718.
100

Kimball, J. (1973). Seven principles of surface structure parsing in natural language. *Cognition*, *2*, 15–47.
42, 49

Kliegl, R., Grabner, E., Rolfs, M., & Engbert, R. (2004). Length, frequency, and predictability effects of words on eye movements in reading. *European Journal of Cognitive Psychology*, *16*(1-2), 262–284.
89

Kliegl, R., Nuthmann, A., & Engbert, R. (2006). Tracking the mind during reading: The influence of past, present, and future words on fixation durations. *Journal of Experimental Psychology: General*, *135*(1), 12.
84

Konieczny, L., & Döring, P. (2003). Anticipation of clause-final heads: Evidence from eye-tracking and srns. In *Proceedings of iccs/ascs.*
89

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2013). *lmertest: Tests for random and fixed effects for linear mixed effect models (lmer objects of lme4 package).* http://CRAN.R-project.org/package=lmerTest.

113

Legge, G. E., Klitz, T. S., & Tjan, B. S. (1997). Mr. chips: an ideal-observer model of reading. *Psychological review*, *104*(3), 524.
89

Levenshtein, V. (1966). Binary codes capable of correcting deletions, insertions and reversals. *Soviet physics doklady*, *10*, 707–710.
92

Levy, R. (2008). Expectation-Based Syntactic Comprehension. *Cognition*, *106*(3), 1126–1177.
81

Levy, R. (2013). *Memory and surprisal in human sentence comprehension.* (draft of book chapter)
82, 83

Levy, R., Bicknell, K., Slattery, T., & Rayner, K. (2009). Eye movement evidence that readers maintain and act on uncertainty about past linguistic input. In *PNAS* (Vol. 106, pp. 21086–21090). National Acad Sciences.
96, 203

Lewis, R. (1992). *Recent developments in the NL-Soar garden path theory* (Technical Report No. CMU-CS-92-141). Carnegie Mellon University: School of Computer Science.
78

Lewis, R. (1993). *An architecturally-based theory of human sentence processing* (Unpublished doctoral dissertation). Carnegie Mellon University, Pittsburgh, PA.
75, 77, 79, 81

Lewis, R. (1998). Reanalysis and limited repair parsing: Leaping off the garden path. In J. Fodor & F. Ferreira (Eds.), *Reanalysis in Sentence Processing* (pp. 247–285). Dordrecht, The Netherlands: Kluwer Academic Publishers.
99

Lewis, R., & Vasishth, S. (2005). An activation-based model of sentence processing as skilled memory retrieval. *Cognitive Science*, *29*, 1–45.
11, 71, 72, 73, 74

Lin, C.-J., Weng, R. C., & Keerthi, S. S. (2008). Trust region newton method for logistic regression. *The Journal of Machine Learning Research*, *9*, 627–650.
39

MacDonald, M. (1993). The interaction of lexical and syntactic ambiguity. *Journal*

of Memory and Language, *32*, 692–692.
87

Manning, C. D., & Schütze, H. (1999). *Foundations of statistical natural language processing* (Vol. 999). MIT Press.
35

Marcel, T. (1974). The effective visual field and the use of context in fast and slow readers of two ages. *British Journal of Psychology*, *65*, 479–492.
18

Marcus, G. (2013). Evolution, memory, and the nature of syntactic representation. In J. J. Bolhuis & M. Everaert (Eds.), *Birdsong, speech, and language.* MIT Press.
86

Marcus, M. (1978). *Theory of syntactic recognition for natural language* (Unpublished doctoral dissertation). MIT.
49

Marcus, M., Hindle, D., & Fleck, M. (1983). D-theory: Talking about talking about trees. In *Proceedings of the 21st annual meeting on association for computational linguistics* (pp. 129–136).
56

Matin, E. (1974). Saccadic suppression: A review and an analysis. *Psychological Bulletin*, 899–917.
17

McConkie, G., & Rayner, K. (1975). The span of the effective stimulus during a fixation in reading. *Perception and Psychophysics*, *17*, 578–586.
18

McElree, B., Foraker, S., & Dyer, L. (2003). Memory structures that subserve sentence comprehension. *Journal of Memory and Language*, *48*(1), 67–91.
72

McLachlan, G., & Peel, D. (2000). *Finite mixture models*. NY Wiley.
95

McRae, K., Spivey-Knowlton, M. J., & Tanenhaus, M. K. (1998). Modeling the influence of thematic fit (and other constraints) in on-line sentence comprehension. *Journal of Memory and Language*, *38*(3), 283–312.
70

Mel'čuk, I. (1988). *Dependency syntax: Theory and practice*. SUNY Press.
84

Meseguer, E., Carreiras, M., & Clifton Jr, C. (2002). Overt reanalysis strategies and eye movements during the reading of mild garden path sentencesClifton. *Memory & Cognition*, 551–561.
89, 102

Miellet, S., O'Donnell, P. J., & Sereno, S. C. (2009). Parafoveal magnification: visual acuity does not modulate the perceptual span in reading. *Psychol Sci*, *20*(6), 721–728.
18

Mitchell, D. (1987). Lexical guidance in human parsing: Locus and processing characteristics. In M. Coltheart (Ed.), *Attention and Performance XII: The Psychology of Reading* (pp. 601–618). Lawrence Erlbaum Associates, Inc.
26

Mitchell, D., Shen, X., Green, M., & Hodgson, T. (2008). Accounting for regressive eye-movements in models of sentence processing: A reappraisal of the Selective Reanalysis hypothesis. *Journal of Memory and Language*, *59*(3), 266–293.
90, 96, 97, 99, 100, 102, 103, 114, 115, 117, 118, 119, 127, 133, 134, 135, 136, 137, 138, 152, 153, 160

Murray, W., & Kennedy, A. (1988). Spatial coding in the processing of anaphor by good and poor readers: Evidence from eye movement analyses. *The Quarterly Journal of Experimental Psychology*, *40*(4), 693–718.
100

Narayanan, S., & Jurafsky, D. (1998). Bayesian models of human sentence processing. In *Proceedings of cogsci.*
69, 70

Narayanan, S., & Jurafsky, D. (2002). A bayesian model predicts human parse preference and reading time in sentence processing. In S. T.G. Dietterich & Z.Ghahramani (Eds.), *Advances in neural information processing systems 14* (pp. 59–65). MIT Press.
69

Needleman, S., & Wunsch, C. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of molecular biology*, *48*(3), 443–453.
93

Newell, A. (1990). *Unified theories of cognition*. Harvard University Press.
75

Nilsson, M., & Nivre, J. (2009). Learning where to look: Modeling eye movements

in reading. In *Proceedings of the thirteenth conference on computational natural language learning* (pp. 93–101).
89

Nilsson, M., & Nivre, J. (2010). Towards a data-driven model of eye movement control in reading. In *Proceedings of the 2010 workshop on cognitive modeling and computational linguistics* (pp. 63–71).
89

Nivre, J. (2004a). Incrementality in deterministic dependency parsing. In *Proceedings of the workshop on incremental parsing: Bringing engineering and cognition together* (pp. 50–57).
38

Nivre, J. (2004b). *Inductive dependency parsing* (Tech. Rep.). Växjö University: Växjö University.
37, 40, 83

Nivre, J. (2006). *Inductive Dependency Parsing*. Springer.
37

Nivre, J. (2009). Non-projective dependency parsing in expected linear time. In *Proceedings of the joint conference of the 47th annual meeting of the acl and the 4th international joint conference on natural language processing of the afnlp: Volume 1-volume 1* (pp. 351–359).
38

Patil, U., Vasishth, S., & Kliegl, R. (2009). Compound effect of probabilistic disambiguation and memory retrievals on sentence processing: Evidence from an eyetracking corpus. In *Proceedings of 9th International Conference on Cognitive Modeling.* Manchester.
72, 82, 123

Pickering, M., & Traxler, M. (1998). Plausibility and recovery from garden paths: An eye-tracking study. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *24*(4), 940–961.
26

Pinheiro, J., & Bates, D. (2009). *Mixed-effects models in S and S-PLUS*. Springer Verlag.
109

Pollatsek, A., Raney, G. E., Lagasse, L., & Rayner, K. (1993). The use of information below fixation in reading and in visual search. *Can J Exp Psychol*, *47*(2), 179–200.
18

Pritchett, B. (1988). Garden path phenomena and the grammatical basis of language processing. *Language*, *64*(3), 539–576.
29

Pritchett, B. (1992). *Grammatical competence and parsing performance*. Chicago, IL: University of Chicago Press.
29, 50, 80

Radach, R., & McConkie, G. (1998). Determinants of fixation positions in words during reading. *Eye guidance in reading and scene perception*, 77–100.
19

Radford, A. (1988). *Transformational grammar: A first course* (Vol. 1). Cambridge University Press.
22

Rayner, K. (1998). Eye Movements in Reading and Information Processing: 20 Years of Research. *Psychological Bulletin*, *124*, 372–422.
26, 128

Rayner, K. (2009). The 35th Sir Frederick Bartlett Lecture Eye movements and attention in reading, scene perception, and visual search. *The Quarterly Journal of Experimental Psychology*, *62*(8), 1457–1506.
128

Rayner, K., Ashby, J., Pollatsek, A., & Reichle, E. (2004). The Effects of Frequency and Predictability on Eye Fixations in Reading: Implications for the EZ Reader Model. *Journal of Experimental Psychology: Human Perception and Performance*, *30*(4), 720.
21, 89

Rayner, K., & Bertera, J. (1979). Reading without a fovea. *Science*, *206*, 468–469.
18

Rayner, K., Carlson, M., & Frazier, L. (1983). The interaction of syntax and semantics during sentence processing: Eye movements in the analysis of semantically biased sentences. *Journal of verbal learning and verbal behavior*, *22*(3), 358–374.
89

Rayner, K., Castelhano, M., & Yang, J. (2009). Eye movements and the perceptual span in older and younger readers. *Psychology and Ageing*, *24*(3), 755–760.
18

Rayner, K., & Frazier, L. (1987). Parsing temporarily ambiguous complements.

The Quarterly Journal of Experimental Psychology Section A: Human Experimental Psychology, *39*(4), 657 – 673.
26, 28

Rayner, K., Inhoff, A. W., Morrison, R. E., Slowiaczek, M. L., & Bertera, J. H. (1981). Masking of foveal and parafoveal vision during eye fixations in reading. *J Exp Psychol Hum Percept Perform*, *7*(1), 167–179.
18

Rayner, K., & McConkie, G. (1976). What guides a reader's eye movements? *Vision Research*, *16*(8), 829–837.
18

Rayner, K., Pollatsek, A., Ashby, J., & Clifton Jr, C. (2012). *Psychology of reading* (2nd ed.). Psychology Press.
17, 19, 89, 98, 99

Rayner, K., Slattery, T. J., Drieghe, D., & Liversedge, S. P. (2011). Eye movements and word skipping during reading: effects of word length and predictability. *J Exp Psychol Hum Percept Perform*, *37*(2), 514–528.
18

Reichle, E., Warren, T., & McConnell, K. (2009). Using EZ Reader to model the effects of higher level language processing on eye movements during reading. *Psychonomic Bulletin & Review*, *16*(1), 1–21.
89

Reilly, R. G., & Radach, R. (2006). Some empirical tests of an interactive activation model of eye movement control in reading. *Cognitive Systems Research*, *7*(1), 34–55.
89

Richter, T. (2006). What is wrong with ANOVA and multiple regression? Analyzing sentence reading times with hierarchical linear models. *Discourse Processes*, *41*(3), 221–250.
108

Roark, B. (2001). Probabilistic top-down parsing and language modeling. *Computational linguistics*, *27*(2), 249–276.
35

Roark, B. (2004). Robust garden path parsing. *Natural language engineering*, *10*(1), 1–24.
35, 36

Roark, B. (2013). *tdparse. An incremental top down parser.* `http://code.google.com/p/incremental-top-down-parser/`.

35, 83, 119, 123

Roark, B., Bachrach, A., Cardenas, C., & Pallier, C. (2009). Deriving Lexical and Syntactic Expectation-Based Measures for Psycholinguistic Modeling via Incremental Top-Down Parsing. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1* (pp. 324–333).
36

Schwarz, G. (1978). Estimating the dimension of a model. *The annals of statistics*, 461–464.
95

Shannon, C. (1948). A mathematical theory of communication. *Bell Systems Technical Journal*, *27*, 379—423.
45

Spivey, M., & Tanenhaus, M. (1998). Syntactic ambiguity resolution in discourse: Modeling the effects of referential context and lexical frequency. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *24*(6), 1521.
89

Staub, A. (2007a). The parser doesn't ignore intransitivity, after all. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *33*(3), 550.
26

Staub, A. (2007b). The return of the repressed: abandoned parses facilitate syntactic reanalysis. *Journal of Memory and Language*, *57*(2), 299–323.
26

Stigler, S. M. (1986). *The History of Statistics: The measurement of uncertainty before 1900*. Cambridge, Mass: Belknap Press of Harvard University Press.
106

Stolcke, A. (1995). An Efficient Probabilistic Context-Free Parsing Algorithm that Computes Prefix Probabilities. *Computational Linguistics*, *21*(2), 165–201.
43, 82

Sturt, P. (1998). *Syntactic Reanalysis in Human Language Processing* (Unpublished doctoral dissertation). University Of Edinburgh.
101, 210

Sturt, P., & Crocker, M. (1996). Monotonic syntactic processing: A cross-linguistic study of attachment and reanalysis. *Language and Cognitive Processes*, *11*(5), 449–494.
10, 55, 56, 58, 59, 61, 210

Sturt, P., & Crocker, M. (1997). Thematic monotonicity. *Journal of Psycholinguistic Research*, *26*(3), 297–322.
55, 210

Sturt, P., & Crocker, M. (1998). Generalized monotonicity for reanalysis models. In J. Fodor & F. Ferreira (Eds.), *Reanalysis in sentence processing* (pp. 365–400). Dordrecht: Kluwer Academic Publishers.
55, 101, 210

Sturt, P., Pickering, M., & Crocker, M. (1999). Structural change and reanalysis difficulty in language comprehension. *Journal of Memory and Language*, *40*, 136–150.
26, 29, 63, 101, 180

Tabor, W., & Hutchins, S. (2004). Evidence for self-organized sentence processing: digging-in effects. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*(2), 431.
180

Tabor, W., Juliano, C., & Tanenhaus, M. (1997). Parsing in a dynamical system: An attractor-based account of the interaction of lexical and structural constraints in sentence processing. *Language and Cognitive Processes*, *12*(2-3), 211–271.
69

Tanenhaus, M., Spivey-Knowlton, M., & Hanna, J. (2000). Modeling thematic and discourse context effects with a multiple constraints approach: Implications for the architecture of the language comprehension system. In *Architectures and mechanisms for language processing* (pp. 90–118). Cambridge University Press.
89

Taylor, S. (1965). Eye movements while reading: Facts and fallacies. *American Educational Research Journal*, *2*, 187–202.
18

Tesnière, L. (1959). *Eléments de syntaxe structurale*. Klincksieck Paris.
24, 33, 84

Traxler, M. (2005). Plausibility and verb subcategorization in temporarily ambiguous sentences: Evidence from self-paced reading. *Journal of psycholinguistic research*, *34*(1), 1–30.
26

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

89

Trueswell, J. C., Tanenhaus, M. K., & Kello, C. (1993). Verb-specific constraints in sentence processing: separating effects of lexical preference from garden-paths. *J Exp Psychol Learn Mem Cogn*, *19*(3), 528-53.
26, 89

Van Dyke, J., & Lewis, R. (2003). Distinguishing effects of structure and decay on attachment and repair: A cue-based parsing account of recovery from misanalyzed ambiguities. *Journal of Memory and Language*, *49*(3), 285–316.
101

van Gompel, R., Pickering, M., Pearson, J., & Liversedge, S. (2005). Evidence against Competition During Syntactic Ambiguity Resolution. *Journal of Memory and Language*, *52*(2), 284–307.
70

van Gompel, R., Pickering, M., & Traxler, M. (2001). Reanalysis in sentence processing: Evidence against current constraint-based and two-stage models. *Journal of Memory and Language*, *45*(2), 225–258.
26, 89

Vasishth, S., Brüssow, S., Lewis, R., & Drenhaus, H. (2008). Processing polarity: How the ungrammatical intrudes on the grammatical. *Cognitive Science*, *32*(4), 685–712.
72, 89

Vasishth, S., von der Malsburg, T., & Engelmann, F. (2013). What eye movements can tell us about sentence comprehension. *Wiley Interdisciplinary Reviews: Cognitive Science*, *4*(2), 125–134.
98, 117

Vitu, F. (1991). Against the existence of a range effect during reading. *Vision research*, *31*(11), 2009–2015.
19

Vitu, F., & McConkie, G. (2000). Regressive saccades and word perception in adult reading. *Reading as a perceptual process*, 301–326.
19, 89

von der Malsburg, T. (2009). Choice of saccade detection algorithm has a considerable impact on eye tracking measures. In *Proceedings of the European Conference on Eye Movements.* Southampton, UK.
94

von der Malsburg, T. (2010). *Scasim.* `http://www.ling.uni-potsdam.de/`

`~malsburg/scasim`.
91, 94, 103, 172

von der Malsburg, T., & Vasishth, S. (2007). A Time-Sensitive Similarity Measure for Scanpaths. In *Proceedings of the European Conference on Eye Movements.* Potsdam, Germany: Proceedings of the European Conference on Eye Movements.
94

von der Malsburg, T., & Vasishth, S. (2008). A New Method for Analysing Eye Movements in Reading that Is Sensitive to Spatial and Temporal Patterns in Sequences of Fixations. In *Proceedings of the CUNY sentence processing conference.* North Carolina.
94

von der Malsburg, T., & Vasishth, S. (2009). Readers use different strategies to recover from garden-paths: A Scanpath Analysis. In *Summer School on Embodied Language Games and Construction Grammar.* Cortona, Italy.
94

von der Malsburg, T., & Vasishth, S. (2011). What is the Scanpath Signature of Syntactic Reanalysis? *Journal of Memory and Language*, *65*(2), 109–127.
91, 94, 102, 103, 172

von der Malsburg, T., & Vasishth, S. (2012). Scanpath patterns in reading reveal syntactic under-specification and reanalysis strategies. *Language and Cognitive Processes*. (In press. Available on request from the first author.)
90, 91, 94, 103

Warner, J., & Glass, A. (1987). Context and distance-to-disambiguation effects in ambiguity resolution: Evidence from grammaticality judgments of garden path sentences. *Journal of Memory and Language*, *26*(6), 714–738.
26

Weger, U., & Inhoff, A. (2007). Long-range regressions to previously read words are guided by spatial and verbal memory. *Memory & Cognition*, *35*(6), 1293–1306.
100, 206

Weinberg, A. (1988). *Locality principles in syntax and in parsing* (Unpublished doctoral dissertation). Massachusetts Institute of Technology, Dept. of Linguistics and Philosophy.
86

Weinberg, A. (1993). Parameters in the theory of sentence processing: Minimal commitment theory goes east. *Journal of Psycholinguistic Research*, *22*(3),

339–364.
86

Wolverton, G., & Zola, D. (1983). The temporal characteristics of visual information extraction during reading. In K. Rayner (Ed.), *Eye movements in reading: Perceptual and language processes.* New York: Academic Press.
17

Wundt, W. (1900). Volkerpsychologie (vol. 1). *Leipzig: Engelmann*.
22

Yang, S. (2006). An oculomotor-based model of eye movements in reading: The competition/interaction model. *Cognitive Systems Research*, *7*(1), 56–69.
99, 102