A Statistical Model of Human Lexical Category Disambiguation

Steffan Corley

Ph.D.
The University of Edinburgh
1998

Contents

Declaration

I declare that this thesis has been composed by myself and that the research reported here has been conducted by myself unless otherwise indicated.

Ste'ffan Corley

Oxford, 13th January, 1998

Acknowledgements

My two supervisors, Matt Crocker and Chris Mellish, deserve primary thanks for both the fact and the shape of this thesis. Discussions with Matt lead to many of the ideas put forward here and his support for my work, generosity with his time and caring attitude have all made a large contribution to the finished product. Chris has been involved and helpful above and beyond the call of duty, and his insights and optimism helped change a fledgling project into a full thesis.

The experiments reported in chapter 8 would have been much poorer without the assistance of Chuck Clifton. His help, both in commenting on my experimental design and in the loan of software, proved invaluable in transforming an idea into a reality. Needless to say, any remaining errors in experimental design or procedure are my own.

Between the first and final submissions I was given the opportunity to make minor improvements to the clarity of presentation (and remove some typos). Thanks are due to my examiners, Suzanne Stevenson and Chris Brew, for pointing out the areas that needed most urgent attention.

The corpus searching software, reported in chapter 5, owes much to discussions following previous collaborative work with Martin Corley. He also provided invaluable assistance in analysing the experiments reported in chapter 8 and in loaning me otherwise unobtainable articles. Holly Branigan also deserves thanks for letting me use her papers collection as my own personal library.

All of the above also deserve thanks for the support they gave me outside work. To this list I would like to add Roger, Brigitte and Kevin Corley, for offering me a respite from my Ph.D., and numerous others, including (in alphabetical order): Lou and Fran Bell, Kim Binsted, Ed Carter, Jeremy Crowe, Alice Drewery, Chris Gathercole, Amanda Hargreaves, Amy Isard, Ali Knott, Hild Leslie, Merce Prat Sala, Patrick Sturt, Graeme Ritchie and Karen Verspoor.

Lastly and most importantly, I would like to thank my wife, Rebecca Bryant, for love and support, for encouraging me to get on with this thesis, and for pointing out that relaxation is at least as important as work.

Abstract

Research in Sentence Processing is concerned with discovering the mechanism by which linguistic utterances are mapped onto meaningful representations within the human mind. Models of the Human Sentence Processing Mechanism (HSPM) can be divided into those in which such mapping is performed by a number of limited modular processes and those in which there is a single interactive process. A further, and increasingly important, distinction is between models which rely on innate preferences to guide decision processes and those which make use of experience-based statistics.

In this context, the aims of the current thesis are two-fold:

- To argue that the correct architecture of the HSPM is both modular and statistical – the Modular Statistical Hypothesis (MSH).
- To propose and provide empirical support for a position in which human lexical category disambiguation occurs within a modular process, distinct from syntactic parsing and guided by a statistical decision process.

Arguments are given for why a modular statistical architecture should be preferred on both methodological and rational grounds. We then turn to the (often ignored) problem of lexical category disambiguation and propose the existence of a pre-syntactic Statistical Lexical Category Module (SLCM). A number of variants of the SLCM are introduced. By empirically investigating this particular architecture we also hope to provide support for the more general hypothesis – the MSH.

The SLCM has some interesting behavioural properties; the remainder of the thesis empirically investigates whether these behaviours are observable in human sentence processing. We first consider whether the results of existing studies might be attributable to SLCM behaviour. Such evaluation provides support for an HSPM architecture that includes this SLCM and allows us to determine which SLCM variant is empirically most plausible. Predictions are made, using this variant, to

determine SLCM behaviour in the face of novel utterances; these predictions are then tested using a self-paced reading paradigm. The results of this experimentation fully support the inclusion of the SLCM in a model of the HSPM and are not compatible with other existing models.

As the SLCM is a modular and statistical process, empirical evidence for the SLCM also directly supports an HSPM architecture which is modular and statistical. We therefore conclude that our results strongly support both the SLCM and the MSH. However, more work is needed, both to produce further evidence and to define the model further.

1: Introduction

1.1 Aims of Thesis

Research in Sentence Processing is concerned with discovering the mechanism by which linguistic utterances are mapped onto meaningful representations within the human mind. Such mapping can be divided into a number of intuitively distinct stages and a number of authors (in particular Fodor, 1983; Forster, 1979 and Frazier, 1979) have argued that these stages are reflected by a modular architecture of mind.[1] In contrast, other researchers argue that the Human Sentence Processing Mechanism (HSPM) does not consist of a number of compartmentalised decision making procedures, but is instead a unitary process with access to an unbounded set of representations and information sources (MacDonald, Pearlmutter & Seidenberg, 1994; Trueswell & Tanenhaus, 1994, and numerous others).

A second debate concerns whether the decision process(es) involved in the HSPM is/ are guided by statistical heuristics. Are decisions determined on the basis of learnt frequency information concerning the target language? Or do they depend on innate preferences, perhaps arising from architectural limitations of the HSPM? The latter position has been largely supported by proponents of a modular model of the HSPM, whereas most non-modular models of processing are statistical.

In this context, the aims of the current thesis are two-fold:

- To argue that the correct architecture of the HSPM is both modular and statistical – the Modular Statistical Hypothesis, introduced in chapter 3.
- To propose and provide empirical support for a position in which human lexical category disambiguation occurs within a modular process, distinct from syntactic parsing and guided by a statistical decision process.

Clearly, evidence supporting a modular, statistical, lexical category disambiguation

[1] We consider what is meant by 'modular' in chapter 3.

module also provides empirical justification for a modular, statistical HSPM. In this way the two aims of this thesis are interrelated – the former provides rational support[2] for the latter, and empirical substantiation of the latter also serves as evidence for the former.

1.2 Organisation of Thesis

In chapter 2, a number of models of human sentence processing are introduced. These serve to exemplify both the variety of empirical data that must be accounted for and the range of explanations that have been offered. In particular, we concentrate on the evidence concerning whether the HSPM is subject to statistical decision procedures, and on previously proposed modular statistical models. Chapter 3 explores the modularity debate in light of the architectures that have been recounted in chapter 2; in this chapter we argue that there are both methodological and rational reasons for preferring a modular model of human sentence processing.

Chapter 4 concerns lexical category ambiguity. We briefly introduce the problem before going on to consider how the models introduced in chapter 2 might be extended to make decisions in the face of such ambiguity. Only one previous model has been proposed in which lexical category disambiguation has a privileged status (Frazier & Rayner, 1987), and we consider this model in chapter 4. Finally, the Statistical Lexical Category Module (SLCM), our own model of lexical category disambiguation, is introduced and explained in this chapter.

Chapter 5 forms a bridge between theory and evidence. This chapter concerns the tools and information sources that were required to generate statistically motivated predictions for the SLCM, and how observable HSPM behaviour might reflect the decisions of the SLCM.

In chapters 6 to 8 we evaluate the SLCM model against empirical data. In chapter 6 we consider whether the initial decisions made by the SLCM tally with existing data concerning lexical category decisions and in chapter 7 we determine whether

[2] Throughout this thesis, the term 'rational' is used to refer to any argument which is essentially philosophical in nature; that is, one which makes no appeal to empirical data. This is not meant to suggest that such argument is necessarily rational, in the more common sense of the word, nor that argument based on empirical data is irrational.

reanalysis behaviour entailed by the SLCM model is evidenced in studies concerning HSPM reanalysis. In chapter 8, we present our own novel experiments. These concern an ambiguity for which SLCM-based predictions about HSPM behaviour differ from those which can be extrapolated from most of the other models we have recounted. Our experimental results concur with our SLCM-based predictions, and therefore provide support for the hypotheses that are central to this thesis.

Finally, chapter 9 sets out the conclusions that can be drawn from this thesis and highlights future work.

1.3 Some Terminology

Throughout this thesis the terms '*frequency-based*' and '*statistical*' are used interchangeably. However, the former more accurately describes the position we are putting forward. A statistical or probabilistic account of human sentence processing may rely on statistics that are not derived from frequency counts concerning an individual's linguistic experience. However, unless otherwise noted, when referring to a statistical model we mean one which is frequency-based.

The definition of the term '*lexical category*' is a recurring topic throughout this thesis – in particular, in chapters 4, 7 and 9. For the time being, we consider lexical category synonymous with part-of-speech. While the definition of this term is also theory-dependent, the majority of the current work concerns parts-of-speech that are acknowledged by most linguistic theories, such as 'noun' and 'verb'.

2: Sentence Processing

2.1 Introduction

Sentence processing research explores the way in which utterances are transformed into completely resolved semantic representations within the Human Sentence Processing Mechanism (HSPM). In this thesis, we concentrate on one part of this process – lexical category decisions. We take the view that these decisions are made independently; previously, many researchers have implicitly or explicitly assumed that such decisions are made as part of syntactic analysis. This chapter therefore reviews models of syntactic structure building and the evidence for and against them. In chapter 4 some of these models are evaluated as candidates for making lexical category decisions.

In section 2.2 we review the basic accepted facts about sentence processing and use these facts to define the space of possible models of the HSPM. In presenting the models, we divide this space in two ways; first between those in which the decision making process has access to only partial knowledge (sections 2.3 and 2.4) and those in which all information can be used to inform initial decisions (section 2.5) and second between those that use statistical or frequency-based knowledge about language to aid in the decision-making process (sections 2.4 and 2.5) and those that don't (section 2.3).

As our model of lexical category disambiguation uses a frequency-based heuristic to make initial decisions, it has most in common with the statistical models presented in section 2.4. The evidence for and against such a model is therefore considered in some detail in this section. The debate about whether a model should make initial decisions based on partial knowledge (a heuristic) is central to this thesis and is therefore dealt with on its own in chapter 3.

While this chapter reviews a cross-section of modern models of sentence processing, it is not intended to be comprehensive. Its purpose is to introduce the models and

exemplify the debates which will feature prominently in the remainder of this thesis.

2.2 The Basics

Language comprehension involves retrieving a largely unambiguous message from a highly ambiguous signal. In the psychological literature, this process has normally been split into four components: lexical access, syntactic analysis, semantic analysis and integrative processes.

Lexical access is the process whereby each lexical item in the input signal is recognized and its associated syntactic and semantic features recovered. Researchers have proposed both serial models, in which the lexicon is searched for a particular representation (Forster, 1976), and parallel models, in which all appropriate lexical entries are activated directly by the input (Morton, 1969; Seidenberg & McClelland, 1989). The experimental evidence largely supports the latter class of models. All entries are initially activated, even in the face of syntactic (Seidenberg *et al.*, 1982; Tanenhaus & Donnenworth-Nolan, 1984) and semantic (Swinney, 1979; Tanenhaus & Donnenworth-Nolan, 1984) biases; however, less preferred meanings are rapidly discarded (Seidenberg *et al.*, 1982, Rayner & Duffy, 1986).

Syntactic processing or parsing involves recovering the super-lexical structure inherent in the input signal. Again there are a number of possible models. However, the evidence in the case of syntactic analysis is far less conclusive. In sections 2.2.1 to 2.2.4 we outline the basic known facts about syntactic processing, and from this we draw conclusions about the possible architectures of the parser.

Semantic processing involves discovering the message – that is, recovering the meaning of the input signal. Comparatively little light has been shed on the mechanisms involved in this transformation, with the exception of its time course (see section 2.2.3). It is generally accepted that recovering the semantic content of an utterance requires both lexical and syntactic information about that utterance.

The final stage of language comprehension is integration. This involves making inferences about the utterance and its intention, and integrating non-linguistic cues into the comprehension process. As such integration is not restricted to language, it is not normally considered as a component of a dedicated HSPM.

In section 4.2 we observe that the evidence supports a model in which lexical category disambiguation occurs after lexical access. It is therefore the relation between syntactic processing and lexical category disambiguation that is central to this thesis. We continue by examining the accepted facts which underpin all modern models of the human syntactic processor.

2.2.1 The HSPM does build Syntactic Structures

Early sentence processing work concentrated on whether a complete grammatical description is formed during normal language comprehension. Alternative proposals included using general heuristics and facts about language to arrive at a pseudo-syntactic structure, and performing a complete syntactic analysis only if these alternative mechanisms failed to lead to successful comprehension (Bever, 1970, see section 2.3.1) or deriving sentence meaning directly from word meaning (Schank, 1972).

However, a number of researchers have demonstrated that such schemes either rely on hidden syntactic categories, or could not account for the full range of language understanding. For instance, Schank's proposal would make no distinction between the two sentences in 2.1 (see Ritchie, 1983, for further discussion):

(2.1) a. The boy saw the girl.
b. The boy was seen by the girl.

Researchers have also produced evidence for the psychological reality of syntactic constituents and structures. Experiments where naïve subjects were asked to indicate where they would put "breaks" into sentences have mainly shown that such subjects' intuitions agree with those of linguists (see Levelt, 1978). Just and Carpenter (1980) produced evidence that extra processing occurs at the end of clauses and sentences, and "click localisation" experiments (e.g. Fodor & Bever, 1965) have also produced evidence supporting the psychological reality of clauses.

Finally, and perhaps most conclusively, a number of experiments have found evidence that syntactic structures can be primed (see Branigan, 1995, chapter 7 and references therein). Put simply, subjects are more likely to assign a particular syntactic structure to a sentence if they have just assigned the same structure to a

previous sentence which is otherwise unrelated (in terms of semantic and lexical content), than if they have assigned a different structure to a previous sentence.

2.2.2 Parsing is not fully Parallel

If we accept that the HSPM does build syntactic structures, the next question must be are *all possible* structures created. For example, consider sentence 2.2, taken from the Brown corpus (Kucera and Francis, 1967).

(2.2) The President spent much of the week-end at his summer home on Cape Cod writing the first drafts of portions of the address with the help of White House aids in Washington with whom he talked by telephone.

Even if we only consider prepositional phrase attachment, there are numerous different syntactically permissible analyses of this sentence.[3] Are all of these available to the human sentence processor?

There is both intuitive and experimental evidence that suggests that parsing is not fully parallel. The intuitive evidence comes from 'conscious garden path' sentences, such as those in 2.3[4]:

(2.3) a. The horse *raced past the barn* fell.
b. The doctor told the patient *he was having trouble with* to leave.
c. After Susan drank *the water* evaporated.
d. Todd gave the boy *the dog* bit a bandage.
e. The old *train* the children.

In all cases the sentence has a grammatically correct reading, but informants reliably report that they experience conscious processing difficulty when trying to interpret these sentences. In some cases (particularly 2.3a), informants judge the sentence to be malformed. The correct (grammatical) reading is not available, suggesting it is either never computed, or discarded too early. This provides strong evidence that not all analyses are constructed.

Similarly, a large number of experiments have been conducted on temporarily

[3] The exact number depends on the specifics of the syntactic theory we espouse.

[4] Throughout this thesis, a part of a sentence which is lexically or structurally ambiguous will appear in italics.

ambiguous sentences, where the region following the ambiguity (the 'disambiguating region') is consistent with just one analysis. Examples are given in 2.4 and 2.5; in each case the disambiguating regions in the (a) and (b) forms favour different readings. Further examples can be found throughout this chapter.

(2.4) a. The cop told the motorist *that he had noticed* his car.
b. The cop told the motorist *that he had noticed* to drive slower.

(2.5) a. I know that the desert *trains* could resupply the camp.
b. I know that the desert *trains* soldiers to be tough.

The vast majority of experimental results have shown that disambiguations favouring one reading take less time to read than all others (see citations in Mitchell, 1994). This suggests that a single syntactic analysis is more available than any alternatives. However, it is worth noting that some experiments have found no reading time difference between disambiguations (Frazier & Rayner, 1987; Clifton, Frazier, Rapoport & Radó, submitted) – such evidence has been interpreted as supporting models which are either partially parallel (see section 2.5) or delay commitment in the face of some ambiguities (see section 2.3.2 and 2.2.4).

See Mitchell (1994) for a more detailed account of the evidence for and against parallel computation and its compatibility with various models.

2.2.3 Processing is Highly Incremental

So far, we have suggested that the HSPM does build syntactic representations, but it does not build all representations in parallel. However, it is also important to determine the timescale of sentence processing. On one end of the scale, it is possible that the processor waits until the end of the constituent, phrase or sentence before trying to create any sort of analysis. The alternative position is that syntactic and possibly other representations are constructed at the earliest possible point in processing ('incrementally'); this position involves the HSPM making far more decisions on the basis of incomplete evidence (see section 2.2.4).

The evidence suggests that processing is highly incremental. Marslen-Wilson (1975) and Marslen-Wilson and Tyler (1987) showed that grammatical errors can be detected and rectified extremely rapidly, apparently on a word-by-word basis. Some

semantic processing also appears to occur highly incrementally: Vonk (1984) found evidence that the antecedent of a pronoun is determined immediately and Tanenhaus, Spivey-Knowlton, Eberhard and Sedivy (1995) found that listeners very rapidly fixated on objects referred to in spoken text. However, Gernsbacher, Hargreaves and Beeman (1989) found that integrating individual clauses into a larger structure may be delayed until clause boundaries.

2.2.4 Early Decisions, Initial Decisions and Reanalysis

If parsing is not fully parallel, then decisions must be made about which syntactic structures to build or discard. Further, if processing is highly incremental, then these decisions must be made very early, often before all the relevant information is available. For example, consider the sentences in 2.6:

(2.6) a. John knew *the man in the hat* very well.
b. John knew *the man in the hat* was happy.

While the lexical realisation of these sentences is very similar, they have differing syntactic structures. These are shown in figure 2.1.

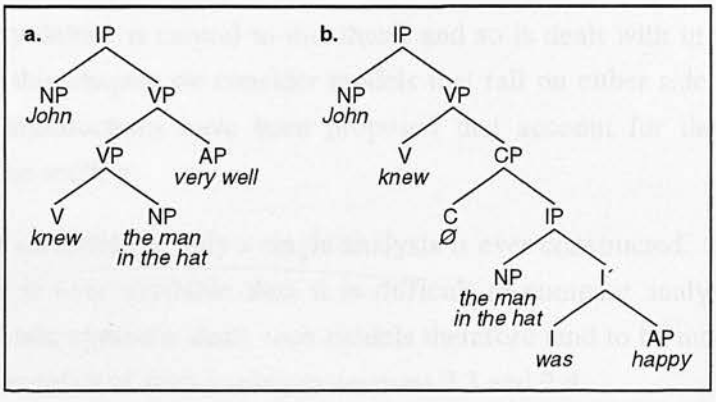

Figure 2.1: Syntactic analyses of sentences 2.6a and 2.6b

These two analyses diverge immediately after the word "knew".[5] However, they are lexically identical until after "hat". The evidence (reviewed in sections 2.2.2 and 2.2.3) suggests that just one of these analyses may be preferred before the disambiguation ("was happy" or "very well") is reached. The HSPM must therefore

[5] The extra VP node dominating "knew" in the analysis of 2.6a could be added later by Chomsky-adjunction.

be capable of making an '*early decision*' and accessing an alternative analysis if this decision proves wrong. However, it is also clear that it is not always possible to access a correct alternative analysis – hence conscious garden path sentences such as those in example 2.3.

One debate has concentrated on whether this early decision is made by a single decision process (an interactive view) or by a number of distinct processing 'modules' (a modular position). In the latter case, the proposal is that an '*initial decision*' is made by the parser, based on limited information, but later processes, making use of other relevant information, may force the parser to produce an alternative analysis ('*reanalysis*') at any time. In modular models the *early decision*, which is the output of the entire HSPM after all incremental processing has occurred for some prefix of a sentence, does not necessarily agree with the *initial decision* the parser makes when integrating the last word of this prefix. In an interactive model there is no clear distinction between *initial* and *early* decisions, and so the concept of an *initial decision*, as used here, is redundant; however, the state of an interactive model may evolve over time, resulting in one or more changes in preferred analysis associated with the processing of a single word.

The modularity debate is central to this thesis and so is dealt with in more detail in chapter 3. In this chapter we consider models that fall on either side of the debate. Three major architectures have been proposed that account for the known facts presented in this section:

- A serial architecture: only a single analysis is ever constructed. If only one analysis is ever available then it is difficult to compare analyses on the basis of non-syntactic cues; such models therefore tend to be modular. We detail a number of such models in sections 2.3 and 2.4.
- Weighted or ranked parallel architectures: all possible analyses are initially constructed, but one is preferred. In the weighted version, this is due to different 'activations' assigned to each analysis. In order to account for conscious garden paths, such a model must also be bounded: some less preferred analyses must be discarded. Such models tend to be interactive, though this is not a necessary consequence of the architecture (the modular

model we propose in chapter 4 could be viewed as a weighted parallel model). We consider such models in section 2.5.

- Minimal commitment or monotonic architectures: the parser often avoids making early decisions; the apparent early preference for a particular analysis is a result of the way the underspecified structures produced by the parser are interpreted by higher levels of processing. Examples can be found in Marcus *et al.* (1983), Sturt and Crocker (1996) and Weinberg (1993). The Delay Strategy (Frazier & Rayner, 1987), which we discuss in section 4.4 and throughout this thesis, could be considered as a (partial) minimal commitment model.

2.3 Non-Statistical Heuristics

If the HSPM only considers a single syntactic structure at a time (a serial model), then there must be some mechanism for choosing which analysis to pursue in the face of ambiguity. For example, in the case of the sentences in 2.6, the processor would have to make a decision immediately upon encountering "the"; either an NP node or a CP node could be created as a sister of the V node dominating "knew".

A number of authors have proposed different *heuristics* that might guide the initial decisions of a serial model. In this section, we first consider heuristics that use purely structural information, then turn to heuristics based on other grammatical content. Finally, we consider a heuristic model in which knowledge of the semantic context plays a primary role. In the next section we consider whether a statistical heuristic might be a better solution.

2.3.1 Structural Heuristics

Early sentence processing research concentrated on the architecture of the parser and short term memory constraints. Architectural limitations on parsing performance provided an explanation of the processing difficulty associated with some syntactic constructions. From these models, heuristics concerning the behaviour of the parser in a variety of situations could be deduced.

These parsing heuristics were formulated in terms of the ongoing tree structure created by the parser. There were two reasons for this. Firstly, architectural

limitations were assumed to arise out of lack of working memory required for structure building. Secondly, the parser was assumed to have access to information about syntactic structure, whereas it was (and is) less clear whether other types of information are available to it. This debate is dealt with in more detail in chapter 3.

More recently, the heuristic has overshadowed the architecture; while a heuristic may still be justified in terms of architectural constraints, research has focused on the heuristic and predictions arising from its use, rather than on the architecture itself.

Early Approaches

Miller and Isard (1963) introduced derivational syntactic theory (Chomsky, 1957) into psycholinguistics. They suggested that the difficulty associated with reading a sentence was related to the number of transformations that need to be applied to retrieve the deep structure analysis of the sentence from the surface structure; however, they put forward no theory of how the surface structure itself was recovered. Fodor, Garret and Bever (1968) believed that transformations were not used in computing deep structure; instead, deep structure relations were determined using cues in the surface structure together with lexical information.

Bever (1970) produced an extensive set of pragmatic heuristics for recovering relevant structural information from an input sentence. Examples include "take the first clause to be the main clause unless there is a subordinating conjunction" and "take a noun-verb-noun sequence as actor-action-object" (Rayner & Pollatsek, 1989). Bever's heuristics do not augment a parser so much as replace it; a complete syntactic analysis of the sentence is only necessary if the heuristics fail to produce an acceptable analysis. However, as discussed in section 2.2.1, the evidence suggests that a full syntactic representation is constructed for all sentences.

It was Kimball (1973) who first proposed a psycholinguistic theory in which surface structure parsing played a central role. This theory makes an assumption shared by many current models of human parsing – that it is the recovery of the syntactic surface structure of an utterance that accounts for much of the observed processing difficulty and so discovering the method by which this structure is recovered is crucial to understanding the HSPM. Kimball proposed six principles which guide surface structure parsing. Two of the principles characterise processing cost ('two

sentences' and 'fixed structure') while the remaining four are heuristics that guide the choices of the parser in the face of ambiguity ('top down', 'right association', 'new nodes' and 'closure'). These latter principles foreshadow the highly influential Garden Path model.

Kimball also suggested a seventh principle ('processing'). This final principle differs from the others in that it defines a two-stage model of the HSPM that embodies some of his other principles. Unfortunately, Kimball's model, and the principles embodied therein, have not stood up to the evidence (see Pritchett, 1992, for discussion). However, Frazier's (1979, 1987a) Garden Path theory could be seen as a simplification of Kimball's principles; it has proved more resilient.

The Garden Path Theory

Frazier's (1979, 1987a) Garden Path theory was also originally justified in terms of a mental architecture, called the "Sausage Machine" (Frazier & Fodor, 1978). The first stage (the Preliminary Phrase Packager or PPP) assigns lexical and phrasal nodes to groups of words, but has a fixed length window onto the sentence. It is therefore up to the second stage (the Sentence Structure Supervisor or SSS) to join these together by adding higher non-terminal nodes.

Frazier and Fodor showed that the Sausage Machine architecture embodies two parsing heuristics. The first they called 'Minimal Attachment' (MA), now often seen as an artefact of a race-bound parser. The definition of MA appears in a variety of wordings in Frazier's later work – one of the most succinct versions is:

Minimal Attachment:

Do not postulate any potentially unnecessary nodes.

(Frazier, 1987a, p.562)

Consider, for example, the sentences in 2.7 (from Frazier & Rayner, 1982).

(2.7) a. The city council argued *the mayor's position* forcefully.
b. The city council argued *the mayor's position* was incorrect.

These sentences have similar structure to those in 2.6. They both contain temporary structural ambiguity, while reading the italicised region of the sentence. The two candidate structures up to the end of the ambiguous region are depicted in figure 2.2.

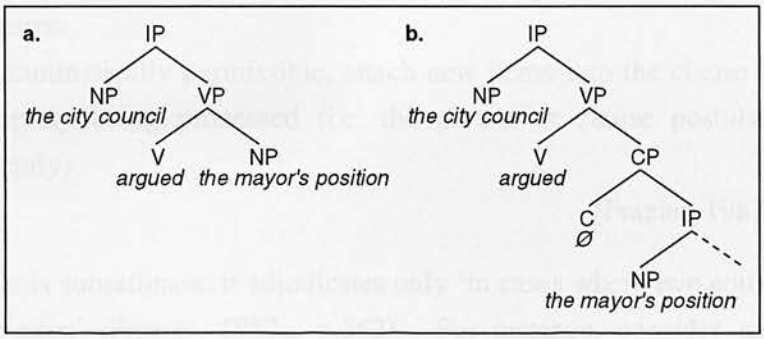

Figure 2.2: Minimal Attachment ambiguity

In these sentences, the parser must choose which structure to create as soon as it encounters the phrase "the mayor's position". To create analysis (b) the parser must create three more nodes than are required for analysis (a) – CP, IP and C. MA therefore predicts that analysis (a) would be preferred. This means that readers will experience more difficulty in the disambiguating region of sentence 2.7b, where the initial preference turns out to be incorrect, than in the corresponding region of 2.7a.

MA makes predictions for a wide range of ambiguous sentences. A few more examples are: complement/relative (2.8), main verb/reduced relative (2.9) and some prepositional phrase attachment (2.10) ambiguities[6]. In each case MA predicts that the (a) form would be easier to process than the (b) form.

(2.8) a. John told the girl *that Bill liked* her.
b. John told the girl *that Bill liked* to leave.

(2.9) a. The horse *raced past the barn* fast.
b. The horse *raced past the barn* fell.

(2.10) a. John hit the girl *with freckles.*
b. John hit the girl *with his hand.*

The second heuristic derived from the Sausage Machine architecture has come to be known as 'Late Closure' (LC). It is similar to Kimball's (1973) 'right association'. Again, different wordings of the definition of LC appear in a number of Frazier's works.

[6] Example 2.8 is adapted from Frazier (1987b); 2.9 and 2.10 are adapted from Frazier (1987a).

Late Closure:

> If grammatically permissible, attach new items into the clause or phrase currently being processed (i.e. the phrase or clause postulated most recently).
>
> (Frazier, 1987a, p.562)

This heuristic is subordinate; it adjudicates only 'in cases where two equally minimal attachments exist' (Frazier, 1987a, p.562). For instance, consider sentence 2.11 (adapted from Kimball, 1973); there are two possible analyses of this sentence, depicted in figure 2.3.

(2.11) Martha expected that it would rain *yesterday*.

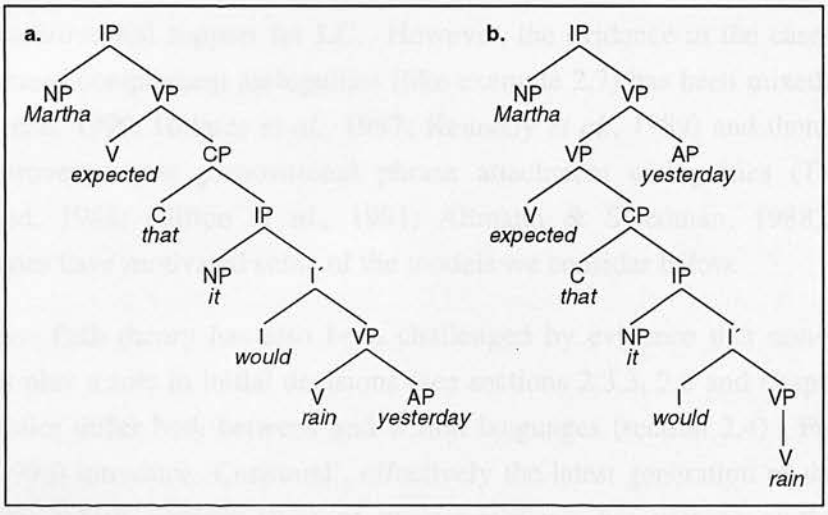

Figure 2.3: Late Closure ambiguity

MA does not prefer either analysis – they involve the creation of an equal number of new nodes. The ambiguity is therefore arbitrated by late closure. The most recently postulated phrase is the VP node dominating 'rain'; LC therefore predicts that the parser will prefer the analysis in which the AP node is integrated into this phrase (a) rather than into the main clause VP (b).

Further examples include subject/object (2.12) and argument prepositional phrase attachment (2.13) ambiguities[7]; in both cases, LC predicts that the (a) form would be

[7] 2.12 and 2.13 are both adapted from Frazier (1987a). However, the LC prediction in 2.13 does not agree with the current author's intuitions.

preferred.

(2.12) a. Since Jay always jogs *a mile* this seems like a short distance to him.
b. Since Jay always jogs *a mile* seems like a short distance to him.

(2.13) a. Jessie put the book Kathy was reading *in the library* down.
b. Jessie put the book Kathy was reading *in the library* collection box.

The Garden Path theory has enjoyed a large amount of support from experimental investigations. Complement/relative (Mitchell, Corley & Garnham, 1992) and main clause/reduced relative (Rayner, Carlson & Frazier, 1983; Ferreira & Clifton, 1986) ambiguities are among a number of types of sentence that have both been shown to follow MA predictions. Subject/object ambiguities (Frazier & Rayner, 1982; Ferreira & Henderson, 1991) are among those where the experimental evidence has provided fairly uncontroversial support for LC. However, the evidence in the case of direct object/reduced complement ambiguities (like example 2.7) has been mixed (Ferreira & Henderson, 1990; Holmes *et al.*, 1987; Kennedy *et al.*, 1989) and there has also been controversy over prepositional phrase attachment ambiguities (Taraban & McClelland, 1988; Clifton *et al.*, 1991; Altmann & Steedman, 1988). These controversies have motivated some of the models we consider below.

The Garden Path theory has also been challenged by evidence that non-structural influences play a role in initial decisions (see sections 2.3.3, 2.5 and chapter 3) and that heuristics differ both between and within languages (section 2.4). Frazier and Clifton (1996) introduce 'Construal', effectively the latest generation of the Garden Path theory, which answers some of these criticisms, but perhaps at the cost of explicit formulation and predictive power.

2.3.2 Grammatical Heuristics

In the Garden Path theory, knowledge of grammatical content (such as thematic roles) is only used in a checking stage that occurs after syntactic structure building (Rayner, Carlson & Frazier, 1983; Frazier, 1990). However, a number of other models have been proposed in which grammatical content plays an important role.

In this section we look at two of these models. Construal (Frazier & Clifton, 1996) distinguishes between arguments and modifiers, but then makes attachment decisions

for (at least) arguments without reference to grammatical content. Pritchett's (1992) model contrasts with Construal in that grammatical content is primary both in initial decisions and reanalysis constraints. The latter approach fits particularly well with Government-Binding (GB) theory (Chomsky, 1981); other GB-based models in which grammatical content plays an important role are proposed in Crocker (1991, 1996), Gibson (1991) and Merlo (1992).

Construal

Frazier and Clifton (1996) noted that the behaviour of the HSPM in the face of some syntactic ambiguities does not seem to fit with MA and LC predictions. These include PP and relative clause attachment ambiguities (see sections 2.3.1 and 2.4.1). They also found anomalies in adjunct predication; for example, consider 2.14 (from Frazier & Clifton, 1996):

(2.14) a. John ate the broccoli *naked.*
b. John ate the broccoli *raw.*

In this case, LC predicts that readers should initially prefer an object adjunct reading (b) rather than a subject adjunct (a). However, on-line experimentation (Clifton, Frazier, Rapoport & Radó, submitted) failed to confirm this prediction; they found no structurally motivated preference. Instead, semantic and aspectual factors appear to influence the complexity of processing and the initial preference in the face of ambiguity. This evidence runs counter to the predictions (and the spirit – see chapter 3) of the Garden Path theory.

What is notable about these exceptions to the Garden Path theory is that they involve modifier or 'non-primary relation' attachment. In contrast, experiments manipulating ambiguities just involving argument or 'primary relation' attachment have tended to support the Garden Path theory. Construal claims that these two types of ambiguity are determined by different mechanisms. Any phrase that can be attached so as to instantiate a primary relation is; MA and LC arbitrate any ambiguities. However, attachment of phrases that can only be analysed as instantiating a non-primary relation is delayed:

Construal Principle:

i. Associate a phrase XP that cannot be analyzed as instantiating a primary relation into the current thematic processing domain.

ii. Interpret XP within that domain using structural and non-structural principles.

(Frazier & Clifton, 1996, p.41)

The eventual attachment of non-primary phrases is constrained to be within the 'current thematic processing domain'. However, they are not attached by the parser. Instead, a later process which has access to semantic and other information resolves any remaining syntactic ambiguities.

Generalised Theta Attachment

Construal still assumes that the primary task of the parser is building syntactic structure. In contrast, Pritchett's (1992) model is based on GB theory; he assumes that "the core of syntactic parsing consists of the local application of global grammatical principles" (Pritchett, 1992, p.68). Pritchett only attempts to characterise conscious garden paths; the heuristic he employs to determine initial decisions is 'generalised theta attachment':

Generalised Theta Attachment:

Every principle of the Syntax attempts to be maximally satisfied at every point during processing.

(Pritchett, 1992, p.138)

The principles of syntax are those proposed by GB theorists; Pritchett makes specific use of $\overline{X}$-theory, the θ-criterion and Case theory (see Sells, 1985). Consider example 2.15, which is similar to the MA example 2.7:

(2.15) a. Bill knew *the woman* well.

b. Bill knew *the woman* hated him.

According to Pritchett, as soon as "knew" is read its maximal theta grid is recovered. In this case, the verb assigns experiencer role to an external argument ("Bill"), and theme role to an internal argument. When "the" is read, the parser must propose an NP; this can either be the direct object of "knew", or the subject of a complement

clause. In the former case, the NP dominating "the" is assigned theme role by the verb, and the θ-criterion is maximally satisfied (every NP has a role and every theta grid role is assigned). In the alternative analysis, the theme role is assigned to the complement clause; in this case, the NP dominating "the" does not (yet) have a role and so the θ-criterion is not maximally satisfied. According to generalised theta attachment, the first analysis should be preferred.

However, 2.15b is not a *conscious* garden path. It must therefore be possible for reanalysis to occur within the HSPM, Pritchett's 'on-line locality constraint' determines when unconscious reanalysis may occur:

On-Line Locality Constraint:

> The target position (if any) assumed by a constituent must be governed or dominated by its source position (if any), otherwise attachment is impossible for the automatic Human Sentence Processor.
>
> (Pritchett, 1992, p.101)

According to this constraint, reanalysis is possible in 2.15b. The NP "the woman" was originally licensed as the direct object "knew". After reanalysis, it is licensed as the subject of the embedded clause. However, the NP is still *governed* by "knew", and so the on-line locality constraint is not violated.

It is not clear how to evaluate Pritchett's model against the evidence. He does not claims to characterise unconscious garden path sentences, though we might assume that the model predicts human processing difficulty in cases where reanalysis must occur. In common with the majority of the other non-statistical heuristic models, his model may prove incapable of accounting for cross-linguistic and intra-linguistic differences (see section 2.4) and may have difficulty explaining evidence that the parser is influenced by non-syntactic factors (sections 2.3.3 and 2.5). It also seems that his model does not actually capture all the 'evidence' he presents; for example, consider 2.16, which Pritchett considers a conscious garden path:

(2.16) Bill warned *the woman* hated him.

While an earlier formulation of the on-line locality constraint (the 'theta reanalysis constraint' – see Pritchett, 1992, p.69) could explain why reanalysis is possible in

2.15 but not 2.16, the final version does not appear to differentiate between these examples and can therefore not explain the data that Pritchett himself presents.

In contrast to the Garden Path theory, Pritchett's model takes into account information specific to lexical items (the maximal theta grid) when making structural decisions. In the next section we consider a model in which both lexical information and semantic world knowledge play an important role in the parsing heuristic.

2.3.3 Cross-Modal Heuristics

The final non-statistical heuristic we consider is the Referential Support model (Crain & Steedman, 1985; Altmann & Steedman, 1988). The models discussed so far have attributed processing preferences to heuristics based on grammatical structure or content, arising from hypothetical limitations of the HSPM. Crain and Steedman's account is far more *functional*; decisions are made to help integrate novel material into the ongoing representation of the discourse. Syntactic decisions are therefore made on the basis of semantic preferences – the heuristic is cross-modal.[8]

Prior to Crain and Steedman (1985), most researchers had presented sentences to subjects without any prior semantic context. This is contrived; in the real world, when we interpret sentences we normally do so in context. Such context may bias our interpretation. Consider 2.18, presented out of context, and then following a complement-supporting context (2.17a) or a relative supporting context (2.17b) (from Altmann & Steedman, 1988):

(2.17) a. A psychologist was counselling a man and a woman. He was worried about one of them but not the other.

b. A psychologist was counselling two women. He was worried about one of them but not the other.

(2.18) a. The psychologist told the woman *that he was having trouble with* her husband.

[8] Although we characterised the models presented in this section as serial (see section 2.2.4), a purely semantic heuristic may not be compatible with a serial parser as "the use of context seems to only allow comparison of analyses" (Crain & Steedman, 1985, p.329). Crain and Steedman suggested that an independent parser generates all possible analyses when faced with ambiguity; a selection mechanism, following referential principles, rapidly selects one of these analyses, and only this analysis is further pursued. We treat this as a serial model because predictions derived from it are identical to those derived from a serial parser motivated by a semantic heuristic.

b. The psychologist told the woman *that he was having trouble with* to visit him again.

The two contexts differ only in the number of women that have been mentioned. Intuitively, it seems logical that when there are two women in the context (2.17b), we would expect the NP "the woman" to be modified by a relative clause in order to pick out an individual referent (2.18b), whereas when there is only one woman (2.17a) we would not expect modification (2.18a). Crain, Steedman and Altmann captured this intuition in their principles of "parsimony" and "referential support":

Principle of Parsimony:

A reading which carries fewer unsupported presuppositions will be favoured over one that carries more.

(Altmann & Steedman, 1988, p.203)

In the null context, the sentence complement interpretation (2.18a) requires that a single woman be postulated in the mental model; the relative clause reading (2.18b) requires a set of women of which one is giving the psychologist trouble. The former reading is therefore preferred; this prediction is the same as that of the Garden Path model. However, when the sentence is presented in context this principle interacts with referential support:

Principle of Referential Support:

An NP analysis which is referentially supported will be favoured over one that is not.

(Altmann & Steedman, 1988, p.201)

In a context in which there is only one woman (2.17a), the sentence complement analysis is referentially supported as "the woman" picks out a unique referent; the relative clause analysis requires the addition of more women to the discourse model and is therefore ruled out by parsimony. In contrast, in the two women context (2.17b), "the woman" does not describe a unique entity in the discourse model. In this case, the analysis is not referentially supported and so a relative clause analysis of the following material is preferred.

There is considerable evidence that discourse effects do influence the final output of

the HSPM (Altmann, 1988; Altmann & Steedman, 1988; Britt *et al.*, 1992; Mitchell *et al.*, 1992), but a number of researchers have suggested that such effects occur in a later semantic processing unit, rather than in the initial decisions of the syntactic parser (Brit *et al.*, 1992; Mitchell *et al.*, 1992; Clifton & Ferreira, 1989). In particular, Mitchell, Corley and Garnham (1992) produced evidence supporting the influence of discourse contexts on the ultimate interpretation, but showed that when the ambiguous region was extremely short, context appeared not to play a role; this evidence suggests that the initial decisions of the parser are made without reference to discourse information.

2.3.4 Summary

In this section we have considered a number of heuristics which might be used to make initial syntactic decisions within the sentence processor. There has been a gradual progression; the decision processes of the early models and the Garden Path theory used purely structural information, whereas those of Construal and Generalised Theta Attachment also made use of grammatical content. Finally, Referential Support required access to semantic and discourse knowledge. The question of what types of information should be available to the parser is discussed in chapter 3.

What all these models have in common is that the heuristic depends on information that is crucial to the interpretive process; structural knowledge, grammatical content and semantic and discourse knowledge must all be used at some point by the HSPM. In the next section we go on to consider statistical models; in these models, extra-linguistic data informs the processing heuristic.

2.4 Statistical Heuristics

Any general model of human sentence processing must apply equally to everyone. All the models outlined so far rely on the linguistic content of the utterance or discourse to inform parsing decisions. If therefore follows that all speakers of a given language should exhibit the same processing behaviour when interpreting the same discourse. Further, speakers of different languages should be using the same strategies to make choices in the face of ambiguity.[9]

[9] Though this does not necessarily mean that the decision will be the same – see section 2.4.1.

Recent work suggests that speakers of different languages, and even individual speakers of the same language, have different initial preferences in the face of the same linguistic input. This has led some researchers to propose that meta-linguistic knowledge plays a crucial role in the processing heuristic; the meta-linguistic knowledge most often proposed is experience-based statistical data about language use – in other words, frequency.

The model of lexical category disambiguation we will develop in chapter 4 is frequency-based. While this model does not attempt to characterise syntactic processing, it is important to understand why the HSPM may be (partially) statistical and how the proposed model fits in with existing statistical models of language processing. This section therefore explores the data and models in some detail.

We begin by examining the cross-linguistic and intra-linguistic evidence that has given rise to the statistical debate. We then go on to consider the small number of existing modular models which could be considered statistical, as well as issues of granularity. In the section 2.5 we explore non-heuristic (or non-modular) statistical models of language processing.

2.4.1 Cross-Linguistic Evidence

Consider example 2.19 (from Cuetos *et al.*, 1996):

(2.19) a. Someone shot the servant of the actress *who was on the balcony.*
b. Alguien disparó contra el criado de la actriz *que estaba en el balcón.*

2.19a is ambiguous between two readings – either the servant is on the balcony or the actress is on the balcony. The Garden Path theory predicts (by LC) that the relative clause will initially be attached low (modifying "actress") rather than high (modifying "servant"). Generalised Theta Attachment and Referential Support both fail to make predictions for this ambiguity (in the null context). The predictions of Construal depend on the particular 'structural and non-structural principles' used to determine non-primary phrase attachment.

However, the Spanish version of the example sentence (2.19b) has the same syntactic and thematic structure as the English version. The discourse context and order of reference is also the same. None of these models predict or suggest that, when faced

with such similar sentences, there should be a difference in the initial decision of the parser between Spanish and English.[10]

Experimental evidence, however, does suggest a difference. Spanish shows a clear high attachment bias in both questionnaire and on-line studies (Cuetos & Mitchell, 1988; Carreiras & Clifton, 1993). In English, the bias is less clear. Questionnaire evidence has supported both low (Cuetos & Mitchell, 1988) and high (Clifton, 1988) attachment. Some on-line experiments on similar constructions have shown an initial preference for low attachment (Clifton, 1988), while others have shown no bias in either direction (Carreiras & Clifton, 1993; unpublished studies by Mitchell and colleagues, cited in Cuetos *et al.*, 1996).

A number of questionnaire and on-line studies in other languages have mainly produced evidence in favour of high attachment (see Cuetos *et al.*, 1996; Corley, 1995). However, there is somewhat controversial evidence for an on-line low attachment preference in Italian (de Vincenzi & Job, 1995), despite data from questions asked after the subjects read the on-line sentences pointing to a high attachment bias (de Vincenzi & Job, 1993). Japanese data (Kamide & Mitchell, 1996, see also Branigan, Sturt & Matsumoto Sturt, 1996) also supports a low attachment bias on-line, despite a high attachment bias in questionnaire studies; however, the left-branching structure of Japanese means that non-statistical models may also predict different preferences to those exhibited in right-branching languages.

It does seem clear that for this construction there is a qualitative difference between the preferences of English and Italian speakers and those of speakers of the other languages tested. This evidence suggests that the parsing heuristic might take into account something other than the linguistic input. In the next section, we consider more evidence supporting this view, this time showing individual differences between speakers of the same language.

2.4.2 Intra-Linguistic Evidence

Evidence of cross-linguistic differences could be explained by a number of theories,

[10] Except, perhaps, Construal – see Gilboy, Sopena, Clifton and Frazier (1995) for an account of how Construal can explain such cross-linguistic differences.

some of which we discuss in the section 2.4.3. This evidence is compatible with, and may be predicted by, a frequency-based account of syntactic processing. It is not predicted by, and may not be compatible with, most of the models presented in section 2.3 (the exception being Construal). However, such evidence does not entail that the HSPM must use statistical mechanisms.

However, evidence for stable differences between speakers of the same language, in combination with the cross-linguistic data, would be far more compelling. Such evidence is beginning to emerge.

Corley (1995) performed a questionnaire study using sentences such as those in 2.20 and 2.21 (see also Corley & Corley, 1995):

(2.20) a. The satirist ridiculed the lawyer of the firm...
b. The satirist ridiculed the firm of the lawyer...

(2.21) a. The artist painted the niece of the patrons...
b. The artist painted the nieces of the patron...

Subjects were presented two questionnaires, three weeks apart. The questionnaires were similar except that where one contained an (a) form from 2.20 or 2.21, the other had the corresponding (b) form. In each case, the subjects were instructed to complete each sentence, beginning the completion with "who" or "which", followed by "was" or "were". Corley found a 63% preference for low attachment, mirroring Cuetos and Mitchell's (1988) figure; however, he also discovered a wide degree of individual variation, which was highly correlated by subject across the two questionnaires.

The results of an on-line experiment (Corley, 1995) are less clear. Two groups of subjects, who had shown a clear high or low attachment bias in the questionnaire study, were presented with novel materials similar to those used in the questionnaire. These materials had an added disambiguating region either immediately after the onset of the ambiguity (2.22) or somewhat later (2.23).

(2.22) a. The judge sentenced the killer of the people *who* was involved in the riot.
b. The judge sentenced the killer of the people *who* were involved in the

riot.

(2.23) a. The traveller visited the wives of the Sultan *who lived in a magnificent palace and* were greatly feared throughout the country.

b. The traveller visited the wives of the Sultan *who lived in a magnificent palace and* was greatly feared throughout the country.

For the early disambiguation materials, Corley found a reliable difference between the high and low attachment groups' reading time in the disambiguating region by subjects but not (quite) by materials. For the late disambiguation materials, this difference was not apparent.

Corley and Caldwell (1996) studied the parsing preferences of Spanish-English bilinguals on sentences similar to 2.19. Their experiment consisted of two parts, both on-line. In one part subjects were presented with sentences in English, in the other the same subjects were presented with sentences in Spanish. They found no difference in initial preferences between languages (ruling out any account where statistics or parameters are learnt separately for different languages). However, they did discover a strong correlation for individual subjects' preferences across languages; that is, stable, cross-linguistic, individual differences.

Such differences could result from some physiological or psychological variation between individuals rather than statistical patterns in the language they have been exposed to. For example, they may be due to differences in the individuals' working memory capacity (MacDonald, Just & Carpenter, 1992). However, such a position cannot account for cross-linguistic differences without postulating a correlation between nationality and working memory.

2.4.3 Statistical Models

The Garden Path theory clearly predicts that individuals should make similar initial decisions when faced with the same linguistic structure, irrespective of language. It is therefore incompatible with both the cross-linguistic and intra-linguistic data we have presented. Referential Support and Generalised Theta Attachment are unpredictive for the ambiguities considered here; they are therefore compatible with the data but cannot explain it. Finally, Construal may offer an explanation for the cross-linguistic data, but does not explain the intra-linguistic variation.

However, a number of heuristic approaches that could be considered statistical or frequency-based might account for this data. In this section we consider three different (partially) statistical models that have been proposed in the sentence processing literature.

Lexical Frame Models

Ford, Bresnan and Kaplan (1982) found that initial choices in prepositional phrase attachment did not follow the predictions of the garden path theory. They therefore suggested that individual lexical items played a guiding role in parsing.[11] In particular, when a verb has a number of alternative lexical subcategorisation frames, these are ranked, and the "strongest" form dictates the initial analysis pursued. We may assume this ranking is based on the previous linguistic experience of the individual, and so this model counts as one of the earliest frequency-based models.

Unfortunately, Ford *et al.*'s analysis relies on a top-down parser which only attaches constituents into the ongoing syntactic analysis once they are complete. The evidence reviewed in section 2.2.3 renders such a model unlikely. They can also give no account for the cross- and intra-linguistic data reviewed in the last two sections, as in both cases the novel phrase is not subcategorised for by the verb.

Parameter Setting

An alternative semi-statistical model has been proposed by Gibson, Pearlmutter, Cansesco-Gonzales and Hickok (1996). They suggest that a number of different strategies compete to determine the attachment of a novel phrase. Consider again example 2.19 (reproduced below as 2.24):

(2.24) a. Someone shot the servant of the actress *who was on the balcony.*
b. Alguien disparó contra el criado de la actriz *que estaba en el balcón.*

In this case, they suggest the competing heuristics are 'recency' (attach to the nearest site) and 'predicate proximity' (attach modifiers as closely as possible towards the sentence head). These heuristics pull in different directions. Gibson *et al.* show that it is possible to model cross-linguistic differences by supposing that at least one of these heuristics is parameterised; its relative weight can have (at least) two different

[11] See section 2.5 for other lexically-driven models.

values. In Spanish, predicate proximity would have greater weight, resulting in high attachment, whereas in English low attachment would be forced by the greater weight attached to recency.

We may reasonably assume these parameters are set by linguistic experience; the heuristic that most often leads to the correct analysis gains the greatest weight. As such, this is a very coarse-grained statistical model.

Parameters could only explain individual difference within linguistic communities if the average weights of the competing heuristics were very similar. This could be the case for these two heuristics in English; however, this would suggest individual differences across a vast range of modifier constructions. If it is found that only some constructions result in individual differences, or that individuals show different preferences for different constructions, then it would be necessary to postulate more parameterised heuristics. In this case, the model would become very similar to the Tuning Hypothesis.

The Tuning Hypothesis

The final statistical model we consider is the Tuning Hypothesis (Mitchell & Cuetos, 1991). In Tuning, the linguistic statistics are not used to assign weights to rationalist heuristics; instead the heuristic itself is fundamentally statistical. Simply put, when faced with linguistic ambiguity an individual will "initially opt for the resolution that has turned out to be appropriate most frequently in the past" (Cuetos *et al.*, 1996, pp.154–155).

The Tuning Hypothesis can therefore be seen as making two claims. These are:

- In the face of structural ambiguity in the linguistic input, the HSPM makes initial decisions based on statistical information.
- This statistical information is derived from the individual's previous experience of language.[12]

For example, consider example 2.24 again. The initial decision of the HSPM would

[12] Note that the authors make no commitment as to whether statistics are amassed individually for each language or across languages. Corley and Caldwell's (1996) results (section 2.4.2) support the latter version of Tuning.

be guided by previous encounters with ambiguities that are syntactically similar. The Tuning Hypothesis makes no commitments as to what constitutes syntactic similarity (though see section 2.4.4); for exposition, we will suppose that statistics are collated over all structures consisting of a noun phrase followed by prepositional phrase followed by a relative clause (NP–PP–RC, for short). A corpus count of English (Corley & Corley, 1995) shows that such sequences are resolved to the low attachment analysis approximately 65% of the time, so the Tuning Hypothesis predicts a preference for low attachment. In contrast, results from a small Spanish corpus (Huergo, cited in Corley & Corley, 1995) indicates a 62% bias towards high attachment, so this is also the prediction of Tuning.

The Tuning Hypothesis cannot be seen as a complete model as it does not characterise reanalysis (and therefore gives no account of why results differed between on-line and questionnaire studies). It also makes no commitment as to what constitutes similarity when collating statistics; we turn to this issue next.

2.4.4 Issues of Grain

In our exposition of the Tuning Hypothesis, we had to make assumptions about what might constitute syntactic similarity. We assumed the HSPM collated statistics over all occurrences of NP–PP–RC. This would mean that the initial decision for 2.25a would be the same as for 2.24a, but that for 2.25b may differ (both 2.25a and b are from Mitchell *et al.*, 1995).

(2.25) a. Someone stabbed the wife of the football star who was outside the house.

b. Someone stabbed the estranged wife of the moviestar outside the house.

It may equally be the case that statistics are collated for NP–PP–modifier; this would entail that the same initial decision is made for all three sentences. There are a large number of other possibilities; the definiteness of the NPs may be taken into account, or statistics could be collated over individual lexical items – in the case of 2.25a, all occurrences of the exact words "the wife of the football star who".

Mitchell, Cuetos, Corley & Brysbaert (1995) dubbed this the "grain problem". Each of the measures mentioned above constitute a different level of granularity over

which statistics may be collated. A fine-grained model distinguishes between a large number of different ambiguities and so necessitates the collation of a large number of different statistics; the most fine-grained proposal above is the one in which individual lexical items play a role. In contrast, a coarse-grained model combines a large number of different ambiguities into a single statistic; NP–PP–modifier is the most coarse-grained of the options we suggested. In this analysis, the garden path theory can also be seen as a statistical model; MA and LC may be seen as statistical measures that capture extremely coarse-grained regularities in the language, provided it transpires that they lead to the correct decision the majority of the time when all relevant ambiguities are considered.

There is a certain amount of evidence that record keeping is not (exclusively) fine-grained. MacDonald *et al.* (1994)[13] proposed that ambiguities such as 2.25a might be determined by the relative frequency of the two head nouns ("wife" and "football star") occurring with post-nominal modifiers. On its own, such a statistic would be unable to account for the many experiments that have discovered a reliable high-attachment bias in such ambiguities in many European languages (except English); the only explanation would be "that every experiment was so badly designed that almost every noun in the first position was more readily modified (on the evidence of statistical records) than each noun in the second position" (Mitchell *et al.*, 1995, p.478).

Consider again the experimental materials used in Corley (1995) – 2.20, reproduced below as 2.26:

(2.26) a. The satirist ridiculed the lawyer of the firm...
b. The satirist ridiculed the firm of the lawyer...

Corley found stable preferences for whether a following relative clause attached to the first or second NP across the (a) and (b) forms of 2.26 for a given subject. This is at odds with the predictions of MacDonald *et al.*'s modifiability account; this predicts that there would be a preference for modifying the same NP ("the lawyer" or "the firm") in both sentences.

There is also evidence that statistical biases from definiteness are ignored in French

[13] See section 2.5.

and gender is ignored in Dutch (Corley, Mitchell, Brysbaert, Cuetos & Corley, 1995; Mitchell *et al.*, 1995). Finally, Traxler and Pickering (forthcoming) produced evidence that subcategorisation preferences of certain verbs are ignored in making initial decisions in English (see also Mitchell, 1987).

This evidence suggests that coarse-grained statistics are (at least in some cases) more important to the outcome of an ambiguity decision than fine-grained alternatives. It does not preclude the use of fine-grained statistics as well. Nevertheless, this is the position of one of the Tuning researchers:

> "While there is nothing in the current general formulation of the tuning hypothesis to rule out the retention and use of more fine-grained statistics as well as (or instead of) such coarse-grained measures, at least one of the present authors (D.C.M.) favors variants of the model which ignore such information in initial decision-making."
>
> (Mitchell, Cuetos, Corley & Brysbaert, 1995, pp.476–477)

The issue of granularity resurfaces in chapter 3, where we discuss modularity.

2.4.5 Evidence Against Statistical Models

We have considered evidence that suggests that parsing decisions are not made on the basis of purely fine-grained statistical information. Recently, evidence against purely coarse-grained record keeping has also been produced. Gibson, Schütze and Salomon (1996) examined sentences in which a conjoined NP can attach to any of three preceding NPs. Examples are shown in 2.27.

(2.27) a. The salesman ignored the customer with the child with the dirty face *and the wet* diaper.

b. The salesman ignored the customer with the child with the dirty face *and the one with the wet* diaper.

c. The salesman ignored the customer with the child with the dirty face *and the one with the* baby with the wet diaper.

In each of these sentences, pragmatic and contextual information favour a particular analysis; in 2.27a, low attachment is preferred, in 2.27b, middle attachment and in 2.27c, high attachment. Gibson *et al.* (1996) performed a questionnaire study in

which subjects were asked to rate such sentences according to how easy or hard they were to understand on first reading. They found that subjects rated high attachment sentences as easier to understand than middle attachment alternatives; however, a corpus study determined that middle attachments occur more frequently than high attachments.[14] This result is not compatible with a structural frequency theory such as Tuning, in which the individual is expected to prefer the most frequent reading.

However, given that Gibson *et al.*'s results were not on-line, there is a possibility that the preference for high attachment does not reflect initial decisions. Gibson and Schütze (1996) therefore performed an on-line study using similar sentences (as in 2.28):

(2.28) a. The sportswriter wrote a column about a soccer team from the suburbs *and one about a* baseball team from the city for the paper's Sunday magazine.

b. The sportswriter wrote a column about a soccer team from the suburbs *and one from the* city for the paper's Sunday magazine.

Again, disambiguation is due to pragmatic and contextual information; in 2.28a the conjunctive phrase must be attached high, whereas in 2.28b disambiguation favours middle attachment. Subjects read such sentences on-line, including comprehension questions which indirectly probed which attachment site had been selected. Gibson and Schütze found that subjects answered questions following high attachment sentences correctly significantly more often than those following middle attachment; they therefore infer that subjects often prefer high attachment even when it is implausible. On-line reading times in the two word region following the disambiguating noun ("baseball team" or "city for") also favoured high attachment.

Brysbaert and Mitchell (1996a) performed an on-line experiment on NP–PP–RC ambiguities in Dutch, such as 2.29:

(2.29) a. De gangsters schoten op de zoon van de actrice *die op het balkon zat met* zijn arm in het gips.
The terrorist shot the son of the actress who was on the balcony with his

[14] This difference was non-significant for the Wall Street Journal corpus and marginally significant for the Brown corpus.

arm in a cast

b. De gangsters schoten op de zoon van de actrice *die op het balkon zat met* haar arm in het gips.
The terrorist shot the son of the actress who was on the balcony with her arm in a cast

They discovered a significant preference for high attachment, in line with the cross-linguistic evidence presented earlier. This evidence was taken as support for the Tuning Hypothesis.

Brysbaert and Mitchell (1996b) then replicated this experiment using slightly modified materials.[15] They again found a significant preference for high attachment in both self-paced reading and eye-tracking studies. However, they also studied two Dutch corpora to determine whether this preference represented a regular pattern in the language. The first corpus was a CD from Dutch publisher Roularta, containing the popular magazine titles "Knack", "Trends" and "Style". The second corpus (of which they only studied a sample) was of the daily newspaper "Volksrant".

The results of this corpus search are shown in table 2.1.

	Knack100		**Knack200**		**Trends**		**Style**		**Volksrant**	
Attach	**die**	**dat**	**die**	**dat**	**die**	**dat**	**die**	**dat**	**die**	**dat**
High	40	14	22	15	27	13	25	24	14	12
Low	104	21	77	17	54	23	91	27	44	11

Table 2.1: RC attachments in Dutch (Brysbaert & Mitchell, 1996b)[16]

The corpus study indicates a 69% bias towards low attachment in Dutch. However, this bias is at odds with the initial decision of Dutch speakers; thus a model in which initial decisions are made solely on the basis of experience-based statistics at this level of granularity is contradicted. Brysbaert and Mitchell report that they have also "completed numerous other analyses using other grains (e.g. classifying heads as

[15] Disambiguation occurred immediately after the relative pronoun to ensure that the experiment was tapping into the *initial* decisions of the subjects – see Mitchell, Corley and Garnham (1992).

[16] For each title, there are two columns; one for ambiguities using the non-neuter relative pronoun "die" and the other for those using the neuter version "dat". The number given is the total of all clearly disambiguated and debatable attachments; however, the ratio of high to low attachments does not change significantly if you exclude debatable items.

human and non-human)" (Brysbaert & Mitchell, 1996b), but they have not found any kind of high attachment preference at any level of grain. If the corpora were truly representative of the linguistic experience of the experimental subjects, this study suggests that initial decisions cannot be made purely on the basis of statistical information.

2.4.6 Summary

The evidence presented in the early parts of this section supports a statistical model of human sentence processing. Cross-linguistic differences are not compatible with heuristic models where decisions are made on purely structural or grammatical grounds. They also suggest that Referential Support alone cannot offer a complete account of the parsing heuristic used by the HSPM. Stable individual differences between speakers within the same community provide further evidence against uniform parsing heuristics; in order to explain them an account must propose that differences in nature or nurture affect parsing preferences. If we wish to give a single account of both cross-linguistic and intra-linguistic differences then a nature (innate) account is difficult to justify; the most likely nurture account is a frequency-based one.

We reviewed the few heuristic models that could be deemed statistical; of these, the Tuning Hypothesis (Mitchell & Cuetos, 1991) is the account that best fits the notion we appeal to when we talk about statistical models throughout the rest of this thesis.

However, there is growing evidence that the Tuning Hypothesis cannot be entirely correct; we considered this in section 2.4.5. This evidence suggests that the decision making processes within the HSPM are not entirely statistical. In chapter 3 we introduce the "Modular Statistical Hypothesis" which captures the notion of a (partially) statistical HSPM. Within this framework, we argue for a position in which statistical information is used only when it offers significant benefit to the sentence processor. We call such use *strategic*.

2.5 Constraint-Based Models

The models proposed so far share two features. Firstly, the parser is serial (though see section 2.3.3). Secondly, the parser makes an initial decision based on some

subset of the available information. The reasons for proposing a model of this kind are explored in chapter 3. In this section we consider constraint-based models, which are parallel and deny the existence of a distinct initial decision mechanism; instead, all possible analyses are computed in parallel, but they are assigned varying weights or activations.

A number of different authors have proposed constraint-based models (St. John & McClelland, 1990; Trueswell & Tanenhaus, 1994; Spivey-Knowlton & Sedivy, 1995; MacDonald, Pearlmutter & Seidenberg, 1994). While these models vary in specific details, it is sufficient for the current work to give an overview of the constraint-based approach. For the purpose of exposition, we concentrate on a single model – that proposed by MacDonald *et al.* (1994) – as this is the most concrete model suggested to date to match human behaviour data.

2.5.1 The Basic Model

MacDonald *et al.*'s model differs from those considered in sections 2.3 and 2.4 in four key ways:

- Multiple access is possible at all levels of representation, but is constrained by frequency and context.
- All levels of representation are available to the language processor at the same time.
- The lexicon is vastly enriched, including frequency and syntactic information; lexical entries are "built" rather than accessed.
- Language processing is viewed as a constraint satisfaction problem between lexical entries, with no subsystem corresponding to a syntactic parser.

The first two differences are those we have already mentioned; instead of a model in which a single analysis is pursued based on the initial decision of a heuristically-guided parser, they propose that syntactic structures are generated in parallel. The HSPM still makes an early decision (see section 2.2) because each analysis has an associated *activation*; the one with the greatest activation is currently preferred. We

assume that conscious garden paths arise from the HSPM discarding analyses that have too low an activation, such that reactivation is costly or impossible.

The enriched lexicon is something that many constraint-based models share – these models are often called 'lexicalist' models. MacDonald *et al.* propose that there is no boundary between lexical access and syntactic structure; lexical entries are constructed rather than retrieved, and they contain partial syntactic structures which are simply 'linked' to create a complete analysis. Analyses gain or lose activation depending on how well they fulfil various (lexically specified?) constraints; the task of the sentence processor is therefore to determine the analysis that best satisfies all constraints.

2.5.2 An Example

This model is probably best understood through an example. Consider 2.30:

(2.30) John examined the evidence.

Figure 2.4 depicts part of the lexical representation of the proper noun "John". This representation includes a number of features: semantics, lexical category, argument structure, X-bar structure and thematic role. "John" can have several thematic roles; the frequency with which each role occurs is encoded in the lexical representation and inhibitory links realise the exclusive nature of alternative readings. For diagrammatic simplicity, we simply enclose the most frequent reading in figure 2.4 in a thick box and other readings in thinner boxes.

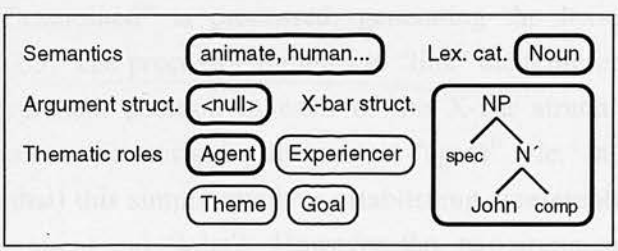

Figure 2.4: Partial lexical representation for "John"
(from MacDonald, Pearlmutter & Seidenberg, 1994)

Figure 2.5 depicts part of the lexical representation of the verb "examined". It

includes semantics, lexical category, voice, argument structure[17], X-bar structure and tense. Again, alternative analyses have attached frequencies, represented by thick and thin boxes. There are also dependencies between a number of the features; for example, if "examined" has past tense then it also has active voice; these are again realised by excitatory and inhibitory links within the representation.

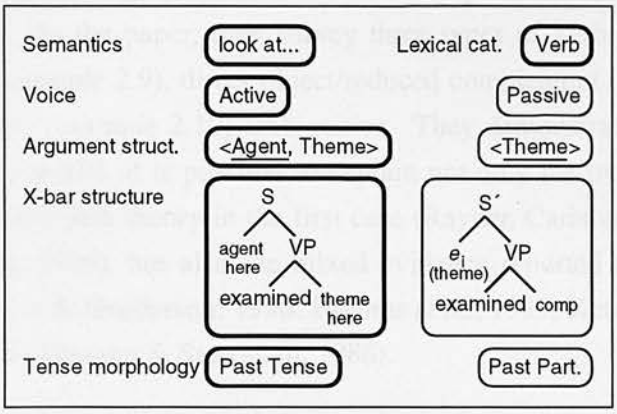

Figure 2.5: Partial lexical representation for "examined"
(based on a figure in MacDonald, Pearlmutter & Seidenberg, 1994)

Using these representations, we can outline how processing occurs. When the processor encounters the word "John", it builds the representation depicted in figure 2.4. Agent is the most frequent thematic role so, in the absence of biasing context, it will be the most activated and inhibit the other possible roles; the early decision of the sentence processor is that "John" will be assigned agent role.

Next, the word "examined" is processed, generating the lexical representation depicted in figure 2.5. The processor attempts to "link" the representation for "John" to the external argument position in each of the X-bar structures. In the two argument case, this involves assigning the subject "agent" role. In the one argument case, (we assume that) this simply involves establishing coreferentiality between the empty external argument and "John". However, the two argument reading is more frequent and therefore initially has greater activation. The agent reading of "John" already has greater activation. This analysis is therefore preferred and excites related features (active voice and past tense); it also inhibits the alternative reading.

[17] The argument structure for verbs also incudes external arguments, underlined in the figure. The position of these arguments is encoded within the X-bar structure.

The final portion of the sentence is "the evidence". This is consistent with the preferred analysis and can be linked into the structure as an internal argument of the verb. The model therefore chooses to assign an active analysis to this sentence.

2.5.3 Explaining the Data

MacDonald *et al.* (1994) argue that their model can explain a wide range of existing experimental data. In the paper, they survey three types of ambiguity: main verb/ reduced relative (example 2.9), direct object/reduced complement (example 2.7) and prepositional phrase (example 2.10) ambiguities. They demonstrate that, within the framework of their model, it is possible to explain not only the overriding evidence supporting the garden path theory in the first case (Rayner, Carlson & Frazier, 1983; Ferreira & Clifton, 1986), but also the mixed evidence reported for the latter two ambiguities (Ferreira & Henderson, 1990; Holmes *et al.*, 1987; Kennedy *et al.*, 1989; Clifton *et al.*, 1991; Altmann & Steedman, 1988).

While it is possible that some of the evidence against the use of statistics (outlined in sections 2.4.4 and 2.4.5) might prove difficult to explain within a constraint-based model, we do not present experimental evidence that disproves this model. Instead, the question of constraint-based and interactive architectures is left until chapter 3, where we argue that it is precisely the ability of these models to account for so much that is their greatest flaw.

2.5.4 Summary

Unlike the heuristic models considered in sections 2.3 and 2.4, constraint-based models do not rely on initial decisions and reanalysis. Instead, all analyses are constructed and they compete, gaining activation by 'linking' with active sub-analyses, and being inhibited by violations of grammatical or other constraints and by each other. At any time, the most preferred analysis is the most active one.

In this section we considered a particular constraint-based model put forward by MacDonald, Pearlmutter and Seidenberg (1994). In common with most constraint-based approaches, it posits that grammatical information is stored in the lexicon, rather than in a separate syntax. We gave a very brief example of how this model might be expected to analyse a simple English sentence and mentioned the data that MacDonald *et al.* (1994) explain within the framework of this model. We defer

criticism of this and other interactivist models until chapter 3.

2.6 Conclusions

In this chapter we have reviewed several models of sentence processing, focussing on their incorporation and explanation of statistical aspects of language processing. These range from extremely simple models relying on a small number of structural heuristics to choose a single preferred analysis (section 2.3) to complex models in which multiple analyses compete for activation (section 2.5). The middle ground was taken by models that still propose that there is an initial decision based on some heuristic, but allow this heuristic to be statistical. We develop such a model in chapter 4.

The models reviewed here are important for a number of reasons. Firstly, many of them may be extended to make predictions in the face of lexical category ambiguity. In chapter 4 we return to this question and throughout this thesis we compare the predictions of these models with those of our own model. Secondly, they exemplify some of the divisions between models of sentence processing; the debate about whether the processor makes initial decisions based on some heuristic is recouched as the "modularity debate" in chapter 3 and we return to the statistical/non-statistical debate throughout this thesis. Finally, it is important to understand the reasons why the use of statistics within models of human sentence processing is becoming increasingly fashionable.

In the next chapter we examine the major division between the heuristic models (sections 2.3 and 2.4) and the constraint-based models (section 2.5) and we argue why a statistical heuristic architecture should be preferred.

3: Modularity and the HSPM

3.1 Introduction

In chapter 2 we explored a number of models of sentence processing. The heuristic models are based on the assumption that the parser makes an *initial decision* based on a subset of the available evidence, subject to later *reanalysis*. In contrast, researchers proposing constraint-based models reject the notions of initial decisions and reanalysis; instead, a single processing mechanism uses all available information to excite or inhibit alternative parallel analyses.

Another way of viewing this distinction is that heuristic models are *modular*, whereas constraint-based models are *interactive*. This chapter concentrates on the modularity debate; we look at the original definition of modularity from Fodor (1983). We then consider what is meant by the term 'modular' in reference to current models of human sentence processing.

In section 3.3 we turn to modular statistical models. We consider what it means for a model to be both modular and statistical, and what characteristics we would expect such a model to have. In doing so, we introduce the Modular Statistical Hypothesis; the remainder of this thesis explores and provides evidence for this hypothesis by motivating one possible modular statistical model of human sentence processing. The evidence presented in chapters 6 to 8 suggests that if we prefer a modular model of mind, this model should be (partially) statistical.

Our reasons for preferring modularity are therefore central to this thesis. We do not interpret or produce any evidence that could not be explained by a constraint-based model, given the correct weighting of different constraints. However, the model introduced in chapter 4 is modular. In section 3.4 we argue why a modular model of mind should be preferred to an interactive one and why a constraint-based model may explain our results, but could never *predict* them. In section 3.5 we consider whether existing evidence already falsifies either modular or interactive positions.

3.2 Expositions of Modularity

The term 'modularity' has been used by a number of different authors to refer to slightly different concepts (see Spivey-Knowlton & Eberhard, 1996). The most famous exposition is Fodor's (1983) book '*The Modularity of Mind*'. However, the notion does not originate with this book – Forster's (1979) 'Autonomy Hypothesis' is an explicit proposal of a modular sentence processor and many early models of sentence processing make tacit assumptions about the modularity of the HSPM (Frazier & Fodor, 1978; Kimball, 1973).

In this section we briefly introduce Fodor's hypothesis and then consider what is meant by modern authors when they refer to a 'modular HSPM'.

3.2.1 Fodor's Modularity of Mind

Fodor (1983) suggested that our cognitive architecture is divided into (at least) two types of faculty: 'input systems', which consist of encapsulated modular processes, and 'central processes', which do not. Fodor claims that "the distinction between the encapsulated mental processes and the rest is – approximately but interestingly – coextensive with the distinction between perception and cognition" (Fodor, 1987, p.27).

Fodor proposes that a 'modular' process differs from a non-modular process in eight key ways. We take four of these to constitute Fodor's definition of modularity:

- A module is informationally encapsulated.
- A module produces 'shallow' output.
- A module is domain specific.
- A module is mandatory.

Most research into syntactic modularity has concentrated on the first of these claims. An 'informationally encapsulated' process has access to only a limited subset of the contextual and stored information available within the individual's brain. That is, it must make decisions based on incomplete knowledge; such a process will sometimes reach an incorrect conclusion even when the information necessary to make a correct decision is available:

> Rushing the hurdles and jumping to conclusions is, then, a characteristic pathology of irrational cognitive strategies, and a disease that modular processors have in spades.
>
> (Fodor, 1987, p.26)

To preserve information encapsulation, a module's input must be 'shallow'. That is, it can only include the representation calculated by the previous module; it should not also include or refer to the information used to calculate this representation. If a module does not have shallow input then it is not truly informationally encapsulated: information supposedly denied to it may be included in its input. Such information could be used to second guess the decisions of a previous module, rendering the system inefficient. It follows that the output of a module must also be shallow, so it does not pass unwarranted information on to a subsequent module.

A modular process is domain specific. In other words, it is dedicated to performing one *and only one* processing task. In contrast, central processes are flexible in that they can be put to different uses at different times. To an extent, the encapsulation of modular processes guarantees their domain specificity; such a process cannot perform any other task because it does not have access to the knowledge required to perform those tasks.

The final differences between an input module and a central process is that the former is mandatory; it is not possible to exert conscious control over the behaviour of a module. For example, we cannot decide whether to understand speech that we hear; once we have heard it we are at the mercy of our perceptual systems. In this way, modular processes act like reflexes.

Why would we expect a cognitive faculty to behave in (what Fodor himself terms) an irrational way? There are a number of answers to this; in section 3.4 we propose our own. Fodor's answer lies in the four characteristics of modularity we have so far ignored:

- Modular systems are fast.
- Input systems exhibit characteristic and specific breakdown patterns.
- Central processes have only limited access to the representations the input

systems compute.

- Input systems are associated with a fixed neural architecture.

While Fodor does not distinguish between the eight characteristics of a modular system, we view the first four as defining modularity, and the latter four as consequences of that definition. The speed of modular systems is taken to account for the apparent rapidity of perceptual processing; for example, if we had to take into account all our knowledge in making parsing decisions, we would be in a position of never knowing when to stop. However, whether modularity is necessary to account for this speed is subject to debate; we return to this question in section 3.4.

It is not relevant to repeat Fodor's arguments for the other characteristics here. Instead, we simply note that Fodor happily mixes rationalist and empiricist arguments, and the latter are based both on intuition and existing experimental results. One of the intuitive arguments Fodor has repeatedly used is that the persistence of some optical illusions despite one's knowledge that they are illusions "strongly suggests that some of the cognitive mechanisms that mediate visual size perception must be informationally encapsulated" (Fodor, 1987, p.25). However, Churchland (1988) has argued that other illusions are "penetrable by higher cognitive assumptions" (p.171). In section 3.5 we consider some of the experimental empirical evidence that supports, and fails to support, Fodor's modularity hypothesis as it applies to the HSPM.

The final part of the modularity argument focuses on which cognitive processes form part of the input system and which are central processes. Fodor divides the language processor at the boundary between "the subject matter of linguistic theory (construed as formal syntax) and the subject matter of disciplines such as pragmatics and discourse analysis" (Marslen-Wilson & Tyler, 1987, p.37). In other words, the parser is a modular input system, but inferential semantic processes are not. According to Fodor, the encapsulated parser has access to only "the acoustics of the input and the grammar" (Fodor, 1987, p.28). In the next section we consider whether this division of labour within language processing is still favoured by modern proponents of modularity.

3.2.2 Modern Views of Modularity

Fodor's modularity hypothesis makes two distinct claims with regard to human sentence processing. The first is that input processes are modular and central processes are not. The second is that syntax constitutes an input process, whereas areas such as pragmatics and discourse analysis do not.

Fodor's position is at odds with a number of modern models of sentence processing. In particular, a number of authors have proposed 'modular' models in which the boundary between input and central processes differs from Fodor's proposal. For example, Frazier (1990) suggests that reference and θ-prediction form 'pseudo-encapsulated modules' which follow syntactic processing (and therefore would be labelled central processes in Fodor's account); she also argues that the bandwidth of communication between modules is extremely low during initial decisions, but there is far more interaction during reanalysis.

It seems that some modern views of modularity disagree with Fodor on exactly what constitutes a module, and on the boundary between modules and input processes. Modular and interactive positions also carry a lot of baggage; Spivey-Knowlton and Eberhard (1996) summed up the situation using a four-dimensional graph, reproduced here as figure 3.1.

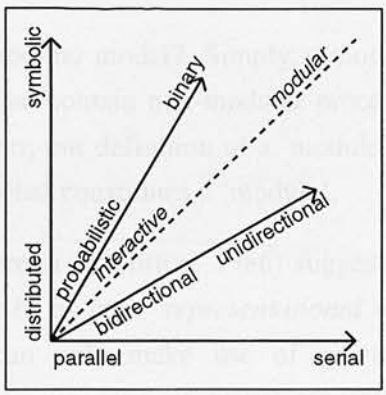

Figure 3.1: The "continuum" of modularity
(from Spivey-Knowlton & Eberhard, 1996)

They argue that modular positions tend to be symbolic, binary, unidirectional and serial; such a model would be placed in the top right of figure 3.1. In contrast,

interactive models tend to be distributed, probabilistic, bidirectional and parallel; they would therefore be placed in the bottom left of figure 3.1. This representation is certainly an accurate reflection of the positions that have traditionally been taken by modular and interactive researchers. However, Spivey-Knowlton and Eberhard suggest that "when a model is specified in enough detail to be associated with a region in this space, that region's projection onto the continuum of modularity indicates the *degree to which* a model is modular" (pp.39–40, their italics).

Spivey-Knowlton and Eberhard's position turns a historical accident into a definition. As we argue in section 3.3, there is no reason a modular model cannot be probabilistic or statistical. Distributed and parallel modular models are also viable. The only criterion which does (to some extent) affect modularity is bidirectionality; while Fodor takes no overt position on reverse information flow between modules, it is clear that if a module bases its decisions on information passed from a later module then it is not fully encapsulated.

It appears that modularity means different things to different people. Fodor's proposition is too restrictive; it does not include many models that would be considered modular, such as that proposed by Frazier (1990). Defining modularity on the basis of characteristics associated with (rather than central to) modular models does not seem a viable solution either.

So what does count as a modular model? Simply, a modular model contains one or more modules (and may also contain non-modular processes, equivalent to Fodor's central processes). However, our definition of a 'module' also differs from Fodor's; we therefore need to state what constitutes a 'module'.

Frazier (1985, cited in Ferreira & Clifton, 1986) suggested that a module can only process information stated in its own '*representational vocabulary*'. For example, the syntactic processor can only make use of grammatical information; it is insensitive, for example, to semantic or visual representations. This stipulation neatly captures a version of information encapsulation; we consider this a defining characteristic of a modular model.

The second defining characteristic is that a module is *independently predictive*. That is, we do not need to know about any other component of the cognitive architecture

to make predictions about the behaviour of a module (provided that we know the module's input).

Finally, a module has *low bandwidth* in both feedforward and (particularly) feedback connections. By this we mean that it passes a comparatively small amount of information (compared to its internal bandwidth) on to consequent modules; more importantly, it receives very little information back from these modules. The feedforward requirement guarantees that subsequent modules get no insight into the internal decision procedures of previous modules; this is equivalent to shallow output. The feedback stipulation guarantees that information flow within a modular model is largely unidirectional.

These three defining properties of a modular architecture overlap. If one module cannot understand the representational vocabulary of another, then information about its internal decision process is useless; thus we would not expect such information to be passed on. Similarly, a module cannot be independently predictive if its decisions depend on representations, constructed by other modules, that are not part of its input – independent prediction is therefore tied to low bandwidth feedback connections.

3.2.3 Summary

Fodor (1983) argued that cognitive faculties are divided into input processes, which are modular, and central processes, which are not. By a modular process, he meant one that is informationally encapsulated, domain specific and mandatory, and has 'shallow' inputs and outputs. He suggested that the divide between input and central processes is roughly coextensive with the divide between perception and cognition; in the case of language, he located this divide between the subject matter of formal linguistics and that of pragmatics and discourse analysis.

We suggest that Fodor's definition no longer reflects what is meant by 'modular' when referring to current models of the language processor. In particular, modern modular models have muddied the divide between input and central processes, and 'modules' have been proposed that do not fulfil all of Fodor's requirements. Instead, we suggest a model is modular if it is composed of individual processes which use their own representational vocabulary (and ignore that of other modules) and are independently predictive, and if the feedforward and (particularly) feedback

connections between modules have low bandwidth. In the next section we consider what it would mean for such a model to be statistical.

3.3 Statistical Modularity

In section 3.2 we introduced the concept of modularity; as Spivey-Knowlton and Eberhard (1996) noted, modularity is normally associated with binary rather than statistical decision procedures. In this section we therefore consider what it means for a model to be both statistical and modular; in section 3.4 we argue why such a model should be preferred to an interactive account.

3.3.1 The Modular Statistical Hypothesis

We define statistical modularity by introducing the 'Modular Statistical Hypothesis' (MSH):

The Modular Statistical Hypothesis:

> The human sentence processor is composed of a number of modules, at least some of which use statistical mechanisms. Statistical results may be communicated between modules, but statistical processes are restricted to operating within, and not across, modules.

This hypothesis encompasses a range of possible models, including the coarse-grained architecture espoused by the Tuning researchers (Mitchell, Cuetos, Corley & Brysbaert, 1995) and the partial model proposed in chapter 4. However, it excludes interactive models such as those proposed by MacDonald *et al.* (1994) and Trueswell and Tanenhaus (1994). The models that fall within the MSH are a subset of those that are modular, as we defined modularity in section 3.2.2.

This hypothesis is the one of the central tenets of the current work. We argue for it in sections 3.4 and 3.5 and the later chapters of this thesis are both situated within the framework encompassed by this hypothesis, and provide evidence supporting it.

3.3.2 A Statistical Module?

We explore the MSH by first considering what it would mean for a module to be statistical. An encapsulated or partially encapsulated (see Frazier, 1990) module does not have access to all cognitive and contextual knowledge; decisions are made

on a heuristic basis using locally available information. In the case of a statistical module, this heuristic is based on statistical knowledge. Such knowledge may be derived by the module itself, or may be imposed on the module by some higher level of processing; however, the latter position implies that higher levels of processing have access to the modules internal representations and is therefore less compatible with full information encapsulation.

Assuming that the module collates statistics itself, it must have access to some measure of the 'correctness' of its decision; this could be derived from whether reanalysis was requested by later processes or not. The most restrictive modular statistical model is therefore one in which modules are fully encapsulated and only offer a single analysis to higher levels of processing.

The statistical measures such a module depends on are thus architecturally limited. They cannot include information pertaining to higher levels of processing, as these are not available to the module. Assuming very low bandwidth feedforward connections or shallow output (see section 3.2), it is also impossible for the module to collate statistics concerning levels of representation that are the province of modules that precede it. A modular architecture therefore strongly constrains the sort of statistical information that may inform a decision; this is in contrast to an interactive architecture, in which there is no such constraint on the decision process (see sections 3.4).

While existing modular models tend to be serial, there is nothing in the MSH that precludes parallel processing. As a statistical module may have to determine which of a number of analyses is more probable, it may well construct several analyses in parallel *internally*. However, as discussed above, a position in which module *output* is serial (only one analysis is passed on to subsequent modules) is more constraining; it therefore makes sense to initially prefer such an architecture.

3.3.3 Statistical Reanalysis

If the HSPM is (partially) probabilistic, then we would expect reanalysis to regularly occur prior to absolute disambiguation. Consider a situation in which there are two statistical modules, A and B. A's output is passed to B; in turn, B may force reanalysis in A. This situation is depicted in figure 3.2.

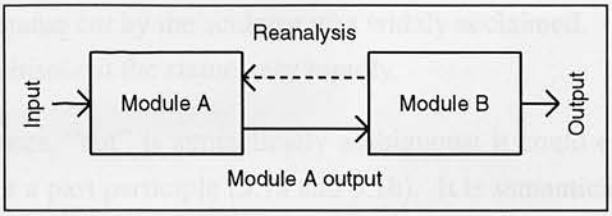

Figure 3.2: A simple modular model

There are two possible scenarios for reanalysis within such a model. The first (more traditional) possibility is that B requests an alternative analysis from A only when B cannot compute any output from its input. Effectively, B forces reanalysis when all outputs have probability zero. However, there is no reason why B cannot request an alternative analysis earlier; the second scenario is that B forces reanalysis in A when it can compute no output with probability greater than a given threshold (itself ≥ 0). Such a model entails that reanalysis would occur when the current analysis becomes unlikely, which may often be prior to absolute disambiguation; this prediction is consistent with experimental results reported in Mitchell, Corley and Garnham (1992).

3.3.4 Back Door Semantics?[18]

Statistical mechanisms record the final outcome of sentence processing for a given input. However, this outcome depends on decisions made by other modules than that which collates the statistics. For example, if we collate statistics at a syntactic level, they will reflect not just the choices of the parser, but also those of semantic and pragmatic processes, because the parser has access to information concerning the correctness of its initial decisions (i.e. whether it is asked for an alternative analysis or not). Crucially, modularity is preserved as the statistics concern only entities in the representational vocabulary of the parser.

A statistical HSPM therefore allows other levels of representation to affect initial decisions while preserving modularity, but only based on previous experience, not, for example, on current context. Exactly what the influence of these levels is depends on the granularity at which statistics are collated. Consider 3.1:

(3.1) a. The chisel *cut* by the sculptor was widely acclaimed.

[18] I am indebted to Matt Crocker for pointing out this characteristic of statistical modularity to me.

b. The statue *cut* by the sculptor was widely acclaimed.

c. The chisel *cut* the statue very rapidly.

In all three sentences, "cut" is syntactically ambiguous; it could either be an active verb (as in 3.1c) or a past participle (3.1a and 3.1b). It is semantically plausible for a chisel to cut something but not for a statue to do so; evidence that we initially choose the active reading of "cut" in 3.1a and 3.1c, but the past participle reading in 3.1b, could therefore support an interactive HSPM in which semantic plausibility plays a role in the initial decisions of the parser.

However, statistical mechanisms offer us an alternative explanation. Suppose the parser collates very fine-grained statistics, including the co-occurrence of words and lexical categories. For example, one relevant statistic might be the probability of "cut" occurring as an active verb following "chisel" occurring as a noun. We can represent this as {chisel, noun}–{cut, active verb}. We would expect this sequence to have a higher probability than {chisel, noun}–{cut, past participle}, as the former sequence occurs more often in language. The decision of the parser would therefore initially be in favour of the active reading in sentences 3.1a and 3.1c, and we would expect a garden path effect in 3.1a when the prepositional phrase "by the sculptor" is read.

In contrast, the sequence {sculpture, noun}–{cut, active verb} is uncommon; the alternative {sculpture, noun}–{cut, past participle} *may* be more frequent. In this case, we would expect a parser guided *solely* by such fine-grained statistics to choose the reduced relative reading in 3.1b; there would then be no garden path effect when the following prepositional phrase was read. The parser would have made an initial decision that is apparently semantically motivated, without access to any explicit semantic representations.

This "back door semantics" is a feature of any statistical model of parsing, using any level of statistical granularity. However, the granularity does restrict the type of semantic effects we would expect the parser to exhibit; for example, lexical effects will only occur if the parser makes explicit use of statistics collated over lexical representations. This 'semantic clairvoyance' is limited in other ways. Firstly, a statistical parser can only have a semantically motivated bias for information that has

an analogue in its representational vocabulary; this restriction rules out, for example, visual context having any effect on the initial decisions of the parser. Secondly, these semantic effects may only refer to *local* context; if we assume that statistics are collated over adjacent representational units[19] (constituents) and a single statistical measure may not include more than n units, then the context which may affect syntactic decisions is the preceding $(n-1)$ units. In particular, it is not possible for information from preceding sentences to affect the initial decisions of the parser (except by priming); evidence for Referential Support (Crain & Steedman, 1985; Altmann & Steedman, 1988) could not be explained by a modular statistical model.

A modular statistical model therefore allows limited semantic and pragmatic preferences to affect the initial decisions of the parser without significant increase in processing cost. The HSPM gains from such limited use of statistics in that reanalysis will occur less frequently. In general, statistical measures within a modular model are, by hypothesis, *strategic*; such mechanisms are not architecturally necessary to the model, nor are they a basic assumption of the framework, and we would therefore not expect them to be used where they do not offer a significant processing advantage. In contrast, a constraint-based model such as that outlined in section 2.5 is critically dependent on statistical information (in that such a model relies on a learning procedure which is sensitive to statistical patterns in language); the use of statistics in such a model cannot be strategic.

3.3.5 Summary

In this section we introduced the Modular Statistical Hypothesis (MSH) and considered what it means to be a statistical module. A key difference between a modular statistical architecture and an interactive one is that the modular architecture imposes a priori architectural limits on the statistics that can be used in processing; an interactive processor imposes no such limits. Thus the MSH imposes predictive restrictions on a model.

We suggested that statistical modules may be *internally* parallel, but the most constraining position is to suppose that communication between modules is serial.

[19] It is in theory possible to collate statistics over non-adjacent units. However, such a model is more complex than one which requires adjacency and does not offer any obvious processing advantage. We therefore assume that the HSPM collates statistics only over adjacent units.

We also demonstrated that we would expect early reanalysis to occur within a modular statistical model.

Finally, we considered the benefits that a modular statistical model of the HSPM offers. We argued that the strategic use of statistical measures in syntactic processing allows semantic and other higher level preferences to have a limited influence on the initial decisions of the parser at extremely low cost. Such a position still preserves modularity; in the next section we consider other reasons why a modular position may be preferable to an interactivist one.

3.4 Rationalist Arguments

The modularity hypothesis is still a hot topic; new modular (Frazier & Clifton, 1996; Crocker, 1996) and interactive (MacDonald *et al.*, 1994; Trueswell & Tanenhaus, 1994) models have recently been proposed. The model proposed in this thesis is modular. While we demonstrate that it is empirically more accurate than alternative modular models, our arguments for preferring it to an interactive architecture are mainly rationalist. In this section we detail those arguments.

3.4.1 'Dominance' and 'Redundancy'

Before we can argue about the merits of various statistical architectures, we need to define some terms. In section 2.4 we considered the 'grain problem'; throughout this thesis we discuss a related issue which we term '*dominance*':

Dominance:

> A particular set of statistical measures is dominant for a particular ambiguity if, in the vast majority of cases, the outcome of ambiguity decisions is the same as it would be if these measures were the only ones used in making these decisions. A statistical measure is dominant if it is a member of the dominant set.

At first glance, dominance sounds similar to granularity; however, dominance and grain are not synonymous. The grain problem is about the structural representations over which statistics are collated – a coarse-grained model makes use of more abstract representations, a fine-grained model uses more concrete ones. Dominance, on the other hand, is concerned with statistical measures rather than representations.

If a particular granularity of representation is used to inform a parsing decision, then it follows that statistical measures collated over that level of granularity are dominant for that decision. If a set of statistical measures have dominance, it does not follow that the decision is made at a particular level of grain. The dominant measures may, for instance, have been collated over a number of different grains, or other grains may be taken into account in making the decision, but have a very limited effect.

The opposite of dominance is '*redundancy*'. If a particular statistical measure makes no difference to any ambiguity decision, then it is redundant. A measure may have a very small influence, and therefore neither dominate nor be redundant. The notions of dominance and redundancy allow as to characterise both modular and interactive models using uniform terminology.

3.4.2 Speed Again

In section 3.2.1 we presented Fodor's version of the modularity hypothesis. One of his primary reasons for advocating a modular position was to explain the apparent speed of the HSPM; language processing occurs automatically and very rapidly. If choices are not based on strictly limited information, then we would not expect rapid processing; the language processor would suffer from a version of the 'frame problem' (Raphael, 1971).

Classically, the frame problem (as stated in Artificial Intelligence) is how to get a robot to appreciate the effects of its actions. When a robot performs an action, it alters the world; the robot should revise its beliefs about the world accordingly. But how can a robot delimit just those beliefs that need reconsidering after a given action? If a real world robot examined all its beliefs to determine which need to be altered, it would be paralysed; thinking too much prohibits action. The problem is when to stop.

A version of the frame problem crops up in decision making. If we attempt to make a decision based on all our knowledge, we would never stop considering and actually decide. Even if we restrict ourselves to relevant knowledge, there is no systematic mechanism for deciding what is relevant! In general, decisions must be based on an arbitrary subset of the available information.

Now consider the language processor. It is highly incremental (see section 2.2.3), so must make decisions extremely rapidly. This means that it simply cannot use all available information when making a single decision, nor can it decide which information is relevant 'on the fly'; such a system would frequently fail to make decisions within a reasonable time limit. There must therefore be some predefined limitation on the information used in decision making. It may be that such a limitation is imposed arbitrarily, or it may arise from architectural considerations. However, both possibilities are consistent with a modular position, and inconsistent with a fully interactive approach. The frame argument therefore supports a modular HSPM.

There are at least two possible replies that an interactivist researcher might make to this argument. We consider each of them below.

Distributed Processing

The first interactivist reply is to claim that the above analysis is true, but only applies to models that process data serially. In a distributed model, all information is considered simultaneously; considering more information does not lead to greater processing time.[20]

This argument is true up to a point – if everything points to the same decision, then the network will normally stabilise very quickly no matter how much information is considered.[21] However, networks are not perfect decision makers; given conflicting information they may take a long time to reach stability, or completely fail to reach stability. Further, the more inputs to the network, the more likely such failure becomes. In effect, the network is suffering from the frame problem; it is failing to terminate when faced with too much information (see Herz, Krogh & Palmer, 1991, for a detailed account of neural networks and network behaviour).

If it's not Broken...

The second interactivist reply is that researchers have implemented interactive partial models that *do* exhibit extremely rapid performance and do terminate (Tanenhaus *et*

[20] A more common version of this argument is "you don't understand distributed processing", often accompanied by a smug smile.

[21] Though even in this case the network can fail to reach a stable state if it repeatedly hops over the minimum.

al., forthcoming; St. John & McClelland, 1990). If they have done it then it must be possible; the frame problem in decision making is fictitious.

However, they would be making the same mistake as McDermott (1986, cited in Fodor, 1987). Just because we can selectively model toy scenarios in which there is a limited and strictly controlled information population, this does not mean that our models scale up to the full problem. The difficulty in knowing what information to take into account in parsing is a function of the diversity of the available information; a model in which this information diversity does not exist will not suffer from the same problem.

3.4.3 Statistical Modularity and Predictiveness

The second argument in favour of modularity relates mainly to statistical models. It is clear that a very simple non-statistical heuristic model, such as the garden path theory, is highly predictive (modulo the choice of syntactic structure). In contrast, it is harder to obtain clear predictions from constraint-based models. There are three possible reasons for this:

- Constraint-based models are relatively new; they need further work to tighten them up.
- Statistical models are less predictive than non-statistical models.
- Interactive models are less predictive than modular models.

We argue that while constraint-based models may need tightening, and obtaining predictions from statistical models may involve more work, the basic problem is that interactive models are inherently less predictive. To do this we compare a modular statistical approach (such as that proposed in chapter 4) with a constraint-based model.

Simply put, the argument is this: if two different types of decisions are dominated by different statistical measures, then a model in which these two decisions are separated into different modules is more constraining and 'simpler'[22] than an

[22] Definitions of 'simplicity' vary. For instance, MacDonald *et al.* (1994) argue that their interactive model is architecturally simpler. We argue in terms of informational or computational simplicity – the number of different pieces of information (or 'parameters') that are involved in making a

interactive version. Such a model is more predictive.

To expand this argument, consider a modular and an interactive (constraint-based) statistical model. In the former case, each module has access only to representations that are relevant to its task – those that are stated in its representational vocabulary. This architectural restriction guarantees that only a limited subset of possible statistical measures may be dominant in the decision process of a module, and *all others must be redundant.* In practise, an architectural definition of 'relevance' is imposed on the model.

In contrast, the single decision process in an interactive model must have access to all levels of linguistic representation. A truly interactive model, one which also uses non-linguistic information such as world knowledge or visual context (see section 3.5.3), must also have access to appropriate non-linguistic representations. Statistics may be collated both within and across any of these representations, and there is *no principled way for a researcher to decide which dominate, and which are redundant.*

Suppose, for example, that the decision making process collates statistics including wall colour. Experiments often take place in rooms with white walls. It could turn out that the results would be completely different if the experiments were carried out in rooms with dark walls! This is not a serious example – but the point behind it is. If we presuppose an interactive architecture then the possible influences on the decision process become too numerous, and cannot be controlled for.

Interactivists might argue that this is being silly. They could say it is obvious that the sentence processor uses statistics that appear relevant to the decision in hand. But this obvious, if tacit, assumption of the interactivist approach is simply a variant of modularity, in which particular informational and representational domains are predicted to be coextensive with the domains over which relevant statistical knowledge is accrued.

Having said that, a number of researchers have presented interactivist accounts in which they identify some dominant statistics (MacDonald *et al.*, 1994; Tanenhaus *et al.*, forthcoming). For example, MacDonald *et al.* (1994) explicitly state that certain statistical measures dominate certain ambiguities in their model. Unfortunately, they

decision.

do not state that other statistics do not also dominate or are redundant, for either the ambiguities they consider or alternative ambiguities; in fact, they introduce new statistical measures at a number of points in the paper. Nor are we told the weights assigned to the different measures. Such a model does well when explaining existing data. However, it is impossible to refute – it can make no firm predictions unless the dominant statistics and associated weights are determined. If these predictions turn out to be wrong, it is the weights and statistics that have been refuted, not the model itself. In summary, the single decision process in an interactive model has too many parameters, leading to unpredictiveness. A complex model such as this can avoid refutation by simply changing the weights associated with different parameters.

In contrast, the number of parameters required for each decision within a modular model may be relatively small. Thus the modular position can have the property of being computationally simpler with respect to both the amount of statistical knowledge which is represented, and the amount of experience, or *training*, required to set such parameters. This in turn means that it is possible to obtain systematic predictions from such a model. A modular architecture defines the space of which statistical measures may influence which decisions and, perhaps more interestingly, which may not. Thus the architecture itself is open to falsification, rather than just the particular statistics and weights used by some 'instance' of the architecture. Indeed, from a methodological standpoint, it seems that the only way to satisfactorily prove the interactivist case is to successfully refute the range of more predictive and falsifiable modular models. In section 3.5 we briefly consider some attempts to do this.

3.4.4 Summary

In this section we have presented two rationalist arguments supporting a modular model of the HSPM. The first stems from an interactivist model's inability to avoid the frame problem, and therefore to perform reliably and efficiently. Our experience of the HSPM suggests it is both reliable and extremely efficient, and therefore an interactivist architecture seems unlikely.

Our second argument is methodological. We demonstrate that modular statistical architectures are more predictive and therefore more falsifiable than interactivist accounts. This predictiveness results from their informational simplicity. It follows

that if we must choose between a modular and an interactive account that can explain the same data we ought, by Occam's razor, to choose the modular version. In chapter 4 we introduce a simple modular model that (as we show in later chapters) can account for data previously thought the province of interactive approaches; if we accept the arguments in this section, we should prefer this model.

3.5 Empirical Results

Our arguments in favour of modularity in section 3.4 were rationalist and methodological – a modular account should be preferred in the absence of empirical evidence falsifying it. In this section we consider whether empirical evidence falsifying the notion of a modular parser[23] already exists. It would be impossible and irrelevant to review all the evidence that has been accumulated; instead, we briefly summarise a part of the existing evidence and then go on to consider one of the most striking results in favour of an interactivist account to date.

3.5.1 Evidence Against Syntactic Modularity

We have already considered some empirical evidence that has been taken to disprove syntactic modularity in section 2.3.3. Evidence in favour of Referential Support (Crain & Steedman, 1985; Altmann, 1988; Altmann & Steedman, 1988) is incompatible with a modular syntactic processor in which syntactic processing is not fully parallel. However, this is only the case if it can be shown that Referential Support affects the initial decisions of the processor, rather than the outcome of processing. Clifton and Ferreira (1989) are among a number of authors who have argued that this is not the case and Mitchell, Corley and Garnham (1992) have produced evidence suggesting that initial processing decisions are made independent of Referential Support.

If we just consider modifier attachment, the evidence against a modular model in which syntactic processing is guided by a structural or grammatical heuristic is more compelling. Clifton, Frazier, Rapoport and Radó (submitted) found that semantic and aspectual factors appear to influence the complexity of processing and the initial

[23] We concern ourselves in this section with evidence for and against modularity of the *syntactic* processor, rather than whether the mind is split into (unspecified) modules. Our rationalist arguments concerned the latter position and did not presuppose particular module boundaries; however, such a position is unpredictive and therefore impossible to investigate empirically.

preference in the face of ambiguity for adjunct modifiers (see section 2.3.2). There is also evidence that decisions in the face of prepositional phrase (Taraban & McClelland, 1988; Altmann & Steedman, 1988) and relative clause (Cuetos & Mitchell, 1988) attachment ambiguities are not strictly structurally motivated. This data may result from an interactive HSPM. However, a weakly modular sentence processor in which the syntactic component sometimes outputs multiple or underspecified representations may predict this data (for example, see Frazier & Clifton, 1996; Sturt & Crocker, 1996). A modular statistical model may also account for these apparently pro-interactive results (see section 3.3.4); however, more work is required to isolate the relevant statistics before a firm conclusion can be drawn.

Marslen-Wilson and Tyler (1987) took a different approach. They argued that processes that map onto discourse representation and those that participate in fixation of perceptual belief share many of the properties[24] that Fodor (1983) considers exclusive to modular input systems. Evidence supporting this position raises questions about the privileged status of input modules, and the boundary the modularity hypothesis draws between these modules and central processes. However, it is fully compatible with the slightly weaker notion of modularity we put forward in section 3.2.2 (see Frazier, 1990, for further discussion).

3.5.2 Evidence Supporting Syntactic Modularity

There is an overwhelming amount of evidence supporting particular modular models of human sentence processing. For example, a huge number of experiments have supported the predictions of the garden path theory (see Frazier, 1987a, and citations therein). However, such evidence does not falsify an interactive approach; in general, it is easy to set the weights of an interactive model so that it emulates the initial decisions of a given modular model.

If interactive models are unpredictive (see section 3.4.3), it is difficult to know how to falsify them. However, it seems unlikely that an interactive HSPM would make an incorrect decision when overwhelming evidence for the correct alternative was available. Evidence that the processor ignores potentially useful information during processing would best be explained by a modular model (information is not available due to architectural limitations) than an interactive account (information is ignored

[24] The major exception is information encapsulation.

for no good reason).

Mitchell (1987) presented evidence showing that the HSPM does ignore verb subcategorisation information in making initial attachment decisions. In sentences such as 3.2 he found reading times in the region following the verb that are consistent with initial attachment of the NP as the direct object of the verb even in the (b) form, where the verb is unambiguously intransitive. However, attempts to replicate these results using eye-tracking have failed (Adams, Clifton & Mitchell, submitted). We offer a novel explanation of this data in chapter 7.

(3.2) a. After the child had visited *the doctor* prescribed a course of injections.
b. After the child had sneezed the doctor prescribed a course of injections.

Traxler and Pickering (forthcoming) have also produced evidence that the subcategorisation preferences of some English verbs are ignored when making initial decisions. Ferreira and Clifton (1986) demonstrated that the parser appears to make initial decisions ignoring potentially helpful thematic information when processing reduced relative ambiguities, and also ignores discourse context when processing both prepositional phrase and reduced relative ambiguities.

3.5.3 Eye-Tracking during Comprehension of Spoken Language

Tanenhaus, Spivey-Knowlton, Eberhard and Sedivy (1995) reported the most striking evidence in favour of an interactivist account to date. Subjects were asked to follow a sequence of spoken instructions while wearing mobile eye-tracking equipment; Tanenhaus *et al.* argued that the eye-tracking evidence supported a model in which the subjects' initial interpretation of language was influenced by their visual perception.

This finding suggests that not only is there no modularity within the HSPM, but the HSPM itself cannot be seen as informationally encapsulated. It is therefore worth considering this result in more detail.

While Tanenhaus *et al.* also demonstrated that word recognition is influenced by visual context, it is their evidence about syntactic interpretation that is relevant here. In their experiment, subjects were given instructions about how to manipulate objects placed on a table in front of them; these instructions were given in both syntactically

ambiguous (3.3a) and unambiguous (3.3b) forms:

(3.3) a. Put the apple *on the towel* in the box.
b. Put the apple that's on the towel in the box.

In 3.3a the prepositional phrase "on the towel" is initially ambiguous – it could modify "the apple" (as in 3.3b) or may itself denote the destination. The final PP ("in the box") acts as a disambiguating region; it must be the destination so "on the towel" must modify "the apple".[25] A number of studies have demonstrated that, in the absence of context, the destination reading of the first PP (VP attachment) is initially preferred (Ferreira & Clifton, 1986; Britt, 1994).

Tanenhaus *et al.* presented these sentences in two different visual contexts. For example, each context might contain two towels, one with an apple on it and one without, and a box. The presence of two towels guarantees that both readings of "on the towel" are possible. In the 'one-referent context' there was an alternative object, such as a pencil; in the 'two-referent' context there was a second apple on something else, such as a napkin.

In the unambiguous one-referent context, subjects normally fixated on the apple after it was mentioned, then on the box at the end of the sentence. In contrast, in the ambiguous one-referent case, subjects typically fixated on the apple, but then looked at the empty towel after hearing the ambiguous PP. They then refixated on the apple during the disambiguating region, and finally looked at the box. The fact that they looked at the irrelevant empty towel suggests that they initially misanalysed the ambiguous PP as a destination.

In both the unambiguous and the ambiguous two-referent contexts, subjects typically looked at one of the apples as soon as they heard the word "apple". If they had looked at the incorrect apple then they refixated on the correct apple very quickly on hearing the word "towel". Finally, they looked at the box. Crucially, subjects fixated on the irrelevant towel comparatively infrequently (compared to the one-referent ambiguous context) and there was no significant difference between fixations on this towel in the ambiguous and unambiguous conditions.

[25] Tanenhaus *et al.* ignore the alternative reading in which "on the towel" is the destination and "in the box" modifies "the towel"; we shall also ignore it.

Tanenhaus *et al.* argue that this evidence demonstrates that visual context does have an effect on syntactic processing. Further, they argue that there can be no initial decision, independent of visual context, which is then reanalysed when it proves incompatible with the visual stimulus. This argument is based on the fact that "the time it took participants to establish reference correctly in the two-referent context did not differ for the ambiguous and unambiguous instructions" (Tanenhaus *et al.*, 1995, p.1634); that is, subjects fixated on the correct apple immediately on hearing the word "towel" in both conditions. Tanenhaus *et al.* assume that if "on the towel" was initially incorrectly interpreted as a destination then there would be a delay (while reanalysis occurs) before the correct referent was determined.

It is at this point that Tanenhaus *et al.* show a crucial misinterpretation of the implications of modularity. They assume that reanalysis will only occur at the end of a phrase; we know of no modular model that makes this restriction. A more probable analysis is that the initial decision of the modular syntactic parser is in favour of the destination reading; this decision will be made immediately upon encountering the preposition "on". However, this reading does not pick out a unique referent for the NP "the apple". We may assume that a later central process embodies the principle of Referential Support (see section 2.2.3); this would deem such a reading improbable and may force reanalysis in the parser. Crucially, reanalysis would have occurred before the word "towel" was heard, so we would expect the subject to already have access to the correct interpretation at this point.

A model such as this, in which reanalysis may occur prior to absolute disambiguation, has found some support in the experimental literature (see Mitchell, Corley & Garnham, 1992). It is also compatible with a modular statistical position (see section 3.3.3).

3.5.4 Summary

The empirical evidence that relates to modularity suggests that Fodor's (1983) 'Modularity Hypothesis' is almost certainly wrong. There are good reasons to suppose that the 'central processes' are more akin to Fodor's 'input systems' than he suggests; some semantic processes may act as modules but not be fully encapsulated (Marslen-Wilson & Tyler, 1987; Frazier, 1990). Such evidence points to a blurring of the distinction that Fodor draws between syntactic input systems and semantic

central processes.

There is also evidence that syntactic decisions are not based only on "the acoustics of the input and the grammar" (Fodor, 1987, p.28). While evidence suggesting that Referential Support affects initial complement/relative decisions is controversial, there is growing evidence that *something* else affects at least modifier attachment decisions. It is not yet clear whether these results point to a model in which the parser has access to statistical information, a weakly modular model in which some decisions are left for later processes, or a fully interactive model.

However, while the letter of Fodor's hypothesis is unlikely to prove correct, the spirit of his proposal is compatible with the data. A modular position such as that espoused in section 3.1.3 is fully compatible with the existing data. In fact, the most striking evidence to date taken to support an interactive position (Tanenhaus, Spivey-Knowlton, Eberhard & Sedivy; 1995) is consistent with the behaviour we might expect from such a model.

While there is ample evidence supporting particular modular architectures, evidence against an interactivist approach is hard to come by. This is largely due to the unpredictiveness of interactivist models (see section 3.4.3). However, there is some data that suggests that the HSPM initially ignores information that would be helpful in making a decision; this is predicted by modular architectures, but within an interactivist approach we would expect the HSPM to make sensible use of all available information. Such data is only compatible with an interactive model if the weights associated with particular information types are surprisingly low; an interactive model can therefore *account for* any data, but fails to *explain* instances where certain post-syntactic information is systematically ignored or delayed.

3.6 Conclusions

A modular model is composed of independently predictive processes that use their own representational vocabulary (and ignore the vocabulary of other modules); such a model must also have low bandwidth feedforward and (particularly) feedback connections.

We suggested this definition in section 3.1. It is important to understand that despite

the historical trend (noted by Spivey-Knowlton & Eberhard, 1996) for such models to be binary, there is no reason why a modular model should not be statistical. In section 3.2 we introduced the Modular Statistical Hypothesis (MSH) and considered what distinctive characteristics a modular statistical model might have. There are four particularly important characteristics, which we reiterate here:

- Architecturally defined relevance: in a modular model, only certain statistical measures may dominate a decision, and all others must be redundant. This contrasts with an interactive approach in which there is no principled way of determining which statistical measures are relevant to a particular ambiguity decision.

- Strategic use of statistics: since statistics are not architecturally required by such a model, we might expect their use to be limited to decisions in which they offer significant benefits. Not all processes in a modular statistical model need be statistical.

- Early reanalysis: such a model has the capability to require reanalysis before absolute disambiguation.

- Back door semantics: even though it is fully encapsulated, such a model may show apparently semantic effects in the initial decisions of the syntactic parser. However, there are strict limitations on the sort of effects that may occur.

We argued that such a model should be preferred to an interactive approach both in terms of speed (following Fodor's argument) and methodologically; a modular model is informationally simpler and more predictive. Finally, we considered whether existing empirical data can convincingly argue for an interactive or modular position; we found evidence both ways, and conclude that both positions are still empirically tenable.

In the next chapter we consider the problem of lexical category ambiguity. We propose a model of sentence processing in which lexical category decisions play a distinguished role; this model is compatible with the MSH. In later chapters we explore the predictions of the model and thereby present evidence supporting both

the privileged role of (statistical) lexical category decisions and the MSH.

4: Lexical Category Disambiguation

> Any theory of syntactic processing would assume that the syntactic classes of individual words are identified (though probably not consciously) as a first step in arriving at an overall structure of a sentence.
>
> (Rayner & Pollatsek, 1989, p.243)

4.1 Introduction

In the previous chapter we introduced the Modular Statistical Hypothesis (MSH). This concerns modular models of human sentence processing in which (at least) some decisions are made on the basis of a statistical heuristic. In this chapter we consider one of those decisions; that of lexical category disambiguation. In section 4.2 we introduce the phenomenon of lexical category ambiguity and argue for its privileged status within the HSPM.

In section 4.3 we consider whether the syntactic models we discussed in chapter 2 could account for lexical category decisions. In the majority of cases, the authors who proposed these models made no specific claims about lexical category ambiguities; we are therefore extending the scope of these models beyond the authors' explicit intentions in order to undertake this analysis. However, the authors of two of these models (Pritchett, 1992; MacDonald *et al.*, 1994) have explicitly contended that their proposals account for (at least) some lexical category ambiguities. Frazier and Rayner (1987) suggested an extension to the Garden Path theory in which lexical category ambiguity decisions are deferred; we consider this model and the evidence that supports it in section 4.4.

Section 4.5 introduces the model that is the central proposition of this thesis. In this model, lexical category decisions have a privileged status, distinct from syntactic processing. There is a separate module concerned with making these decisions; we call this the Statistical Lexical Category Module (SLCM). As its name implies, the SLCM makes use of a frequency-based heuristic. Any model that includes this

module therefore falls within the MSH and so evidence for the existence of this module can be seen as evidence supporting the MSH. We consider whether existing and novel evidence supports such a model throughout the remainder of this thesis.

4.2 Lexical Category Ambiguity

Lexical category ambiguity occurs when a word can be assigned more than one grammatical class (noun, verb, adjective etc.). These classes are also known as lexical categories.[26] Consider, for example, the sentences in 4.1:

(4.1) a. Time *flies like an arrow.*
b. He saw *her duck.*

Both of these sentences are ambiguous, and in each case the different readings arise from lexical category ambiguity. In sentence 4.1a, "flies" is ambiguous between noun and verb readings and "like" is ambiguous between preposition and verb readings.[27] Syntactically valid readings can be constructed if "flies" is a verb and "like" is a preposition (in which case the meaning is as in 4.2a), and if "flies" is a noun and "like" is a verb (cf. 4.2b); however, only the former reading is semantically plausible.

(4.2) a. Time flies by like an arrow.
b. House flies like an apple.

In contrast, both readings of 4.1b are plausible. In one reading, "her" is a possessive pronoun and "duck" is a noun (cf. 4.3a); in the other reading, "her" is a personal pronoun and "duck" is a verb (cf. 4.3b).

(4.3) a. He saw her apple.
b. He saw her leave.

4.2.1 Lexical Category Ambiguity and Lexical Access

In section 2.2 we briefly mentioned lexical access – the stage of processing at which lexical entries for input words are retrieved. The evidence suggests that all possible

[26] We use the term (grammatical) class and lexical category interchangeably.

[27] "Like" may also be an adjective, an adverb, a conjunction or a noun, but we ignore these alternatives in this example.

meanings for a given word are retrieved even when semantic context biases in favour of a single meaning (Swinney, 1979; Seidenberg *et al.*, 1982). However, it seems plausible that grammatical classes may be resolved during lexical access; that is, if the *syntactic* context favours one reading of a word, then only that reading is retrieved.

The evidence does not support the determination of grammatical class during lexical access. Tanenhaus, Leiman and Seidenberg (1979) found that when subjects heard sentences such as those in 4.4, containing a locally ambiguous word in an unambiguous syntactic context, they were able to name a target word which was semantically related to either of the possible meanings of the ambiguous target (e.g. SLEEP or WHEEL) faster than they were able to name an unrelated target. This suggests that words related to both meanings had been 'primed'; both meanings must therefore have been accessed, despite the fact that only one was compatible with the syntactic context.[28]

(4.4) a. John began to tire.
b. John lost the tire.

Seidenberg, Tanenhaus, Leiman and Bienkowski (1982) replicated these results, and Tanenhaus and Donnenworth-Nolan (1984) demonstrated that they could not be attributed to the ambiguity (when spoken) of the word "to" or to subjects' inability to integrate syntactic information fast enough prior to hearing the ambiguous word.

Interestingly, Tanenhaus *et al.* (1979) found that, while words related to all meanings of an ambiguous word were primed immediately following that word, a gap of 200ms or more between the word and the target resulted in priming of only the contextually appropriate meaning. This suggests that while all classes and meanings of a word are initially accessed, the ambiguity is very quickly resolved.

The evidence we have considered favours a model in which lexical category disambiguation occurs after lexical access. The tacit assumption in much of the

[28] Note that Tanenhaus *et al.'s* methodology meant that only homonyms (words with a number of unrelated meanings) were investigated. It is therefore possible that categorially ambiguous polysems (words with a number of related meanings) are not subject to parallel access. As there does not appear to be any good reason to prefer this more complex alternative, we assume that categorially ambiguous homonyms and polysems are both accessed in parallel.

sentence processing literature has been that grammatical classes are determined during parsing (see, for example, Pritchett, 1992). If grammar terminals are words rather than lexical categories, then such a model requires no augmentation of the parsing mechanism. We consider such models in section 4.3.

However, there are alternative possibilities. Frazier and Rayner (1987) proposed that lexical category disambiguation has a privileged status within the parser; different mechanisms are used to arbitrate such ambiguities from those concerned with structure building. We consider Frazier and Rayner's proposal in more detail in section 4.4.

Finally, lexical categories may be determined after lexical access, but *prior to* syntactic analysis. That is, lexical category disambiguation may constitute a module in its own right. This is the position we propose in this thesis. In the remainder of this section, we argue why such a position is worth further investigation.

4.2.2 Examples and Types

The sentences in example 4.1 exhibited lexical category ambiguity. This resulted in global ambiguity at the syntactic level; however, such examples are unusual. It is far more common for lexical category ambiguities to be disambiguated by their syntactic context (as in 4.2a and 4.4). If the disambiguating context occurs *after* the ambiguity or there is no disambiguating context, then there is ambiguity on both the lexical category and syntactic levels; two (or more) lexical category decisions lead to different (temporarily) valid syntactic structures. We will call such examples 'lexical-syntactic' category (LSC) ambiguities. The sentences in 4.5 to 4.9 exemplify this type of lexical category ambiguity.

(4.5) a. The *old train* whistled through the station.
b. The *old train* the young.

(4.6) a. Bill told Sarah *that* the man was poor.
b. Bill told Sarah *that* man was poor.

(4.7) a. The army *bases* are overcrowded.
b. The army *bases* their decisions on long reports.

(4.8) a. The candidate went *to* place his vote.
b. The candidate went *to* Westminster.

(4.9) a. Without *her friends* are hard to find.
b. Without *her friends* the wedding would have been a disaster.

Alternatively, disambiguating syntactic context may occur *before* the ambiguity. In this case, there is no ambiguity at the syntactic level; we will call such examples 'non-syntactic' category (NSC) ambiguities. Examples of NSC ambiguities are given in 4.10 to 4.12.[29]

(4.10) a. The *train* whistled through the station.
b. The boys *train* the dogs.[30]
(4.11) a. The army's *bases* are overcrowded.
b. The armies *base* their decisions on long reports.
(4.12) The candidate wanted *to* place his vote.

Any model in which initial decisions in the face of lexical category ambiguities are made with reference to all relevant syntactic information will predict that NSC ambiguities will not cause the processor to garden path; in these models such sentences are simply not ambiguous. All existing work therefore concentrates on LSC ambiguities. In chapter 7, we show that some existing evidence regarding the late availability of subcategorisation information can be reinterpreted as an NSC ambiguity. In chapter 8, we present novel experiments that test whether sentences that contain NSC ambiguities do cause processing difficulty.

4.2.3 Frequency of Occurrence

We can make a very rough estimate of the frequency of lexical category ambiguity by determining how many words occur with more than one category in a large text corpus. DeRose (1988) has produced such an estimate from the Brown corpus; he found that 11.5% of word types and 40% of tokens occur with more than one lexical category. A full breakdown of his results for word types is given in table 4.1. As the mean length of the sentences in the Brown corpus is 19.4 words, DeRose's figures suggest that there are 7.75 categorially ambiguous words in an average corpus

[29] In these examples, we italicise the ambiguous word even though there is no *syntactic* ambiguity.

[30] Pritchett (1992) presents sentences such as those in 4.10b as containing temporary ambiguity, suggesting that readers may consider the Saxon genitive form (cf. "the boy's train was black"). We do not concur with his analysis; however, a noun compound reading may be syntactically permissable in this example (see the discussion in section 6.2.2).

sentence.

Number of tags	Number of word types
1 (unambiguous)	35340
2	3760
3	264
4	61
5	12
6	2
7	1

Table 4.1: Degrees of category ambiguity (from DeRose, 1988)

Our own study leads to slightly different results. Using the Treebank version of the Brown corpus, we discovered 10.9% ambiguity by type, and a staggering 65.8% by token. To obtain these results, we used the coarsest definition of lexical category possible – just the first letter of the corpus tag. As DeRose (1988) gives very little detail about how his results were obtained, it is difficult to guess why our results differ. It could be due to the fact that the corpus was retagged for Treebank, or because we omitted punctuation from our count. However, both results suggest that lexical category ambiguity is extremely frequent in normal English text.

4.2.4 The Privileged Status of Lexical Category Ambiguity

As discussed in section 4.2.1, lexical category ambiguities may either be considered syntactic, or may be viewed as a distinct processing problem. In the first case, terminals in the grammar are words and it is the job of the parser to determine the lexical category that dominates each word. This is the model that has been tacitly assumed by many researchers; in section 4.3 we consider the predictions existing parsing models would make when extended to arbitrate lexical category ambiguities.

If we take the latter view of lexical category ambiguities, one possibility is that a pre-syntactic modular process makes lexical category decisions. These decisions would have to be made on the basis of a simple heuristic, without the benefit of syntactic constraints. In common with all modules, such a process is irrational; it will make incorrect decisions when available information should force the correct decision

(NSC ambiguities).[31] It does, however, offer an extremely low cost alternative to syntactic arbitration; as we shall see in section 4.5, a statistical model of lexical category disambiguation is both computationally simple and extremely accurate.

Given the high frequency of lexical category ambiguity (see section 4.2.3), a separate decision making process makes computational sense. As much ambiguity is resolved prior to parsing, the job of the parser is significantly simplified. As we shall see in chapters 6 and 7, a number of common 'syntactic' ambiguities can be recast as LSC ambiguities, particularly if we adopt a more fine-grained definition of lexical category.

There are a number of qualitative differences between lexical category and syntactic ambiguities.[32] These also lead to the conclusion that these two types of ambiguity should be considered distinct:

- Lexical category ambiguities tend to be disambiguated locally, normally within the phrase in which they occur. In contrast, disambiguation in the case of syntactic ambiguities regularly spans phrasal nodes.
- Immediate lexical context is extremely relevant in determining the most probable lexical category for a word. In the case of syntactic preference, lexical context is often not highly predictive.
- In a serial model of syntactic structure building, the parser is unaware of alternative syntactic analysis. However, as we reported in section 4.2.1, studies of lexical access (Seidenberg *et al.*, 1982; Tanenhaus & Donnenworth-Nolan, 1984) demonstrate the simultaneous availability of different lexical categories. We may therefore expect lexical category decisions to be guided by a comparative, rather than 'blind', heuristic.
- Word boundaries may be identified prior to lexical categories being assigned, and a lexical category normally spans a single word. In contrast, determining what part of a sentence is dominated by a syntactic node is dependent on the node being assigned. Lexical categories immediately dominate words in a non-branching manner, whereas syntactic nodes may

[31] We investigate this claim experimentally in chapter 8.

[32] See MacDonald, Pearlmutter and Seidenberg(1994), for a contrasting view

immediately dominate a number of other nodes or lexical items.

In summary, while the status of lexical category ambiguity is still open to debate, there are sufficient qualitative differences between it and syntactic ambiguity to motivate the proposal of a distinct lexical category disambiguation module. The computational benefits of such an approach also support further investigation. In section 4.5 we introduce our Statistical Lexical Category Module (SLCM).

4.2.5 Summary

Lexical category ambiguity is one of the most frequent forms of ambiguity in language. DeRose (1988) estimated that 40% of word tokens in the Brown corpus are lexically ambiguous; our own study produced a figure of 65.8%. The issue of how lexical category decisions are made is therefore not a small footnote in the syntactic ambiguity literature; it is a problem that deserves independent study.

Experimental evidence suggests that categorially ambiguous words are retrieved from the lexicon in parallel, even in the face of strong syntactic bias. Lexical category decisions must therefore occur after lexical access; either as part of parsing or as a distinct pre-syntactic process. The latter position has computational benefits and is supported by qualitative differences between lexical category and syntactic ambiguity. It is the position that we take in this thesis.

Before introducing a model in which lexical category ambiguity has privileged status, we consider the behaviour of existing syntactic models if lexical category ambiguity was added to their remit.

4.3 Syntactic Models and Lexical Category Decisions

In chapter 2, we recounted a range of models of human parsing. As discussed in section 4.2.4, such a model may be sufficient to account for lexical category ambiguity. Very few authors have explicitly argued that their proposed models *do* account for such ambiguity (though see MacDonald *et al.*, 1994 and Gibson, 1991), but at least one (Pritchett, 1992) has presented evidence about lexical category decisions in support of his model.

In this section we extend some of the models outlined in chapter 2 to cover lexical

category ambiguity, and consider what sort of behaviour they would display. In section 4.4 we examine the only explicit modular model of lexical category ambiguity that has been proposed in the sentence processing literature.

4.3.1 Non-Statistical Heuristic Models

The Garden Path Theory

The Garden Path theory is very simple, and can therefore be highly predictive. However, its predictions depend strongly on details of syntactic formalism. This can be exemplified by considering 4.14, sentences used in Frazier and Rayner's (1987) study (see section 4.4).

(4.14) a. I know that the desert *trains* young people to be especially tough.
b. I know that the desert *trains* are especially tough on young people.

In this example, "trains" may either be a verb (4.14a) or a noun (4.14b). In the former case, "desert" is a noun. Frazier and Rayner consider "desert" to be an adjective in the latter case, basing their argument on the fact that English requires compound nouns to have stress on the left-hand member, but such stress is not obligatory on phrases such as "desert trains". The noun interpretation of "desert" requires the construction of fewer new nodes; MA therefore predicts that this analysis will initially be preferred. As, according to Frazier and Rayner, this reading is only congruent with the verb analysis of "trains", they predict an initial decision favouring the verb reading of the ambiguous word.

However, MacDonald (1993) assumes that the correct reading for 4.14b is as a noun compound – both "desert" and "trains" are nouns. We concur with this linguistic analysis, and base our own discussion of existing and novel experimental results on it (see section 6.2 and chapter 8). The argument in favour of such an analysis also depends on sentence stress; we argue that it is *normal* to assign a compound noun stress pattern to "desert trains", and this is therefore the preferred analysis. Consider the contrasting pair of sentences in 4.15 (both adapted from MacDonald's, 1993, experimental materials); in 4.15a the normal stress pattern suggests a noun compound reading whereas in 4.15b an adjective–noun reading is indicated.

(4.15) a. The computer *programs* are slower than anticipated.

b. The official *documents* are in the post.

If we accept this noun compound analysis for sentence 4.14b, then the Garden Path theory predicts an initial commitment to the noun analysis of "trains"; this analysis is supported by both MA and LC.

This dependence on syntactic structure means that, in common with many other models, pinning down solid predictions for particular sentences based on the Garden Path model can be difficult. However, we can make a couple of observations about the general behaviour of the model when faced with lexical category ambiguities:

- In any two cases where there is ambiguity between the same lexical categories, and in which the structural context is identical, the HSPM will make the same initial decision, regardless of the frequency with which the different categories occur.
- NSC ambiguities (recall section 4.2.2) will never cause any processing difficulty.

Construal

Lexical category ambiguities cross the boundaries laid down by the Construal hypothesis (Frazier & Clifton, 1996); while some may affect whether a phrase is primary or not, or affect the attachment of a primary phrase, others make no difference to primary phrase attachment. Consider the examples in 4.16, 4.17 and 4.18:

(4.16) a. I discovered *that young* dog under the table.
b. I discovered *that young* dogs were under the table.

(4.17) a. The adolescent *rages* will soon pass.
b. The adolescent *rages* at his parents.

(4.18) a. Bill walked a mile *to* the shops.
b. Bill walked a mile *to* cool off.

In 4.16, "that" is ambiguous between determiner and sentence complement readings. Both these readings involve primary phrase attachment; by MA, the determiner reading should be preferred as it involves the construction of fewer new nodes. In 4.17, "rages" is ambiguous between noun and verb readings. Choosing the verb

reading instantiates the primary subject–verb relation; the noun reading instantiates no primary relation. The verb reading should therefore be preferred. Finally, in 4.18, "to" is ambiguous between preposition and infinitival readings. In neither case can the resultant phrase be primary. The decision is therefore made on the basis of "structural and non-structural principles" (Frazier & Clifton, 1996, p.41).

While the predictions of Construal vary across different lexical category ambiguities, every ambiguity we consider in chapters 6, 7 and 8 involves at least one possible primary phrase attachment. In these cases, the behaviour patterns of Construal and the Garden Path theory are similar; we therefore pay very little attention to Construal for the remainder of this thesis.

Generalised Theta Attachment

The predictions of Generalised Theta Attachment (Pritchett, 1992) depend largely on the maximal theta grid of individual lexical items. This allows for some variation in initial decisions between lexical category ambiguities in syntactically similar contexts, and may account for results that appear to be due to statistical mechanisms. For example, consider 4.19 and 4.20:

(4.19) a. The baby *sneezes* all night.
b. The baby *sneezes* kept me awake all night.

(4.20) a. The baby *saw* its parents for the first time.
b. The baby *saw* was not big enough to cut the wood.

In all cases, the italicised word is ambiguous between noun and verb readings. In 4.19, choosing the verb reading allows all verb theta roles to be assigned, and every NP to have a role. The θ-criterion is therefore maximally satisfied. In contrast, if the noun reading of "sneeze" is chosen, then "sneeze" itself does not receive a theta role and the θ-criterion is not maximally satisfied. In this case, Generalised Theta Attachment predicts that the verb reading is initially preferred.

While the verb "sneeze" can only be intransitive (though see section 7.4), the verb "saw" may be used transitively. Its maximal theta grid therefore contains an experiencer *and* a theme role. Choosing the verb reading in 4.20 leaves the theme role initially unassigned, whereas choosing the noun reading leaves one NP without a role. Pritchett (1992) does not state which transgression of the θ-criterion is to be

preferred; Generalised Theta Attachment is therefore unpredictive in this case.

Unfortunately, much of our data deals with exactly this ambiguity (see chapters 6 and 8), and the majority of the verbs used may be transitive. While Generalised Theta Attachment is not contradicted by this data, it does not predict any frequency-based variation. As we shall see, Generalised Theta Attachment (alone) is not sufficient to explain the observed behaviour of the HSPM.

Referential Support

When discussing the Referential Support model (Crain & Steedman, 1985; Altmann & Steedman, 1988) in section 2.3.3, we noted that the theory is concerned with the interpretation of sentences presented in context. However, the only experimental data we consider in this thesis occurs in the null context. According to the model, in this case ambiguity resolution will be arbitrated by the Principle of Parsimony. Consider example 4.21.

(4.21) a. The cheese *spreads* are inedible.
b. The cheese *spreads* straight from the fridge.

It is not clear which reading of the ambiguous material should be preferred by parsimony. Both introduce one novel discourse entity; the correct reading for 4.21a introduces "cheese spreads" whereas 4.21b introduces "cheese". Importantly, the former reading does not also necessitate a discourse entity representing a particular cheese; in fact, a cheese spread need not even contain cheese (cf.. "hedgehog crisps"). Parsimony therefore makes no clear prediction about which reading would be initially preferred.

The Referential Support model is similarly unpredictive for all the ambiguities explored in this thesis. It is therefore consistent with the data explained in this thesis, but not explanatory. Alone, Referential Support cannot explain the observed behaviour of the HSPM.

However, it would be a mistake to assume that Referential Support is therefore irrelevant to lexical category ambiguity resolution. If subjects were presented with sentences in context, it could transpire that this context has a strong effect on their initial decisions in the face of lexical category ambiguity. The fact that we consider

no such experimental results simply means that we have no evidence that discourse context does not affect lexical category decisions. To test this hypothesis is a matter for future research.

4.3.2 The Tuning Hypothesis

The most coarse-grained variant of the Tuning Hypothesis (Mitchell & Cuetos, 1991) makes sweeping predictions for lexical category ambiguities, similar to those of the Garden Path model. If statistical decision mechanisms only affect, for example, the formation and attachment of phrasal nodes, then we would expect that similar decisions would always be made in the same syntactic context, regardless of the frequency bias of individual words. As a model for determining lexical category ambiguity, coarse-grained Tuning stands or falls with the Garden Path theory.

In contrast, more fine-grained variants of Tuning may predict that initial decisions in the face of lexical category ambiguity are determined by preferences associated with individual words. In this case, it will be difficult to differentiate the predictions of Tuning and our proposed statistical module (see section 4.5). In particular, as we argue in section 7.5, the behaviour of our model may, in fact, approximate that of a statistical parser.

In order to determine whether lexical category disambiguation does occur before syntactic parsing, we must examine human behaviour when processing constructions in which lexical co-occurrence preferences point to one decision, but syntactic preferences point another way. Examples may occur when a word with two or more possible lexical categories occurs immediately following the realisation of a syntactic gap; for instance, the word "trains" in example 4.22:

(4.22) The man that John knows trains pigeons.

Here, the syntactic environment may favour a verb reading of "trains", as a putative direct object for the verb "knows" is already available; however, a presyntactic lexical category decision model would be oblivious to the presence of a gap and may determine that the noun reading is more probable, given the particular lexical context. Unfortunately the postulation of a syntactic gap may cause processing delay (see, for example, Stowe, 1986) and so evaluating experimental results would be

difficult. In chapter 8 we present novel experimental evidence concerning a different construction in which syntactic and lexical preferences differ.

4.3.3 Interactive Models

Interactive models of sentence processing tend to make use of fine-grained lexical data (see, for example, MacDonald *et al.*, 1994). They are therefore ideally situated to account for lexical category ambiguity data. Much of the evidence we consider in chapters 6 and 7 originates from proponents of interactive models.

However, as we argued in chapter 3, there are problems with such models. In particular, they tend to be unpredictive. We therefore do not argue against these models in terms of their power to account for the data. Instead, we demonstrate that our proposed model, which is far simpler and more predictive, offers explanations for a spectrum of ambiguities previously considered as evidence for an interactive approach. While we consider how interactive models might account for the data we present, and highlight where such accounts appear contrived, our argument against them is ultimately reliant on Occam's razor.

4.4 The Delay Strategy

In section 4.3 we considered models in which lexical category ambiguity is not distinguished from other types of syntactic ambiguity. In this section, we review the only model that has been proposed in the literature in which lexical category ambiguity does have a privileged status.

4.4.1 Evidence Supporting the Delay Strategy

Frazier and Rayner (1987) studied lexical category ambiguity using sentences similar to those in 4.14. The full set of conditions is exemplified in 4.23.

(4.23) a. I know that the desert *trains* young people to be especially tough.
b. I know that the desert *trains* are especially tough on young people.
c. I know that this desert trains young people to be especially tough.
d. I know that these desert trains are especially tough on young people.

The (a) and (b) forms are temporarily syntactically ambiguous (LSC ambiguities); 4.23a is disambiguated towards the verb reading of trains, whereas 4.23b is

disambiguated to the noun reading. As discussed in section 4.3.1, Frazier and Rayner consider that the correct reading of 4.23b involves categorizing "desert" as a derivative adjective; if the Garden Path theory were extended to include lexical category ambiguities, they would predict that the verb analysis of "trains" would be initially preferred in both the (a) and (b) forms. Our preferred noun compound analysis of (b) would lead to an MA prediction for both sentences in which a noun reading of "trains" is initially preferred.

The (c) and (d) forms of 4.23 are NSC ambiguities – Frazier and Rayner therefore consider them essentially unambiguous. In (c), the violation of number agreement between the determiner and "trains" forces the verb reading. In contrast, the verb reading in (d) is ruled out by the disagreement between the determiner and "desert". In the following discussion, we refer to the ambiguous word ("train") and the word preceding it as the 'critical region'. We refer to the remainder of the sentence as the disambiguating region, even in the unambiguous conditions.

Frazier and Rayner identify two alternatives to the Garden Path theory that might apply to lexical category ambiguities. Firstly, the processor might construct parallel syntactic analyses until the ambiguity is resolved. If this hypothesis was correct, we would expect an increased reading time for the critical region of the ambiguous sentences (compared to the unambiguous ones), as more than one syntactic analysis must be constructed and maintained. The second option is that the HSPM suspends syntactic processing until disambiguating material becomes available; in this case we would expect reduced complexity and therefore processing time while reading the ambiguous critical regions; Frazier and Rayner call this option the "delay strategy". In both these cases increased reading time would be expected in the disambiguating region for both ambiguous materials – as the parser would have to discard one or more analyses (in the parallel model) or construct a delayed analysis. In contrast, the Garden Path theory predicts no reading time increase in the critical region, and processing delay in only one ambiguous condition in the disambiguating region.

Frazier and Rayner performed two eye-tracking experiments on these materials. In experiment 1, the meaning of the critical words was not systematically related; in experiment 2, it was. In both experiments, they found a reduced reading time in the critical region of the ambiguous sentences compared to the unambiguous items.

They also discovered an increased reading time in the disambiguating region of both ambiguous conditions compared to the unambiguous versions.

So far, these results support the delay strategy. It seems that neither the Garden Path theory nor a parallel processing model can account for the reduced reading time in the ambiguous region; the Garden Path theory also suggests no explanation of the increased reading time in the disambiguating region of *both* conditions. However, Frazier and Rayner's experiment 1 did produce one result that is apparently consistent with the Garden Path theory: reading times in the disambiguating region of the noun items (b) were greater than those for the verb items (a). Similarly, but in contrast to the Garden Path predictions, strikingly long fixations occurred on the first word of the critical region in the unambiguous noun condition (d). As these processing delays were not found in experiment 2, Frazier and Rayner attribute them to "the need to construct a salient semantic relation between the adjective and noun without the benefit of thematic constraints" (Frazier & Rayner, 1987, p.514). They therefore conclude that their results support the delay strategy; the architecture of the sentence processor accords with the Garden Path theory, but lexical category ambiguities are accorded special treatment: syntactic processing is suspended until disambiguation.

4.4.2 Reasons to Doubt

Unfortunately, this proposal has both theoretical and empirical shortcomings. From a theoretical perspective, it is unclear how the sentence processor might determine when a lexical category ambiguity has been disambiguated *unless it continues to build syntactic structures*. Disambiguation occurs when one of the alternate analyses violates a syntactic constraint; if syntactic processing is suspended then the HSPM will remain ignorant of this violation. In sentence 4.23a, "young" is categorially ambiguous; if syntactic processing is suspended then the HSPM will be unaware that, given that "people" is a noun, all grammatically licit structures spanning the entire sentence involve assigning the grammatical class 'adjective' to "young".[33]

MacDonald (1993) suggests an alternative analysis of Frazier and Rayner's (1987) empirical results, which underpin the delay strategy. MacDonald points out that "this" and "these", used to create Frazier and Rayner's unambiguous conditions,

[33] For the purposes of exposition, we ignore the reading where "people" introduces a reduced relative.

have a deictic function and appear awkward in sentences without prior context. Long reading times immediately following these determiners – on the critical region in the unambiguous condition – may simply reflect subject's confusion about the infelicitous use of these determiners.

MacDonald used materials similar to 4.24 and 4.25 to determine whether the choice of determiners resulted in an artefactual effect in Frazier and Rayner's study. The sentences in 4.24 are analogous to those in 4.23, except that changes in tense are used to manipulate ambiguity; 4.25 is identical to 4.24 except that the determiners "this" and "these" are used in place of "the" before the critical words.

(4.24) a. I know that the desert *trains* could resupply the camp.
b. I know that the desert *trains* soldiers to be tough.
c. I know that the deserted trains could resupply the camp.
d. I know that the desert trained soldiers to be tough.

(4.25) a. I know that these desert trains could resupply the camp.
b. I know that this desert trains soldiers to be tough.
c. I know that these deserted trains could resupply the camp.
d. I know that this desert trained soldiers to be tough.

MacDonald presented these sentences to subjects in a moving window self-paced reading study. Her results for the examples in 4.24 did not support a model incorporating the delay strategy; while she did find greater reading times in the disambiguating region for the ambiguous conditions, she did not find increased reading times in the critical region for the unambiguous noun condition (c) compared to ambiguous version (a). However, the pattern of reading times MacDonald obtained for 4.25 did resemble Frazier and Rayner's results. Crucially, the reading times for the critical regions in all sentences in 4.25 were significantly longer than those for the analogous condition in 4.24, suggesting that the infelicitous use of the determiners "this" and "these" did have an artefactual effect on Frazier and Rayner's results.

4.4.3 Summary

The delay strategy is an interesting proposal, in that it is the only previous attempt to consider lexical category ambiguity as a phenomenon in its own right. However, the

empirical evidence supporting a delay model appears to rely on an artefact of the experimental design. In chapter 6 we consider further results from MacDonald's (1993) investigation of lexical category ambiguity and suggest that her conclusions may also be unwarranted. In chapter 8 we present novel experimental evidence that is also inconsistent with the delay strategy.

4.5 A Statistical Model of Lexical Category Disambiguation

This section introduces our own model. The model differs from those considered so far in that lexical category disambiguation is postulated as a distinct modular process, which occurs prior to syntactic processing but following lexical access. We call this the Statistical Lexical Category Module (SLCM).

4.5.1 Why Statistical?

In section 4.2.4 we argued that there are good reasons to consider lexical category disambiguation as a process distinct from syntactic processing. However, the model we are proposing is not only modular, it is also statistical. Why should a pre-syntactic lexical category module be statistical?

Many of our reasons are given in section 3.3, where we introduced the MSH. We support a model of human sentence processing that is (at least partially) statistical on both rational and empirical grounds: such a model appears sensible and has characteristics which may explain some of the behaviour patterns of the HSPM. It is the purpose of this thesis to provide evidence supporting this position, as well as to propose a concrete theory of lexical category disambiguation. We take the MSH as a starting point and we therefore propose that some processing modules are statistical.

If we are to investigate a partially statistical modular HSPM, it makes sense to begin with modules that occur early on in the processing chain, such as the SLCM. There are two reasons for this. Firstly, the initial decisions made by earlier modules may affect those of later modules, but not visa versa; without discovering the behaviour of earlier processes, it is difficult to predict the behaviour of later processes with respect to a particular input to the language processor. Secondly, mathematical models underlying such processes as lexical category disambiguation and parsing have been widely explored (see section 4.5.3). In contrast, while there is no a priori reason

why, for example, semantic integration may not rely on statistical processes, there has been very little exploration of what these processes might be.

4.5.2 What Statistics?

If we accept that the SLCM is statistical, the next question must be what statistics dominate its decisions. Limitations of the modular architecture we are proposing constrain the choice. The SLCM has no access to structural representations; structurally motivated statistics could therefore not be expressed in its representational vocabulary. We will assume that the input to the module is extremely shallow – just a word and a set of candidate grammatical classes. In this case, the module also has no access to low level representations including morphs, phonemes and graphic symbols; the module may only make use of statistics collated over words or lexical categories, or combinations of the two.

It seems likely that the SLCM collates statistics concerning the frequency of co-occurrence of individual words and lexical categories. One possible model is therefore that the SLCM simply picks the most frequent class for each word; for reasons that will become apparent, we will call this the 'unigram' approach.

The SLCM may also gather statistical information concerning context. For example, decisions about the most probable lexical category for a word may be dependent on the previous word. Alternatively, such decisions may only depend on the category assigned to the previous word, or both word and category may be used. For reasons that will become clear in section 4.5.3, we will call these the 'word bigram', 'tag bigram' and 'combined bigram' approach respectively.

In section 4.5.4 we consider which of these models should be preferred. In the next section we approach the same problem using probability theory, and introduce the notion of a part-of-speech 'tagger'.

4.5.3 Probability Theory and the Tagger

The problem faced by the SLCM is to assign the most likely set of lexical categories to a given sequence of words. This task must be performed incrementally. Research in Natural Language Processing (NLP) has concentrated on a (non-incremental) version of this problem and a number of successful and accurate 'part-of-speech

taggers' have been built (e.g. Garside, 1987; Cutting *et al.*, 1992).

A number of different approaches to the problem have been suggested by different authors. Klein and Simmons (1963) and Greene and Rubin (1971, cited in DeRose, 1988) both made use of large dictionaries, morphological and special case rules and hand-written context frame rules. More recently, Brill (1992) has also produced a rule-based tagger; however, Brill's tagger learns context and morphological rules automatically using information derived from a training corpus and achieves much greater accuracy than the earlier versions.

However, the majority of modern taggers use statistical information about language. The CLAWS tagger (Garside, 1987), used to assign part-of-speech 'tags' to the British National Corpus (see section 5.2), makes use of a number of different information sources including tag co-occurrence probabilities and very limited word–tag frequency information. In other taggers (DeRose, 1988; Church, 1988; Cutting *et al.*, 1992), statistical information plays a primary role. It is this last set of taggers that is most suitable as a model of Human Sentence Processing; they provide a straight-forward learning algorithm based on prior experience, are comparatively simple, do not make use of arbitrary or *ad hoc* rules, and can be used to assign preferred lexical category tags incrementally.

In section 4.5.5 we introduce the notion of a Hidden Markov Model (HMM) and show how the job of a tagger can be seen as finding the best path through an HMM. In this section we consider the problem of tag assignment from the perspective of probability theory, and derive the equations that underlie an HMM tagger. While these equations underpin HMM tagging, none of the authors cited above explicitly derive them; however, derivations very similar to that given below can be found elsewhere in the literature (see, for example, Charniak *et al.*, 1993 and/or Charniak, 1993).

Suppose we have a sentence of length n, containing the words w_1 to w_n. We need to calculate the most probable sequence of lexical category tags $T_{1,n}$, given the words $w_{1,n}$. That is, we need to find:

$$T_{1,n} = \arg\max_{t_{1,n}} P(t_{1,n} \mid w_{1,n}) \qquad (4.1)$$

where $\arg\max_x f(x)$ is the value of x that maximises $f(x)$[34] and $t_{1,n}$ is any possible sequence of lexical category tags. We can simplify equation 4.1 as shown in 4.2; the first step relies on the definition of conditional probability. As the denominator of our new equation is constant, and we are only looking for the maximum value, the further simplification is also valid.

$$T_{1,n} = \arg\max_{t_{1,n}} P(t_{1,n} \mid w_{1,n}) = \arg\max_{t_{1,n}} \frac{P(w_{1,n}, t_{1,n})}{P(w_{1,n})}$$

$$= \arg\max_{t_{1,n}} P(w_{1,n}, t_{1,n}) \tag{4.2}$$

In order to find the most likely sequence of tags, we therefore need to calculate the value of $P(w_{1,n}, t_{1,n})$ for all tag sequences $t_{1,n}$. This is clearly an intractable task – it requires prior knowledge of the frequency of every possible sentence of English. However, if we expand this expression using the axiom in 4.3, we obtain one of two equations (4.4 and 4.5); further simplification of these equations is then possible if we make certain assumptions.

$$P(x_{1,n}) \equiv P(x_1)P(x_2 \mid x_1)\ldots P(x_n \mid x_{1,n-1}) \tag{4.3}$$

$$P(w_{1,n}, t_{1,n}) = P(t_1)P(w_1 \mid t_1)P(t_2 \mid w_1, t_1)\ldots P(w_n \mid t_{1,n}, w_{1,n-1}) \tag{4.4}$$

$$P(w_{1,n}, t_{1,n}) = P(w_1)P(t_1 \mid w_1)P(w_2 \mid t_1, w_1)\ldots P(t_n \mid w_{1,n}, t_{1,n-1}) \tag{4.5}$$

The assumptions we must make in order to obtain a useful language model concern context; effectively, we must decide that only immediate linguistic context has any effect on the probability of a particular word and tag co-occurring. In the NLP literature, this is called the 'Markov assumption' (Charniak, 1993). This is clearly only partially valid; in the majority of cases, the ranking of different probabilities is unaffected by this assumption, but there are some language constructs in which non-immediate context plays a primary role (consider, for example, the cases of long distance dependencies and second arguments of ditransitive verbs).

The type of model we will eventually end up with depends on our definition of immediate context. The simplest approach is to assume that no context except the current word plays any role; this is a unigram model, which can be derived from equation 4.5. The appropriate Markov assumptions are formalised in equation 4.6

[34] This notation is borrowed from Charniak (1993).

and 4.7, and the resulting approximation in equation 4.8.

$$P(t_k \mid w_{1,k}, t_{1,k-1}) \approx P(t_k \mid w_k) \tag{4.6}$$

$$P(w_k \mid t_{1,k-1}, w_{1,k-1}) \approx P(w_k) \tag{4.7}$$

$$\begin{aligned} P(w_{1,n}, t_{1,n}) &\approx P(w_1)P(t_1 \mid w_1) \ldots P(w_n)P(t_n \mid w_n) \\ &\approx \prod_{i=1}^{n} P(w_i)P(t_i \mid w_i) \end{aligned} \tag{4.8}$$

The most probable tag sequence can therefore be found by substituting equation 4.8 into equation 4.2. As the sequence of words $w_{1,n}$ is nonvariant in this equation, the probability of each word is also fixed. The unigram model can therefore be further simplified as shown in equation 4.9.

$$\begin{aligned} T_{1,n} &= \arg\max_{t_{1,n}} \prod_{i=1}^{n} P(w_i)P(t_i \mid w_i) \\ &= \arg\max_{t_{1,n}} \prod_{i=1}^{n} P(t_i \mid w_i) \end{aligned} \tag{4.9}$$

In section 4.5.5 we consider how this equation might actually be used to determine the preferred tag sequence for a sentence. However, first we turn our attention to other possible definitions of immediate context.

One alternative possibility is that the relevant context spans two tags but only a single word (a tag bigram model). We can derive this model from equation 4.4; our new Markov assumptions are formalised in equations 4.10 and 4.11, and the resultant equation is given as 4.12.

$$P(t_k \mid w_{1,k-1}, t_{1,k-1}) \approx P(t_k \mid t_{k-1}) \tag{4.10}$$

$$P(w_k \mid t_{1,k}, w_{1,k-1}) \approx P(w_k \mid t_k) \tag{4.11}$$

$$\begin{aligned} P(w_{1,n}, t_{1,n}) &\approx P(t_1)P(w_1 \mid t_1)P(t_2 \mid t_1) \ldots P(t_n \mid t_{n-1})P(w_n \mid t_n) \\ &\approx P(t_1)P(w_1 \mid t_1) \prod_{i=2}^{n} P(t_i \mid t_{i-1})P(w_i \mid t_i) \end{aligned} \tag{4.12}$$

Equation 4.12 can be further simplified if we introduce a pseudo-tag, t_0, such that $P(t_1 \mid t_0) = P(t_1)$. The simplified version is given in equation 4.13; by substituting

this into equation 4.2 we derive the tag bigram model, shown in equation 4.14.

$$P(w_{1,n}, t_{1,n}) \approx \prod_{i=1}^{n} P(t_i \mid t_{i-1}) P(w_i \mid t_i) \tag{4.13}$$

$$T_{1,n} = \arg\max_{t_{1,n}} \prod_{i=1}^{n} P(t_i \mid t_{i-1}) P(w_i \mid t_i) \tag{4.14}$$

Other contextual models are possible. These include a tag trigram model; in this case the context is taken to consist of three tags. While trigram models and variable length *n*-gram models are common in the NLP literature (Church, 1988; Cutting *et al.*, 1992), the behavioural predictions of such a model differ little from those of the tag bigram model and the increased complexity is not empirically warranted. We therefore pay these scant attention in this thesis.

Definitions of immediate context which include neighbouring words as well as (or instead of) the tag assigned to them are uncommon in the NLP literature. However, some psycholinguists have suggested that word co-occurrence frequencies affect the initial decisions of the HSPM (for example MacDonald, 1993). In terms of the SLCM, such possibilities include a word bigram model (equation 4.15), in which the relevant context spans two words but only a single tag, and the combined bigram model (equation 4.16), in which the context spans two words and two tags.

$$\begin{aligned} T_{1,n} &= \arg\max_{t_{1,n}} \prod_{i=1}^{n} P(t_i \mid w_{i-1}, w_i) P(w_i \mid w_{i-1}) \\ &= \arg\max_{t_{1,n}} \prod_{i=1}^{n} P(t_i \mid w_{i-1}, w_i) \end{aligned} \tag{4.15}$$

$$T_{1,n} = \arg\max_{t_{1,n}} \prod_{i=1}^{n} P(t_i \mid w_{i-1}, w_i, t_{i-1}) P(w_i \mid w_{i-1}, t_{i-1}) \tag{4.16}$$

In section 4.5.4 we consider which of the various context models we have presented should be preferred on rational grounds; we then go on to explore (in section 4.5.5) how these equations might underpin the decision making process of the SLCM.

4.5.4 What Context Model?

By both rational argument (section 4.5.2) and probability theory (section 4.5.3), we have arrived at a number of possible models of the SLCM. These include unigram

and tag, word and combined bigram models. The difference between these models depends on what is taken to constitute context, and this can vary over two dimensions.

We start by studying the word/tag dimension. If we just consider the tag and word bigram models, it is clear that we should prefer the tag variant. Tag context will frequently be useful in determining category ambiguities; in many cases, word context will not be. This problem is exemplified by sentence 4.26:

(4.26) The despicable boasts annoyed Bill.

The word pair "despicable boasts" is uncommon; it is quite likely that an individual may never have heard it before. Where such 'scarcity of data' occurs, some strategy must be used to make a decision in the absence of the relevant statistics. Two techniques have been proposed: 'smoothing' (see Charniak, 1993) and 'backing-off' (Katz, 1987). Both involve the use of coarse-grained statistics where more fine-grained alternatives are not available. Consider again the equation underlying the word bigram model (equation 4.15, reproduced below as 4.17).

$$T_{1,n} = \arg\max_{t_{1,n}} \prod_{i=1}^{n} P(t_i \mid w_{i-1}, w_i) \qquad (4.17)$$

We suggested that the individual might have encountered insufficient evidence to estimate the term $P(t_i \mid w_{i-1}, w_i)$ for all t_i. However, the more coarse-grained term $P(t_i \mid w_i)$, based on a unigram Markov assumption as used in equation 4.9, may be available. In the case of 'smoothing', we always use some part of both the unigram and bigram terms; when there is no evidence on which to base the bigram term, we simply give it a value of 0. If λ is used as a smoothing constant (where $0 \leq \lambda \leq 1$), then the new equation including smoothing is given in equation 4.18:

$$T_{1,n} = \arg\max_{t_{1,n}} \prod_{i=1}^{n} \left[(1-\lambda) P(t_i \mid w_{i-1}, w_i) + \lambda P(t_i \mid w_i) \right] \qquad (4.18)$$

In contrast, 'backing off' only uses the more coarse-grained term when the finer-grained alternative cannot be reliably estimated. For example, where $P(t_i \mid w_{i-1}, w_i)$ was not available it might be approximated from the unigram probability $P(t_i \mid w_i)$.

Returning to example 4.26, we now have a strategy for dealing with the case where

the word pair 'despicable boasts' has not been seen; use unigram probabilities instead. As "boasts" occurs more frequently as a verb than a noun, a word bigram model may initially decide that "boasts" is a verb. In contrast, an individual is extremely likely to have encountered the tag sequence adj–noun before; this statistic will dominate the decision in a tag bigram model. There is therefore a difference in the empirical predictions of the two models in this case. Intuitive evidence suggests that the prediction offered by the tag bigram model is more plausible.

The combined bigram model will also suffer from scarcity of data. However, smoothing or backing off may be an appropriate strategy to use when relevant data is unavailable, and we would expect more plausible predictions from such a model. Having said that, the problem with such approaches from the point of view of a psychological model is that the outcome of a decision may depend on even more parameters and, as we argue below, the word and combined bigram models already suffer from informational complexity.

The other dimension of variance of the possible SLCM models is context length. Unigram models take no account of surrounding context, whereas bigram models span two words; models that take into account trigram or longer contexts are also possible. However, longer contexts require a greater number of parameters. In a unigram model which assigns x different tags to y different words, we would need to record up to xy probabilities for $P(t_i \mid w_i)$[35]. The number of parameters required by a tag bigram model is somewhat larger – x^2+xy – and a tag trigram model requires even more – x^3+xy. Bearing in mind the arguments in favour of informational simplicity outlined in chapter 3, it is clear that we should prefer models that make use of shorter context. We therefore initially propose a unigram SLCM; however, bigram and trigram models are possible if the evidence militates against our initial option.

We can now, once again, consider context type. A word bigram model (as described in equation 4.15) requires up to xy^2 parameters; a combined bigram model (equation 4.16) requires $x^2y^2+xy^2$. As the number of words (y) far exceeds the number of tags (x), the informational complexity of both these models far exceeds that of even the

[35] As many words and tags do not co-occur, xy is the upper limit on the number of probabilities that are required. A similar argument holds for the number of tag co-occurrence and word co-occurrence parameters in bigram models.

tag trigram model. By Occam's razor, we should prefer more complex models only if simpler ones prove defective. In chapters 6 onwards, we consider the empirical evidence and argue that a model which makes use of only tag context is sufficient.

In the next section we consider how the most likely tag sequence for a given sentence might be incrementally determined.

4.5.5 Incrementality and Hidden Markov Models

In section 4.5.3, we derived an equation that allows us to determine the most likely set of lexical category tags ($T_{1,n}$) for a given sentence. If we assume a tag bigram SLCM, this set is described by equation 4.14, reproduced below as 4.19:

$$T_{1,n} = \arg\max_{t_{1,n}} \prod_{i=1}^{n} P(t_i \mid t_{i-1}) P(w_i \mid t_i) \tag{4.19}$$

We can use this equation to determine the most probable tag sequence for a given sentence simply by considering the probability assigned to every possible tag sequence $t_{1,n}$. However, such an algorithm is computationally inefficient and psychologically implausible. It is inefficient because a vast number of different tag sequences must be considered when many are clear losers. The implausibility is due to the non-incremental nature of the algorithm; the lexical category tags for a given sentence can only be discovered once the entire sentence has been heard or read (or, at best, when the next categorially unambiguous word is encountered). However, as reported in chapter 2, the HSPM appears to process language in a highly incremental manner. A plausible SLCM must therefore assign a lexical category to each word as soon as that word is encountered. In this section we consider the unigram and tag bigram SLCM variants, and describe by example an incremental and efficient algorithm for assigning lexical category tags to words (Viterbi, 1967).

Consider sentence 4.27.

(4.27) That old man cries.

We will suppose that each of the words has (at least) two possible lexical categories; some of the words are, in fact, more ambiguous than this. The word "that" may be either a sentence complementiser or a determiner, "old" may be an adjective or a noun, and both "man" and "cries" may be either a verb or a noun. We can represent

this situation as a finite-state automaton, as shown in figure 4.1. We call each possible set of transitions or arcs from the start to the end of the sentence a '*tag path*'; in this example, there are 16 possible tag paths. The solid transitions in figure 4.1 indicate the most probable tag path, while tag paths including dotted arcs are less likely.

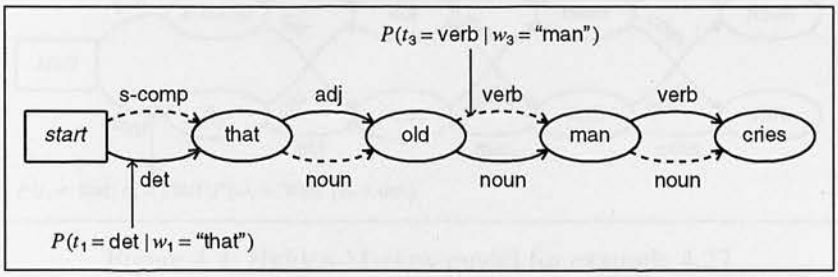

Figure 4.1: Finite-state automaton for example 4.27

Such an automaton, augmented with probabilistic information, is called a *Markov chain*. In figure 4.1, we have annotated a couple of the arcs with their unigram probabilities; we could also have added word bigram probabilities to figure 4.1.

It should be clear that we can calculate the unigram probability (equation 4.7) of a particular tag path for this sentence simply by multiplying together the probabilities of traversing the relevant arcs. For example, the probability of the most probable tag path (which we will call $T_{1,n}$) is given in equation 4.20:

$$\begin{aligned} P(w_{1,4}, T_{1,4}) \approx\ & P(t_1 = \text{det} \mid w_1 = \text{"that"}) P(t_2 = \text{adj} \mid w_2 = \text{"old"}) \\ & P(t_3 = \text{noun} \mid w_3 = \text{"man"}) P(t_4 = \text{verb} \mid w_4 = \text{"cries"}) \end{aligned} \quad (4.20)$$

If the probability assigned to a tag path for an entire sentence is calculated simply by multiplying together the probabilities associated with transitions between the words in the sentence, then a preferred tag can be assigned to each word as soon as the word is read. That is, the sentence can be tagged incrementally and in linear time; the tag initially assigned to the word is simply the one that occurs in the most probable partial tag path up to that word. In the unigram case, the probability of a tag being assigned to a word is independent of context; the most probable tag for word i is simply the one that maximises $P(t_i \mid w_i)$.

The bigram models are slightly more complex. The tag bigrams case can be

represented by a finite-state automaton in which states represent lexical categories and transitions represent words. The automaton for example 4.27 is given in figure 4.2; some of the transitions are annotated with tag bigram probabilities.

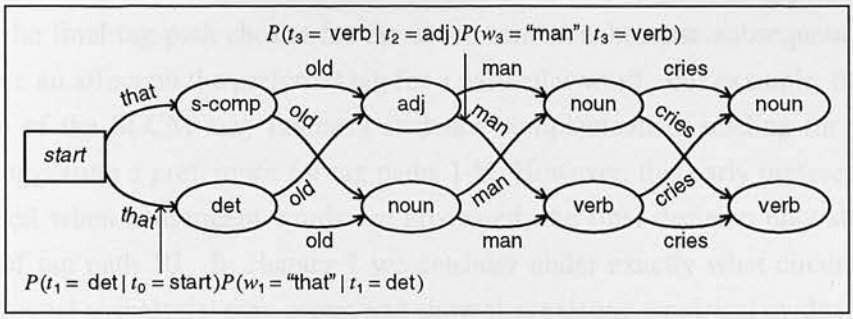

Figure 4.2: Hidden Markov model for example 4.27

In such a model, several transitions out of the same state may represent the same word. Because the tag path through such an automaton cannot be determined from knowledge of only the labels (i.e. words) attached to transitions, the probabilistic version is called a *Hidden Markov Model* (HMM). The problem of finding the most probable tag path through such a model is best viewed as a search problem; the search space for the HMM in figure 4.2 is represented by the tree diagram in figure 4.3. Each tag path in this figure is numbered; the most probable tag path, indicated by the solid line in figure 4.1, is number 10.

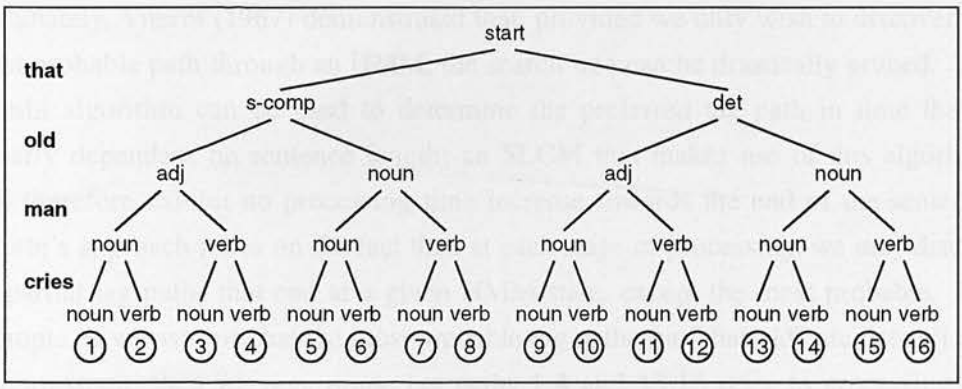

Figure 4.3: Search space for figure 4.2

The probability for a given tag path can be calculated in a similar way to that proposed for the unigram model: simply multiply together the probability associated

with each arc traversal. If the tree given in figure 3 is searched in a breadth-first manner (a word at a time), then an incremental lexical category decision may be made for each word; simply assign the tag that occurs in the most probable partial tag path up to that word. However, unlike the unigram case, this partial tag path may not lead to the final tag path chosen for the entire sentence, because subsequent context may have an affect on the preferred tag for a particular word. For example, the initial decision of the SLCM may favour a sentence complementiser reading for the first word, suggesting a preference for tag paths 1-8. However, this early preference may be revised when subsequent words are processed; the final decision may still be in favour of tag path 10. In chapter 7 we consider under exactly what circumstances such 'internal reanalysis' may occur, and show that existing empirical evidence about the behaviour of the HSPM can be explained by an incremental SLCM that may revise recent decisions.

From the exposition so far, it would appear that the entire search tree (figure 4.3) must be traversed in order to determine the most likely tag path in a bigram SLCM. The size of this tree grows quadratically with sentence length; we would therefore expect any search algorithm to be of complexity dependent quadratically on sentence length. Such an algorithm appears incompatible with intuitive data about the HSPM, as it suggests that processing time required by the SLCM should increase for each word as we process a sentence.

Fortunately, Viterbi (1967) demonstrated that, provided we only wish to discover the most probable path through an HMM, the search tree can be drastically pruned. The Viterbi algorithm can be used to determine the preferred tag path in time that is linearly dependent on sentence length; an SLCM that makes use of this algorithm will therefore exhibit no processing time increase towards the end of the sentence. Viterbi's approach relies on the fact that, at each stage of processing, we may discard all partial tag paths that end at a given HMM state, except the most probable. For example, if we assume that the most probable tag paths for "that old" are det–adj and s-comp–noun, then we may prune tag paths 1-4 and 13-16 prior to processing the next word.

Viterbi's algorithm is correct because the possible completions of any partial path through an HMM ending at the same state are identical; the probability associated

with each completion must also be identical. The probability of the complete tag path is simply the probability of the partial tag path multiplied by the probability of the completion. If one partial tag path has greater probability than a second, and they both share a set of possible completions, then it follows that the most probable complete tag path cannot include the less probable partial tag path.

In the case of the SLCM, where reanalysis may be forced by other modules, it is not so clear that less probable tag paths may be discarded in such a wholesale fashion. However, the SLCM may also be less rational in its approach to determining the most likely tag path for a sentence; if an initially highly improbable tag path turns out to be the correct (and most probable) alternative, then we might still expect processing breakdown. In other words, it is plausible that the SLCM discards a correct analysis if it appears improbable at an early stage of processing. In practise, the SLCM might pursue the n most probable tag paths, rather than strictly follow Viterbi's algorithm, but such an approach remains linear in complexity. Whether the SLCM does keep track of the most probable tag paths, or follows Viterbi's algorithm, has no effect on its initial decision behaviour (chapter 6) and very little effect on its internal reanalysis behaviour (chapter 7), provided n is not too small. However, the former version will more often be able to offer correct reanalysis when an initial decision is rejected by later modules some time after it is made. We do not present any evidence that distinguishes between the two possibilities in this thesis.

4.5.6 Summary

In this section we have considered the problem of lexical category disambiguation from a number of different perspectives. If we consider that the SLCM is a modular process with no access to structural representations, or to representations that are the province of earlier modules, then there are a limited set of statistical measures that may dominate lexical category decisions and all others must be redundant. The measures that might dominate are word–category co-occurrence probabilities, and specifications of prior context in terms of words and lexical categories.

Turning to the NLP literature, we observed that these statistical measures are exactly those used by Hidden Markov Model (HMM) taggers. Such a tagger is a plausible component of the HSPM, as it is a low-cost process which can be used to assign tags incrementally. An SLCM based on an HMM tagger therefore appears worth further

investigation.

However, a number of variants of HMM tagger are possible, depending on our definition of context. A priori, we use Occam's razor to determine a preferred model: the simplest. This is the unigram model in which context plays no role in lexical category assignment. However, the tag bigram model is plausible, and not much more complex; word and combined bigram models require far more parameters and are therefore (prior to examining the evidence) substantially dispreferred.

4.6 Conclusions

In this chapter we have considered lexical category ambiguities and a number of possible processing models for their arbitration. Evidence from DeRose (1988) suggested that such ambiguities are extremely common, perhaps more so than genuine syntactic ambiguity. Understanding the behaviour of the HSPM when faced with such ambiguities is therefore paramount if we are to form a complete processing model.

While lexical category ambiguities have often been treated as a subclass of syntactic ambiguity, we argued (in section 4.2.4) that there are a number of qualitative differences which suggest that such ambiguities may warrant special treatment. Empirical evidence suggests that disambiguation does not occur during lexical access. We therefore propose a model in which lexical category disambiguation occurs in a separate module, prior to syntactic processing but after lexical access.

Chapters 6, 7 and 8 present existing and novel empirical evidence that offers support for our proposed model. Evidence for this model is also evidence for the MSH (see section 3.3.1), as any model that includes the SLCM must be consistent with the MSH. The discussion in this chapter, and the empirical evidence presented in later chapters, therefore not only explores the notion of a separate module concerned with lexical category disambiguation, but also has ramifications for the gross architecture of the HSPM. We return to this issue in chapter 9.

In the next chapter, we introduce the methodologies that we use to explore the SLCM and MSH, and review the computer simulation and tools developed to this end.

5: Methodologies and Tools

5.1 Introduction

In chapter 4 we introduced our own model of lexical category disambiguation – the SLCM. We have argued that lexical category disambiguation may have a privileged status with the HSPM and that the particular statistical model we are proposing is, a priori, plausible. Whether either the model or the wider hypothesis are correct can only be determined by appeal to empirical results; we evaluate the model against experimental data in chapters 6 to 8.

This chapter forms an interlude, or perhaps a bridge, between theory and data. The less avid reader may skip to chapter 6 without losing the argument or missing any empirical results. However, it is in this chapter that we document how the empirical predictions about the behaviour of the SLCM were obtained. In section 5.2, we consider the use of currently available large text corpora as an approximation of the linguistic experience of a community. We highlight some of the problems associated with such an approach, but conclude that it is the best available option.

Section 5.3 introduces the tools created for determining lexical co-occurrence frequencies from text corpora. While it is skipped over briefly in this chapter, tool creation constituted a significant proportion of the work involved in this thesis; further information about the capabilities of these tools can be found in Corley (1996).

Once we have obtained statistical information about a community's prior linguistic experience, precise and stable predictions can be made concerning the behaviour of the SLCM. We make this task somewhat easier by creating a computational simulation of the SLCM. This model allows the relative probabilities assigned to a number of tag paths to be output at each stage of processing; it is therefore possible to determine not only which tag path is preferred, but also by how much it is preferred.

The final, and crucial, stage of evaluation is to compare the behavioural predictions of the SLCM with data pertaining to human performance in the face of lexical category ambiguities. In section 5.4 we briefly consider where we might obtain such data, and how this data might map on to SLCM predictions.

5.2 Estimating Probabilities

The equation used by a tag bigram SLCM to determine a preferred tag sequence was presented as 4.14, and is reproduced below as 5.1.

$$T_{1,n} = \arg\max_{t_{1,n}} \prod_{i=1}^{n} P(t_i \mid t_{i-1}) P(w_i \mid t_i) \tag{5.1}$$

In order to determine the behaviour of this SLCM variant for a particular linguistic utterance, we require access to two statistics about language: $P(t_i \mid t_{i-1})$ and $P(w_i \mid t_i)$ for all (relevant) values of t_{i-1}, t_i and w_i. The definition in 5.2 can be used to estimate both of these probabilities.

$$P(X = x \mid Y = y) \stackrel{\text{def}}{=} \frac{|X = x, Y = y|}{|Y = y|} \tag{5.2}$$

The numerator in this equation, $|X = x, Y = y|$, is a count of the frequency with which two random variables X and Y have values x and y simultaneously, while the denominator, $|Y = y|$, is a count of the number of times Y has value y independent of the value of X. We can use this equation to estimate the word–tag co-occurrence probability (the last part of equation 5.1) for some word w^x and some tag t^y as shown in equation 5.3.

$$P(W_i = w^x \mid T_i = t^y) \stackrel{\text{def}}{=} \frac{|W_i = w^x, T_i = t^y|}{|T_i = t^y|} \tag{5.3}$$

For example, we could calculate the probability of the word "throw" occurring as a noun in some sample of English by counting all occurrences of "throw" as a noun, and dividing this by the total number of nouns in the sample. If our sample is representative of the English usage of a particular community and it is sufficiently large, then we would expect this probability to be a good estimate of the probability with which "throw" occurs as a noun throughout the community.

The equation for estimating the tag co-occurrence probability for two tags t^x and t^y is very similar and is presented in equation 5.4.

$$P(T_i = t^x \mid T_{i-1} = t^y) \stackrel{\text{def}}{=} \frac{|T_i = t^x, T_{i-1} = t^y|}{|T_{i-1} = t^y|} \quad (5.4)$$

Again, we can calculate the probability with which, for example, a determiner is followed by an adjective in some sample of language by counting the number of occurrences of any determiner followed by any adjective in the sample, and dividing this by the number of occurrences of any determiner.

We have suggested that the statistical values used in human sentence processing are estimated from an individual's prior linguistic experience. In the case of the SLCM, learning may be based on the frequency with which a particular tag output for a given word, and following a particular previous output, has been accepted by higher levels of processing. In order to make predictions about SLCM behaviour, we therefore need a sample of language that is representative of the linguistic experience of a particular individual or community, or some more direct method for determining the probabilities used by the various SLCM variants. In this section we consider possible sources of such data.

5.2.1 Questionnaire Studies

A number of researchers have used questionnaire studies to approximate the final production or comprehension preferences of the HSPM. Intuitively, production preferences should be highly correlated with linguistic frequency; the more likely speakers are to produce a particular utterance type, the more often it will occur in the language. A questionnaire study could therefore be used to estimate the relative frequency of occurrence of two or more linguistic entities.

Such a study could take one of two forms. Either subjects would be asked to invent sentences containing ambiguous words, such as "tire". Alternatively, subjects would be asked to complete a partial sentence ending in an ambiguous word, such as 5.1. In either case, responses would then be rated for the reading that the subject assigned to the ambiguous word.

(5.1) I believe that rubber tires...

Both forms of study suffer from a similar problem. In the latter case it is obvious that comprehension processes may well bias the subjects response; plausibility, word co-occurrence frequency or parsing heuristic may all have an effect, depending on the true architecture of the HSPM. As estimating frequencies in this way depends crucially on comprehension processes, the behaviour of the model is likely to reflect HSPM preferences, whether or not the model is correct. In other words, if we use such a method for data collation, the statistical data on which we base our psycholinguistic predictions reflects final comprehension biases, rather than the linguistic experience of any individual or community.

The former technique – asking subjects to invent whole sentences – is less obviously flawed. However, the subject must still comprehend a word before producing an utterance containing that word. Moreover, the language is produced in an artificial environment, and may therefore fail to reflect normal usage patterns. This method of data collection is also laborious and could therefore only be used to obtain data for a very small portion of language. It would be far better to collect examples of unrestricted and naturally occurring language use and count the number of occurrences of the interesting linguistic entities. Large corpora offer us exactly this facility.

5.2.2 Corpora

There are a number of large text corpora available for research use. Several of these are annotated in various ways, often including information about the part-of-speech of each corpus word. Using such corpora, it is possible to establish the frequencies required for equations 5.3 and 5.4. If the corpus is representative of the linguistic experience of a particular community and it is sufficiently large, then we can use this data to predict normal SLCM behaviour for that community. Comparing these predictions with the results of psycholinguistic experiments should allow us to establish whether the SLCM does in fact exist, and which variant is (closest to) correct.

In this section we briefly survey a number of corpora that were available for this project and then, in section 5.2.3, consider which of these corpora is most appropriate for the current work.

The Brown Corpus

The Brown corpus contains 1,017,139 words of American English, and is documented by Kucera and Francis (1967). A more recent version (Francis & Kucera, 1982) includes part-of-speech mark up, and the Penn Treebank 1 (Marcus *et al.*, 1993) distribution contains a retagged and parsed version. It is this latest version that was used for this project. It was tagged automatically by a program called PARTS (Church, 1988) and then corrected by human annotators; we would therefore expect the part-of-speech mark up to be very accurate.[36]

The texts in the Brown corpus cover a wide and representative range of written materials (summarised in table 5.1) but do not contain any spoken language. They also avoid written forms of language that are mainly composed of dialogue, such as drama. All the texts in the corpus were first printed in the United States in 1961.

Category Code	Type of Text	Number of Texts
A	Press: Reportage	44
B	Press: Editorial	27
C	Press: Reviews	17
D	Religion	17
E	Skills and Hobbies	36
F	Popular Lore	48
G	Belles Lettres, Biography, etc.	75
H	Miscellaneous	30
J	Learned and Scientific Writings	80
K	Fiction: General	29
L	Fiction: Mystery and Detective	24
M	Fiction: Science	6
N	Fiction: Adventure and Western	29
P	Fiction: Romance and Love Story	29
R	Humour	9

Table 5.1: Texts in the Brown corpus

The SUSANNE Corpus

The SUSANNE corpus, documented by Sampson (1995), is composed of a subset of

[36] Experience suggests that this assessment is optimistic; however, the tagging is more accurate than that of other corpora of comparable size.

the texts in the Brown corpus; these are taken from the A, G, J and N sections in table 5.1, and total 131,294 words. The corpus has extremely detailed part-of-speech mark up as well as a syntactic information that includes details of both surface and deep structure. All annotation was performed by hand.

The Wall Street Journal Corpus

1,117,250 words of tagged text from the Wall Street Journal are distributed as part of the Penn Treebank 2 (Marcus *et al.*, 1993; 1994) and 1,009,471 of these are parsed. The part-of-speech mark up was again performed automatically by PARTS (Church, 1988) and then corrected by a human annotator; the estimated error for this mark up is 3%.

Unfortunately, while this corpus comprises a slightly larger sample of American English than the Brown corpus, the language it contains cannot be viewed as representative. We discuss this further in section 5.2.3.

The British National Corpus

The British National Corpus (BNC)[37] is by far the largest of the corpora available for this study, containing 100,106,008 words. It was tagged and segmented into sentences by the CLAWS tagger (Garside, 1987). The majority of the corpus was not corrected by human annotators; the estimated error rate for the automatic tagging is 1.7%. Unfortunately, this high precision was achieved by assigning two tags in any case where the tagger could not be sure of a decision; a further 4.7% of words are marked with two tags. A further release of this corpus, expected in 1997, is intended to address this problem.

The texts in the corpus are composed of both spoken and written language; the former form 10% of the corpus. The vast majority of these texts were composed by British authors and published in Britain; in contrast to all the corpora considered so far, this corpus can therefore be seen to represent British English. It is intended to represent the language as it is used today; the majority of the texts date from after 1975. The corpus can therefore be seen as an approximation of the linguistic experience of a young adult speaker of British English.

[37] For more information see http://info.ox.ac.uk/bnc/.

The LOB Corpus

The Lancaster/Oslo-Bergen (LOB) corpus (Johansson, *et al.*, 1986) contains approximately 1 million words of British English, and has been tagged by the CLAWS tagger (Garside, 1987). The genres represented mirror those in the Brown corpus and the texts all date from around 1960. While substantially smaller than the BNC, it represents a 'cleaner' source of data as much more manual correction has been performed. Unfortunately, no tagged version was available for use in the current study.

5.2.3 Which Corpus?

If we are to derive statistical data from a corpus in order to predict the behaviour of the SLCM, the corpus must be representative of the linguistic experience of the community for which we wish to make these predictions. All adults will have been exposed to a vast quantity of language and, for most, the majority of this language will have been spoken. As none of the corpora considered consist mainly of spoken language, we cannot claim that any of them accurately reflect the linguistic experience of an average individual. However, it is unlikely that word frequency and co-occurrence patterns differ vastly between spoken and written language (though frequency of syntactic constructs may). We therefore assume that, for our purposes, data about mainly written language can be used to estimate a linguistic experience composed largely of spoken language.

It is clear that a corpus of British English is more appropriate than an American English corpus for estimating the experience of British speakers. Consider, for example, the noun compound "car park". This is one of the most common noun compounds in the BNC, occurring 1651 times in the 100 million words. However, it does not occur at all in the Brown corpus. It seems plausible that grammatical constructs also occur with different frequencies in British and American English.

It is also likely that exposure to linguistic phenomena will vary regionally, depending on accent, dialect and socioeconomic background. However, few psycholinguistic studies are limited to subjects from one regional area, and none of the corpora available to us are annotated with regional or socioeconomic information. It is therefore necessary to assume that, within national boundaries, individuals have a similar linguistic experience.

It is clear that the best corpus to use when predicting SLCM behaviour for British English subjects is the BNC. It contains mainly British English, and is large enough that we can anticipate reliable data for both unigram and bigram statistics. However, the fact that the corpus was automatically tagged, and has not been post-edited, gives pause for thought. Any tag biases found in the corpus reflect the preference of the tagger, trained on a smaller data set. In other words, relative tag frequencies obtained from the BNC may be no more reliable than those obtained from the data set on which the tagger was trained. Nevertheless, the sheer size of the BNC makes its use appealing and we shall therefore use it whenever possible. Ideally, results concerning British English would be replicated on the LOB corpus; however, this corpus was not available for the current study.

Determining SLCM predictions for American English is more problematic. While even the SUSANNE corpus is large enough to obtain reliable statistics for events which vary over a comparatively small set of possibilities, such as lexical category co-occurrence, none of the American English corpora are sufficiently large to reliably estimate word co-occurrence (or, for less common words, word–tag co-occurrence). We therefore use the BNC to make primary SLCM predictions for American as well as British English and replicate these, where possible, using the Brown corpus. The Brown corpus was chosen in preference to the Wall Street Journal corpus as the latter contains mainly language relating to financial institutions, and this can not be seen as representative of the linguistic experience of an average American citizen.

5.2.4 Summary

In order to make behavioural predictions for the SLCM, frequency data concerning an individual's prior linguistic experience is required. Questionnaire studies do not provide such data, as they cannot differentiate production preferences (which may mirror linguistic frequencies) from comprehension behaviour. Such studies are also laborious to run and provide comparatively little data.

Large text corpora may approximate the shared linguistic experience of a community. However, the majority include mainly written material and do not encode regional variation. Nonetheless, they are the best approximation currently available. We examined a number of corpora and concluded that, despite concern over the accuracy of the part-of-speech mark up, the BNC is the best corpus to use when predicting

SLCM behaviour for British English speakers. The available American English corpora are much smaller than the BNC, so it was decided to make initial predictions for the SLCM behaviour of American English speakers using the BNC, and replicate these where possible using the Brown corpus.

5.3 Tools

While corpora allow us access to large quantities of language that can be seen as fairly representative of a native speaker's linguistic experience, they do not directly encode frequency information. Instead, such data must be collated from the corpus. For all but the smallest corpus, gathering frequency counts requires the use of a tool, in the form of a computer program.

Some corpus frequency counts have been published. We first consider the applicability of these to the current study, before going on to look at available tools. We conclude that all of these are insufficient for the current task, and so introduce our own tools, created specially for this project.

5.3.1 Frequency Tables

Kucera and Francis (1967) published word and sentence length frequency data based on the untagged Brown Corpus and Francis and Kucera (1982) published tables of word–tag co-occurrence frequencies for the tagged version of the same corpus. The latter are also available in machine readable format as part of the MRC linguistic database (Coltheart, 1981).

Unfortunately, while these statistics would be sufficient to model the behaviour of a unigram SLCM for a limited subset of American English, they do not include lexical category co-occurrence counts or transitivity information (see chapter 7). Nor are similar data available for other corpora; in section 5.2 we concluded that our corpus of choice should be the BNC, but no pre-compiled frequency tables are available for this corpus. We therefore consider tools with which we can create our own frequency tables.

5.3.2 Existing Tools

There has recently been a dramatic increase in the number of tools available for

manipulating large text corpora. A list can be found at the IMS tools web site[38] and Schulze and Heid (1994) have produced a comparative study. However, the majority of these tools are designed for lexicographic work and so are aimed more at searching for instances of individual words (in context) than collating lexical co-occurrence frequencies.

At the start of the data gathering phase of the current study (Summer 1995), we had a list of tool requirements which were not met by any available tool or combination of tools. These were:

- Must run on a UNIX system.
- Must allow frequency information to be gathered from a range of corpora, including the BNC.
- Must allow arbitrary transformations on corpus tag sets (such as transitivity marking – see chapter 7) prior to or during searching.
- Must allow collation of co-occurrence statistics that include sentence breaks (see chapter 6.3).
- Must be able to produce output in a variety of formats (as this formed input for a number of small tools, and for the SLCM simulation documented in section 5.3.4).
- Must be highly flexible, so a number of different co-occurrence statistics can be considered.

The only available system that run on UNIX computers was the IMS Corpus Workbench (Christ 1993; 1994). In common with other available tools for other platforms, this was primarily designed for lexicographic work; it was therefore biased towards corpus searching and only included rudimentary facilities for frequency collation. It also did not support the BNC (BNC support became available in late 1996). Finally, there was no support for arbitrary transformations on the corpus tag set and sentence boundary information was (and is) only encoded for some corpora. In summary, its facilities were insufficient for the current project; it

[38] http://www.ims.uni-stuttgart.de/euralex/tools/Concordancer.html

was therefore clear that it would be necessary to write our own tool.

5.3.3 The lstats Tool

The tools used to gather all the statistical information presented in this thesis is called 'lstats'.[39] For speed reasons, it is written entirely in C. The main tool achieves corpus independence by making use of corpus specific filters to transform the raw corpus into its own native format. These filters are based on a code library, and so new filters can be written quickly and easily.

The lstats tool allows collation of arbitrary lexical co-occurrence statistics and these can be presented in a flexible output format. Translation tables allow simple transformations to be performed on the corpus prior to searching, and more complex manipulations may be achieved by writing dedicated filters. While the creation of this tool took a significant part of the time allotted to this project, it would be irrelevant to give a full review of its features in this thesis; documentation can be found in Corley (1996), available, together with the program, from the author.

5.3.4 SLCM Simulation

While it is possible to 'hand run' the SLCM equations in order to make predictions about its behaviour, a computational model is extremely useful if we wish to compare the probabilities assigned to a number of tag paths, or consider SLCM behaviour over time. It is also invaluable for 'what if' analyses. We do all three of these in chapter 7; we therefore constructed a computational implementation of the SLCM to aid in making behavioural predictions.

While the SLCM strongly resembles a traditional part-of-speech tagger, it was not possible to adapt an existing tagger to simulate SLCM behaviour for a number of reasons:

- A tagger is designed to output a preferred analysis at the end of processing. Making behavioural predictions for the SLCM requires access to the relative probabilities assigned to a number of tag paths during processing.

- Most taggers include smoothing functions and strategies for handling

[39] Earlier versions were called 'ngrams'. It was originally intended to add further corpus manipulation tools to create a toolkit, called CORSET (Corley, 1996).

absent data. We have not proposed that such strategies are used by the SLCM; we therefore required a simpler model.

- The mathematical models underlying various existent taggers are notoriously poorly documented (see Charniak *et al.*, 1993). Short of working through the code of a number of taggers, it would be difficult to determine whether any existent tagger uses exactly the model we have proposed in chapter 4.

We therefore constructed our own SLCM simulation, designed to allow exploration of the tagging process rather than just output the final preferred tag path. This simulation can be trained on any corpus for which the lstats tool can collate statistical data. It models only the tag bigram and tag trigram SLCM variants – as we shall see in chapter 6, other variants do not appear to be empirically and rationally justifiable. The simulation also allows counterfeit frequency counts to be added to the training data at run time; this is used in section 7.4 when we consider what would happen if some word–tag frequencies were slightly different from those found in our training corpus.

5.3.5 Summary

Collating statistical data from large corpora such as the BNC requires efficient tools. At the time data was gathered for this project, only one available tool worked on UNIX systems – the IMS Corpus Workbench – and this was considered unsuitable largely due to its inability to search the BNC (at that time). Other deficiencies also contributed to the decision not to use this tool. We therefore constructed our own tool for collating statistics, called 'lstats', which was used to gather all the statistical data used in this thesis.

A computer simulation of the SLCM was also constructed. This is used to generate all the predictions presented in chapter 7.

5.4 Simulation and Evaluation

Frequency data can be obtained from corpora, and, based on this data, it is possible to produce precise and stable predictions about the behavioural characteristics of the SLCM. In order to determine the psychological plausibility of the model, we must

evaluate how well these characteristics match human performance; we therefore need to determine how SLCM predictions might be realised as observable human behaviour.

As the SLCM occurs early in the module chain, SLCM predictions concern the initial decisions of the HSPM, rather than the final outcome of processing. Off line experimental methodologies access only this final outcome, so we would expect the results of such studies to be less relevant to determining whether the hypotheses put forward in this thesis are correct. The results of on line experiments will be most informative for the current study.

A large number of on line experiments measure reading time for words or segments of a sentence, either using self-paced methodologies or eye-tracking. One of the suggested causes of anomalous reading time increases is garden pathing. This occurs when new evidence renders an initially preferred analysis of a sentence implausible or impossible and the sentence processor must construct or access an alternative interpretation. In the case of the SLCM, we might expect reading time increases when a lexical category is initially incorrectly assigned to a word, and following context renders this analysis unlikely. Detection of the anomaly may either occur in later modules (see chapter 6) or in the SLCM itself (see chapter 7).

Because the SLCM is a modular process, we would also expect a reading time increase when the initially preferred analysis of the SLCM is rejected immediately by a later module. For example, consider 5.2:

(5.2) The people who book dinner earliest get the best tables.

The initial decision of the SLCM might favour the noun interpretation of the ambiguous word "book", but the parser cannot construct a licit syntactic structure containing the noun "book" (as only a plural noun could occur at this position)[40]. In this case, reanalysis would be forced while the word is being read; we would expect increased reading time on the ambiguous word itself. In section 6.2, we argue that MacDonald's (1993) results can be explained by exactly this effect, and in section 8.3 we present our own experiment for which we predict processing difficulty of this

[40] For the purposes of exposition, we ignore the reading in which "book" introduces a compound noun.

nature.

5.4.1 Summary

Predictions about SLCM behaviour may be evaluated against the results of existing studies or we may perform novel experiments to test the model. Results of the former type are presented in chapters 6 and 7, and novel experimental data in chapter 8.

In order to evaluate behavioural predictions concerning SLCM decisions against reading time data, we require a theory of when such decisions will lead to processing delays, manifested as reading time increases. We suggest that SLCM decisions may lead to garden pathing within the HSPM when following context is not compatible with an SLCM decision; when the incompatible material is encountered, we would expect reading time increases. We also predict reading time increases on the ambiguous word when SLCM decisions are deemed implausible by later modules. Evidence supporting the existence of such processing delays is presented in chapters 6 and 8.

5.5 Conclusions

In order to make predictions concerning the behaviour of the SLCM, it is necessary to have access to an individual or a community's linguistic experience. Large text corpora are a good approximation of this experience; however, they are only an approximation and so results based on them should be treated with mild caution. Of the available text corpora, the BNC is the largest and the best suited to the current study; however, any predictions made using the BNC for American English should be replicated on a corpus of American English, such as the Brown corpus.

If we are to collate statistics from corpora, we require a computer program to aid in the task. No existing computer program was available that fulfilled all our requirements; we therefore constructed our own tool called 'lstats'. Documentation for this tool can be found in Corley (1996). We also wrote a computer simulation of the SLCM, which is used extensively in chapter 7.

However, the test of the SLCM is in its ability to predict empirical data. In this chapter we considered how behavioural predictions about the decisions of the SLCM

might be reflected in reading time data collected in on line experimental studies. In chapters 6, 7 and 8 we evaluate whether existing and novel experimental data does support the SLCM model.

6: Existing Evidence – Initial Decisions

Mistakes are always initial.

(Cesare Pavese, quoted in The Faber Book of Aphorisms)

6.1 Introduction

In chapter 4 we presented four variants of a statistical model of lexical category disambiguation within a modular HSPM. We called this model the Statistical Lexical Category Module (SLCM) and the variants were the unigram, tag bigram, word bigram and combined bigram SLCMs. So far, our reasons for advocating the SLCM have been based on rational arguments; these were explored in some depth in chapters 3 and 4. However, any proposed psychological model must be tested against empirical evidence. In this and the next chapter, we present existing evidence about the behaviour of the HSPM when faced with words that exhibit lexical category ambiguity, and consider whether this data matches SLCM predictions. In chapter 8, we present our own experimental results.

It is important to note that the SLCM has a very high accuracy. In general, part-of-speech taggers assign the correct lexical category to a word in an English sentence 95% of the time (Charniak, 1993). As the basic architecture of the SLCM is identical to that of a 'standard' HMM tagger, we would expect it to display similar performance. We know that the eventual output of the HSPM is highly accurate; it therefore seems likely that, in the vast majority of cases, it makes the correct initial decision in the face of ambiguity. If this were not true, we would not be able to rapidly understand even simple linguistic utterances. Thus we already know that the SLCM is empirically viable for the vast majority of language; as psycholinguistic models are often tested mainly against empirical evidence about situations in which the HSPM makes incorrect initial decisions, it is not clear whether many other models make correct predictions for the vast quantity of unproblematic examples.

However, the SLCM does have distinctive breakdown and repair characteristics.

Breakdown occurs when it assigns an incorrect lexical category to a word when that word is first encountered; effectively, the SLCM makes an incorrect initial decision. Repair occurs when the SLCM corrects that decision. This may be due to a subsequent module indicating that the initial decision of the SLCM is not viable; as we shall see in chapter 7, repair may also occur when subsequent context leads the SLCM itself to reconsider an earlier decision.

In this chapter we consider the breakdown characteristics of the SLCM – where it makes an incorrect initial decision. We look at evidence from MacDonald (1993) concerning sentences containing noun–verb ambiguities, and propose an analysis of her results that is compatible with all SLCM variants, including the unigram and tag bigram versions which are ruled out by MacDonald's own analysis. We then consider evidence about the ambiguous word "that", from Juliano and Tanenhaus (1994), and demonstrate that this data is incompatible with a unigram SLCM; however, the predictions of a tag bigram SLCM match Juliano and Tanenhaus' results.

6.2 Noun–Verb Ambiguities

Following on from Frazier and Rayner's (1987) work, MacDonald (1993) considered lexical category ambiguities in which the ambiguous word may be either a verb or a noun. In section 4.4 we considered her first experiment, which shows that Frazier and Rayner's results were probably an artefact of their choice of materials. In this section, we turn to her second experiment, which apparently demonstrates that the initial decisions of the HSPM when faced with such ambiguities are influenced by semantic plausibility. Such a result is clearly incompatible with the proposed SLCM, as it suggests that lexical category disambiguation is not a modular task, encapsulated from semantic representations.

6.2.1 MacDonald's Results

MacDonald's materials resemble Frazier and Rayner's, in that the ambiguous word follows a determiner–noun sequence. However, all her materials are disambiguated to favour the verb reading of the ambiguous word; she manipulates the plausibility of a noun compound reading between conditions. Example materials are shown in 6.1 and 6.2:

(6.1) a. The union told reporters that the warehouse *fires* many workers each spring...

b. The union told reporters that the corporation *fires* many workers each spring...

(6.2) a. The union told reporters that the warehouses fire many workers each spring...

b. The union told reporters that the corporations fire many workers each spring...

We shall call the ambiguous word and the word preceding it the 'critical region'; we refer to the two words in this region as c_1 and c_2. The region following the critical region is the 'disambiguating region'; we refer to the words in this region as $d_1 \ldots d_n$. In 6.1a, the critical region ("warehouse fires") forms a plausible noun compound. As the disambiguating region always forces a verb reading for c_2, MacDonald calls this an 'unsupportive bias'. In contrast, the noun compound reading of the critical region ("corporation fires") in sentence 6.1b is implausible; the verb reading is favoured by a 'supportive bias'.

The items in 6.2 are unambiguous versions; in both the (a) and (b) forms the noun phrase reading is ruled out on syntactic grounds (though see section 6.2.2). These therefore constitute control materials with which to compare performance in both the critical and disambiguating regions.

MacDonald's hypothesis is that, if semantic bias does affect the initial lexical category decisions of the HSPM, then in the ambiguous unsupportive bias condition (6.1a) the processor will initially assign a noun reading to c_2. We would therefore expect processing difficulty in the disambiguating region in this condition, and this should be reflected by greater reading time than in the unambiguous control (6.2a). In contrast, in the ambiguous supportive bias condition (6.1b), semantic constraints militate against a noun reading and so we would expect the initial decision of the HSPM to favour a verb reading. As a verb reading will also be initially preferred in the equivalent unambiguous condition (6.2b), we would expect no reading time difference for the disambiguating region of these two conditions.

Finally, MacDonald notes that disambiguation is not forced immediately on reading

d_1; English allows reduced relative clauses to follow a noun as in 6.3. The category ambiguity therefore persists for some words – a switch in preferred analysis, and the associated reading time differences, may not be apparent until the a few words into the disambiguating region.

(6.3) The warehouse *fires* many workers were killed in burnt all night.

MacDonald tested her hypothesis using a moving window self-paced reading study. This methodology allows estimation of reading times for individual words; length-adjusted reading times[41] for d_5 are shown in figure 6.1 (d_5 is the fifth word of the disambiguating region, which we did not include in examples 6.1 and 6.2). The results MacDonald obtained apparently back up her hypothesis. Summing the reading times for all but the first word in the disambiguating region reveals an interaction between ambiguity and bias; in the supportive bias condition, there was no effect of ambiguity in this region, whereas in the unsupportive bias condition, ambiguous reading times were a mean of 20ms/word longer than their unambiguous counterparts. Significant reading time differences (in the unsupportive bias condition) were also found individually for words d_5 and d_8.

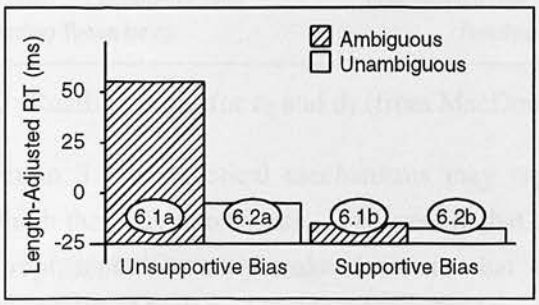

Figure 6.1: Reading times for d_5 (from MacDonald 1993)[42]

However, MacDonald also found significant differences in the *opposite* direction for c_2 and d_1 in the unsupportive bias condition; the reading times for these two words are shown in figure 6.2. MacDonald argues that the difference at c_2, which she calls

[41] Length-adjustment involves subtracting a predicted reading time for a word (based on the length of that word) from the actual reading time; in theory, this allows reading times for words of different lengths to be systematically compared. For more details of the procedure, see chapter 8.

[42] Figures 6.1 and 6.2 were created by approximating the values from the graphs in MacDonald's (1993) paper. They are intended for exposition rather than accuracy.

a 'reverse ambiguity effect', can be attributed to the fact that the ambiguous unsupportive bias condition is the only one in which a verb phrase structure is not constructed at this point; as the work done in creating a verb phrase is greater than that involved in integrating a noun into an existing noun phrase, we might expect increased reading times when a verb phrase is constructed (MacDonald, 1993; see MacDonald, 1994, for further discussion). The persistence of this effect to the following word (d_1) is due to 'spill-over', common in self-paced reading studies. This analysis is supported by work showing that reading times are sensitive to syntactic complexity (Just & Carpenter, 1980). If we accept this explanation, then it appears that MacDonald's results are consistent with an effect of semantic plausibility on the initial lexical category decisions of the HSPM.

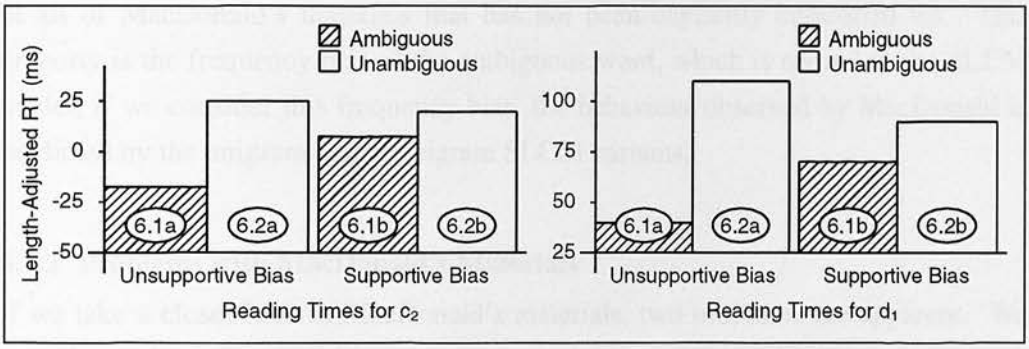

Figure 6.2: Reading times for c_2 and d_1 (from MacDonald 1993)

As discussed in section 3.3.4, statistical mechanisms may capture regularities in representations to which they have no access. It is feasible that a process that has no access to semantic representations may make decisions that appear semantically-motivated. In her discussion, MacDonald suggests just such a statistical explanation of her results. She shows a correlation between 'unsupportive bias' and some fine-grained statistical measures, including word–word co-occurrence frequencies (e.g. the noun compound "warehouse fires" occurs more frequently than "corporation fires") and the head–modifier preference of the first noun (e.g. "warehouse" is more likely to occur as a modifier than "corporation", which is more likely to occur as a head noun). MacDonald suggests that the initial decisions of the HSPM are made on the basis of these statistics.[43]

[43] As MacDonald's own proposed model (MacDonald *et al.*, 1994) allows the unconstrained interaction of multiple information sources, the model may make direct use of plausibility

Such a fine-grained statistical account is compatible with the word and combined bigram SLCMs, both of which make use of word–word co-occurrence statistics. However, in section 4.5 we outlined reasons why these versions should be less preferred. A unigram or tag bigram model does not make use of word–word co-occurrence statistics; as the only difference between the supportive bias and unsupportive bias condition is the word at c_1, it appears that such a model cannot account for MacDonald's data.

In the next section, we consider some possible problems with MacDonald's experiment. Alternative analyses are suggested by potential flaws in her materials. We conclude that the results she presents are probably valid nonetheless – however, as we shall see in section 6.2.3, these results may be attributed to a crucial property of all of MacDonald's materials that has not been explicitly controlled for. This property is the frequency bias of the ambiguous word, which is central to the SLCM model; if we consider this frequency bias, the behaviour observed by MacDonald is predicted by the unigram and tag bigram SLCM variants.

6.2.2 Problems with MacDonald's Materials

If we take a closer look at MacDonald's materials, two problems are apparent. We consider each of these in this section.

Are MacDonald's Unambiguous Materials really Unambiguous?

The first problem with MacDonald's materials has to do with the manipulation she uses to produce unambiguous control items. It is unclear whether the critical regions in example 6.2 are completely unambiguous; there is no *syntactic* constraint violated by forming a noun compound out of a plural noun followed by a singular noun. In fact, there are a number of such noun compounds in common usage, including "sports car", "antiques fair" and "teachers meeting".

MacDonald might well argue that "sports car" is an idiomatic expression equivalent to a single lexical item; in general, noun compounds composed of a plural and a singular noun are unusual, and the HSPM would therefore be foolish to initially

information. It is therefore unclear why MacDonald apparently prefers an account based on local statistics.

adopt this analysis. However, occasional 'foolish' initial decisions are a primary characteristic of modular systems.

Suppose that we accept that a noun compound analysis is ruled out on either syntactic, semantic or pragmatic grounds. Does this then mean that we can be sure that the subject will have decided on the verb analysis by the end of the critical region? Unfortunately, the subject's behaviour is still not clear; an alternative analysis of both sentences in 6.2 is to assume that the critical region is treated as a Saxon genitive lacking an apostrophe (cf. "warehouse's fire").

However, if we take the view that MacDonald's unambiguous materials are flawed and do not favour a verb analysis, then we are left unable to explain why reading times for the majority of the disambiguating region were found to be *longer* in the ambiguous unsupportive bias condition than in the unambiguous equivalent. While there is a potential problem with MacDonald's materials, it is apparently ruled out by her results.

A Single Lexical Item?

The second problem with MacDonald's materials is that many of the critical regions in the ambiguous unsupportive bias condition occur extremely frequently as noun compounds in American English. The estimated frequency per million words of each is shown in table 6.1.

Noun Compound	Frequency per million words
prison guard	0.12
miracle cure	0.27
fraternity house	0.01
bank account	4.81
office supply	0.10
computer program(me)	3.48
official document	0.00
window frame	1.28
employee benefit	0.10
college loan	0.01
business contact	0.48
tape measure	0.72
army base	0.34
grocery store	0.25
warehouse fire	0.03
tax return	0.42

Table 6:1: Frequency of MacDonald's experimental items (from BNC)

Unfortunately, many of MacDonald's experimental items did not occur in the largest representative corpus of American English available to us – the Brown corpus. Frequencies were therefore estimated from the BNC; however, the statistics discovered may not be accurate for American English. In particular, "fraternity house" and "college loan" are American but not British English. "Tax return" and "employee benefit" both occurred infrequently in the BNC, but 6 and 4 times respectively in the million word Brown corpus.

Nonetheless, MacDonald's noun compounds are fairly frequent. It seems plausible that the HSPM may treat extremely frequent noun compounds as single lexical items; examples include "interest rate" and "world war". In this case, the noun compounds MacDonald used in her experiment may also be treated as unitary entities during sentence processing.[44]

If this is the case, we would expect an initial decision favouring a noun compound analysis of the ambiguous unsupportive bias materials (and, at least in some models, a verb analysis of the others) without the necessity of supposing that semantic

[44] I am grateful to Chuck Clifton for first pointing out this analysis to me.

information interacts with lexical category decisions, or that these decisions rest on fine-grained statistical information. This prediction matches MacDonald's reported results for the disambiguating region.

Further support for this hypothesis is provided by the reading times reported for c_2 (figure 6.2). If the two words in the critical region are treated as a single lexical item, then we might expect lexical access and syntactic integration of the second word to be extremely rapid.

It seems then, that MacDonald's results may be entirely explained by postulating that many of the noun compounds in the ambiguous unsupportive condition are treated as single lexical items. However, the average frequency of these noun compounds (0.78 occurrences per million words) is far lower than that of "interest rate" (31.48) or "world war" (37.17). A single lexical item analysis is therefore possible but does not seem highly probable. In chapter 8 we conduct our own study, which partially replicates MacDonald's experiment, but avoids the pitfalls outlined here. In the next section we consider exactly what the predictions of a unigram or tag bigram SLCM would be, and show these are, in fact, consistent with MacDonald's results.

6.2.3 SLCM Predictions

It is clear that both a unigram and tag bigram SLCM would make the same initial decisions for analogous items across the two ambiguous conditions and across the two unambiguous conditions. In the unigram case, this is because the decision is dominated by a statistic depending only on the ambiguous word. In the bigram case, the dominating statistic depends on the ambiguous word and the probability of each tag sequence for the prefix of the sentence preceding the ambiguous word; as this prefix is identical across conditions with the same bias, this latter probability will not change.

Considering the tag bigram SLCM first, what will its initial decision be? If the ambiguous word (c_2) is the k^{th} word in the sentence, then we may calculate the SLCM's initial decision for this word using equation 6.1:

$$T_{1,k} = \arg\max_{t_{1,k}} \prod_{i=1}^{k} P(t_i \mid t_{i-1}) P(w_i \mid t_i) \qquad (6.1)$$

However, in all MacDonald's materials c_1 is either unambiguously a noun, or is far more frequently a noun than any other grammatical class. A noun reading is also congruent with c_1's syntactic context. We would therefore expect by far the most probable tag sequence for the prefix of the sentence up to c_1 to end in a noun. A noun reading for c_1 is also compatible with both noun and verb readings for c_2; we can therefore be sure that the most probable tag sequence up to c_2 also involves assigning a noun reading to c_1.

Given this information, we need not calculate equation 6.1 for the entire prefix ending at c_2; the qualitative results of this calculation will be identical to simply calculating which possible tag for c_2 is most likely to follow a noun. The new formula for determining the initial decision of the tag bigram SLCM for c_1 is given in equation 6.2.

$$T_k = \arg\max_{t_k} P(t_k \mid t_{k-1} = \text{noun}) P(w_k \mid t_k) \tag{6.2}$$

It so happens that $P(t_i = \text{noun} \mid t_{i-1} = \text{noun})$ and $P(t_i = \text{verb} \mid t_{i-1} = \text{noun})$ are roughly equal: using the SUSANNE corpus and the 'lstats' tool described in chapter 5, we can estimate the former at 0.183 and the latter at 0.207;[45] obtaining these statistics from a variety of other corpora leads to similar results. The initial decisions of the tag bigram SLCM will therefore largely depend on $P(w_k \mid t_k)$ – they will be similar to those of a unigram SLCM.[46]

The initial decisions of the unigram SLCM depend entirely on the category bias of the individual words used. We can estimate this bias from a corpus, using equation 6.3; figure 6.3 shows the bias of each of the ambiguous words used in MacDonald's experimental items, estimated from the BNC and the Brown corpus.

$$bias = \log\left(\frac{noun\ count}{verb\ count}\right) \tag{6.3}$$

[45] These figures were obtained using a broad definition of 'noun' and 'verb'. Narrower definitions tend to increase the probability of the noun–noun sequence.

[46] The bigram and unigram variants of the SLCM use different measures of lexical probability ($P(w_k \mid t_k)$ and $P(t_k \mid w_k)$ respectively). While the latter depends solely on the category bias of the individual word, the former also depends inversely on the overall frequency of each tag considered. However, the overall frequency of nouns and verbs is similar, so we would expect any strong frequency bias to dominate the ranking for these values of t_k in both the unigram and bigram cases.

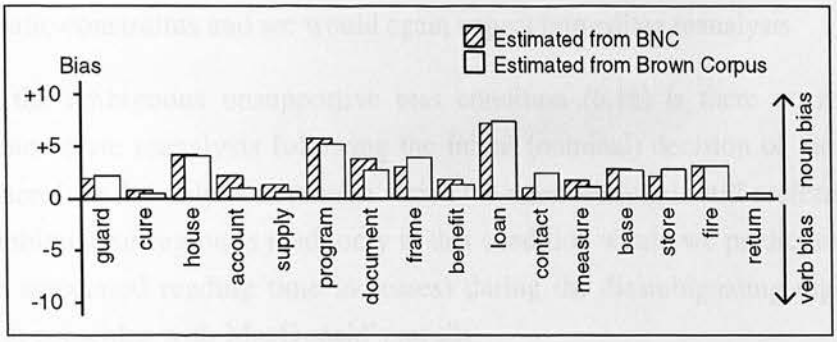

Figure 6.3: Bias of ambiguous words in MacDonald's (1993) experiment

For the BNC data, the mean bias is 2.66 (about a 14:1 ratio in favour of noun occurrences) and the standard deviation is 1.79; for the Brown corpus, the mean bias is 2.60 (13:1) and the standard deviation 1.82.[47] We group together the singular and plural ("-s") forms of the ambiguous word to collate this data; evidence that the SLCM does not distinguish between these is presented in chapter 8. However, even if we consider singular and plural forms separately, we discover that all items except the plural "returns" exhibit a clear noun bias; "returns" is biased towards a verb reading in the BNC (bias = -0.09) but towards a noun reading in the Brown corpus (bias = 1.18).

In summary, all MacDonald's ambiguous words were biased towards a nominal reading; most were strongly biased. We would therefore expect both unigram and tag bigram SLCM models to make an initial decision favouring noun category for the ambiguous word in all conditions. As the SLCM precedes syntactic and semantic processing, this includes the 'unambiguous conditions'.

However, the SLCM is just one stage in the sentence processor. Later modules may reject the analysis it proposes. In the unambiguous conditions (6.2), syntactic (or possibly semantic or pragmatic – see section 6.2.2) constraints rule out the noun reading; we would therefore expect later modules to force immediate reanalysis, *prior to reading the disambiguating region.* That is, the *early* decision of the entire HSPM would not be identical to the *initial* decision of the SLCM. In the ambiguous supportive bias condition (6.1b), the noun reading is similarly ruled out by semantic

[47] The alternative spelling "programme" was included in the count for "program". In the Brown corpus, the word "loan" never occurred as a verb. To avoid a gap in the data, we estimated its verb frequency as 0.05 occurrences per million words; this estimate is based on the BNC data.

or pragmatic constraints and we would again expect immediate reanalysis.

Only in the ambiguous unsupportive bias condition (6.1a) is there no reason to expect immediate reanalysis following the initial (nominal) decision of the SLCM. This is therefore the only condition in which the noun reading is still preferred when the disambiguating region is read; only in this condition would we predict reanalysis (and the associated reading time increases) during the disambiguating region; this prediction coincides with MacDonald's results.

However, the explanation we have put forward leads to a further prediction. We suggested that there is immediate reanalysis (forced by later modules) on reading the ambiguous word in all but the unsupportive bias ambiguous condition. As is common in self-paced reading studies, evidence of processing difficulty at c_2 may spill over onto the next word (d_1). We would therefore expect increased reading times for both these words in the unambiguous conditions (6.2), and in the ambiguous supportive bias conditions (6.1b). In particular, an SLCM model leads to the prediction of increased reading time in the unambiguous unsupportive bias condition (6.2a), compared to the ambiguous unsupportive bias condition (6.1a), but no such difference between the supportive bias conditions (6.1b and 6.2b). This prediction explains MacDonald's reported results for these two words (c_2 and d_1) – her 'reverse ambiguity effect' – which she attributed to syntactic complexity. We detailed these in section 6.2.1 and they were summarised in figure 6.2, repeated below as figure 6.4. The predictions of the unigram and tag bigram SLCM variants exactly match MacDonald's reported results without the need to appeal to the independent 'reverse ambiguity' effect.

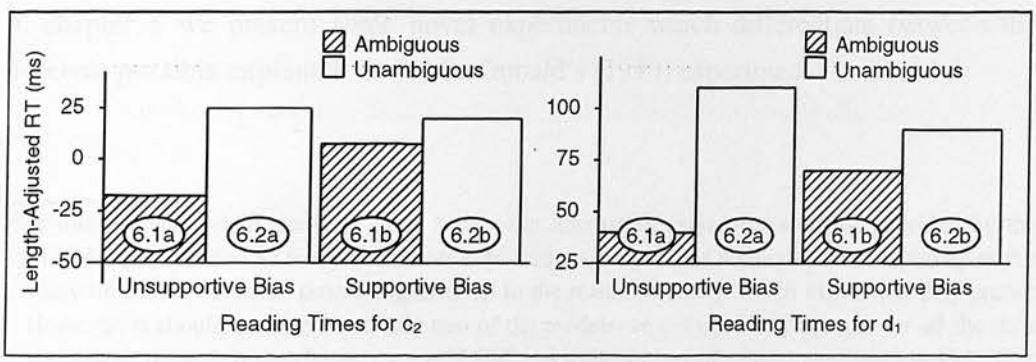

Figure 6.4: Reading times for c_2 and d_1 (from MacDonald 1993)

6.2.4 Alternative Explanations[48]

While we have shown that MacDonald's (1993) results are consistent with all SLCM variants, this does not mean that they cannot also be explained by other models of sentence processing. MacDonald's own explanation of her results, in combination with the 'reverse ambiguity effect', has not been disproved. The latter may be tested by considering whether a similar 'reverse ambiguity effect' occurs when the ambiguous word is biased towards a verbal reading; we do this in chapter 8.

MacDonald's findings are also consistent with any model that initially assigns a noun reading to all the ambiguous words, provided that model is modular and makes the same assumptions about reanalysis as we have done. Thus the Garden Path theory may account for MacDonald's data (see chapter 4 for further discussion).

Finally, and despite MacDonald's protests to the contrary (MacDonald, 1993, p.703), these results are consistent with the delay strategy. Frazier and Rayner (1987) suggest that "the processor delays syntactic integration of new input items (until disambiguating information is encountered) under circumstances where alternative (stored) representations of an input are activated" (p.507); however, they make no claims as to what constitutes "disambiguating information". If disambiguation may be pragmatic as well as syntactic, then they may expect MacDonald's supportive bias conditions to act as immediate disambiguation. In this case, the delay strategy would predict that suspended syntactic integration would lead to a processing delay during disambiguation only in the ambiguous unsupportive bias condition; in this condition, reading times for the ambiguous word would also be reduced due to the suspension of parsing. This tallies with MacDonald's results.

In chapter 8 we present some novel experiments which differentiate between the different possible explanations of MacDonald's (1993) experiment.

[48] In this chapter and in chapters 7 and 8 we offer alternative explanations to that provided by the SLCM. The fact that so many alternative explanations are possible highlights the complexity of the study of human sentence processing; it is up to the reader to decide which explantion they prefer. However, it should be noted that only two of the models we consider may account for all the data we present in these three chapters – the SLCM and some variant of a constraint-based model. Our arguments against the latter are given in chapter 3 and are reiterated throughout these three chapters.

6.2.5 Summary

MacDonald's (1993) investigation of noun–verb ambiguities led her to the conclusion that lexical category decisions are influenced by semantic plausibility or fine-grained statistical mechanisms. This conclusion is compatible with the word and combined bigram SLCM models, but not with the coarser-grained unigram and tag bigram versions, which predict the same initial decision for all of MacDonald's materials.

However, if we consider the SLCM as part of a larger model, we are led to the conclusion that, while the initial decision of a unigram or tag bigram SLCM would always favour the same reading, later processes would force immediate reanalysis in three out of four of MacDonald's experimental conditions. In this case, the behavioural predictions for the disambiguating region match MacDonald's results. We also predict extra processing delays at the ambiguous word when immediate reanalysis is forced; MacDonald found exactly this effect, which she attributed to the greater syntactic complexity of the verb reading.

While the SLCM analysis is both simple and appealing, a number of different accounts may be offered for MacDonald's data. In chapter 8 we present novel experiments that distinguish between the different possibilities.

6.3 "That" Ambiguity

The experiment presented in section 6.2 does not differentiate between the unigram and tag bigram variants of the SLCM, as the lexical category context of the ambiguous word does not vary between conditions or bias the initial decisions of the tag bigram SLCM. However, Juliano and Tanenhaus (1993) have presented evidence concerning the effect of syntactic context on lexical category ambiguity resolution. In this section, we consider that evidence and show that it offers support for a tag bigram SLCM.

6.3.1 Juliano and Tanenhaus' Primary Results

Juliano and Tanenhaus (1993) investigated the initial decisions of the HSPM when faced with the ambiguous word "that" in two different syntactic contexts – sentence initially and following a verb. Example materials for these two cases are given in 6.4

and 6.5 respectively:

(6.4) a. *That experienced* diplomat would be very helpful to the lawyer.
b. *That experienced* diplomats would be very helpful made the lawyer confident.

(6.5) a. The lawyer insisted *that experienced* diplomat would be very helpful.
b. The lawyer insisted *that experienced* diplomats would be very helpful.

In all four conditions, the critical region ("that experienced") is ambiguous; "that" may either be a sentence complementiser or a determiner. In examples 6.4a and 6.5a (the 'NP conditions'), the disambiguating region (which immediately follows the critical region) forces the determiner reading; this is because a singular noun such as "diplomat" cannot occur in a noun phrase that does not contain a determiner. In contrast, the disambiguating region in 6.4b and 6.5b (the 'complement conditions') forces the sentence complementiser reading; a plural noun such as "diplomats" may occur without a determiner, but not with a singular determiner such as "that".

Juliano and Tanenhaus hypothesize that the initial decisions of the HSPM follow the regular pattern in the language. In other words, the preferred analysis will depend on the contingent frequency with which "that" occurs in different syntactic contexts. In an analysis of the Brown corpus, they discovered that sentence initially "that" is most frequently a pronoun (54%), then a determiner (35%) and finally a complementiser (11%). In contrast, "that" is most frequently a complementiser (93%) following a verb; determiner (6%) and pronoun (1%) readings have comparatively low frequency.

According to Juliano and Tanenhaus' account, for the sentence initial examples (6.4), the initial decision of the HSPM will favour a determiner reading (though this prediction appears at odds with their statistics, above, which support a pronoun reading; the authors do not explain this anomaly). In the NP condition (6.4a), this analysis is congruent with the disambiguating material, and so no processing disruption is expected. In the complement condition (6.4b), reanalysis is forced when the disambiguating material is read, and Juliano and Tanenhaus therefore predict a reading time increase at this point.

In contrast, the initial decision reading for "that" following a verb is the sentence

complementiser reading. This is consistent with the disambiguation in the complement condition, but not in the NP condition (6.5a); Juliano and Tanenhaus therefore predict a reading time increase in the disambiguating region for the latter condition.

Juliano and Tanenhaus' results for the first and second words of the disambiguating region (d_1 and d_2) are shown in figure 6.5. While reading time differences on d_1 were negligible, the combined reading times for d_1 and d_2 showed a clear interaction between condition and the position of "that". In the sentence initial conditions, reading times for d_2 were significantly slower for complement condition items than for NP condition ones; when the ambiguous word followed a verb, the reverse effect was significant by subjects and just significant by items (p=0.1).

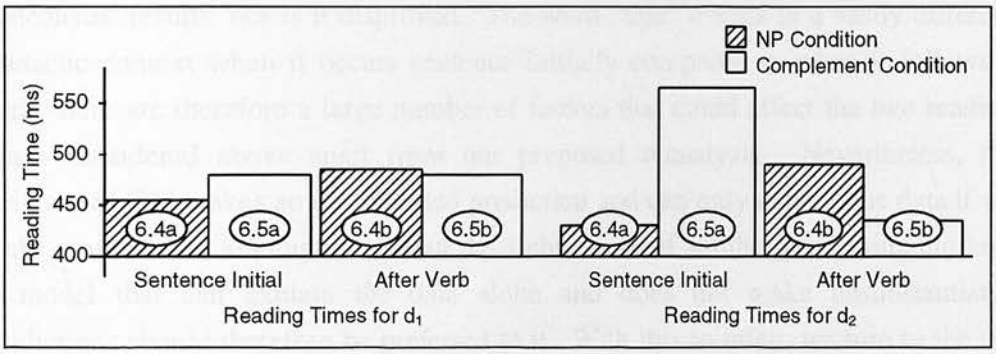

Figure 6.5: Reading times for d_1 and d_2 (from Juliano & Tanenhaus, 1993)

These results apparently support Juliano and Tanenhaus' thesis: when disambiguation does not favour the regular pattern increased reading times are observed.

Juliano and Tanenhaus' (1993) second experiment shows that, unlike Frazier and Rayner (1987), their results cannot be attributed to the incongruency of using a deictic determiner. Their third experiment concerns "that"-preference, an effect which, we suggest in section 7.5, could be explained by SLCM internal reanalysis.

6.3.2 SLCM Predictions

On first glance, it would appear that a unigram SLCM is incapable of explaining Juliano and Tanenhaus' results. Such a model simply predicts that the HSPM always initially chooses the most probable tag for a given word; in the case of "that", this tag

is 'complementiser'.

However, in our analysis of MacDonald's (1993) experiment (section 6.2), we suggested that an initial decision of the SLCM may be revised immediately if it is ruled out on syntactic or pragmatic grounds. While sentence initial complement clauses (as in 6.4b) are syntactically permissible, they are comparatively rare. We might suggest that a parser that makes use of statistical information would force immediate reanalysis when a sentence complementiser was proposed at the start of a sentence. Such a model would predict an increased reading time for the word "that" in the sentence initial condition; Juliano and Tanenhaus' data does not support this prediction (mean reading time sentence initially is 390ms, after verb is 411.5ms).

While we have not found support for the unigram SLCM from Juliano and Tanenhaus' results, nor is it disproved. The word "that" occurs in a vastly different syntactic context when it occurs sentence initially compared to when it follows a verb; there are therefore a large number of factors that could affect the two reading times considered above apart from our proposed reanalysis. Nevertheless, the unigram SLCM makes an unsupported prediction and can only explain the data if we make unwarranted assumptions about the architecture of another processing module. A model that can explain the data alone and does not make unsubstantiated predictions should therefore be preferred to it. With this in mind, we turn to the tag bigram SLCM.

On first glance, it would appear that the tag bigram SLCM is very similar to Juliano and Tanenhaus' proposal, and should therefore make the same predictions. Juliano and Tanenhaus suggest that initial decisions depend on the *contingent* frequency of a word occurring with a particular lexical category in a particular syntactic position. The SLCM is oblivious to syntactic structure, and instead uses an estimate of the probability of a word occurring with a given tag following a particular lexical category context. However, the syntactic context is entirely determined by the category of the immediately preceding lexical item (in Juliano and Tanenhaus' analysis) and so we would expect these two approaches to give the same result.

If we are to determine the predictions of the tag bigram model, we must first consider what tags are likely to be assigned to the word preceding the ambiguity. In the

sentence initial conditions, there is no problem. Juliano and Tanenhaus (1993) do not include a full list of their experimental materials; we therefore simply assume that the word preceding the ambiguity was always either an unambiguous verb, or far more likely to be a verb than any other lexical category. In this case, we can ignore all other possibilities in our analysis.

Unfortunately, no tagged corpus (that we know of) makes a distinction between the pronoun and determiner uses of "that". The most detailed tag markup available to us is in the SUSANNE corpus – but even in this corpus both uses receive the tag "DD1a". It is therefore not possible to use frequencies established directly from a corpus to determine the behaviour of an SLCM that does distinguish between these uses.

One possibility is to assume that the linguistic theory apparently espoused by corpus developers is correct, and there is no distinction between the two uses. This would be consistent with Juliano and Tanenhaus' analysis. In this case, the SLCM need only decide between determiner and sentence complement readings.

In order to determine the behaviour of the tag bigram SLCM, we need to estimate the appropriate word given tag probabilities ($P(w_i \mid t_i)$) and the appropriate tag co-occurrence probabilities ($P(t_i \mid t_{i-1})$). An estimate of the word given tag probability for each possible reading (from the BNC) is shown in table 6.2; the tag co-occurrence results are in table 6.3.

t_i	**Complementiser**	**Determiner**
w_i = "that"	1.0	0.171

Table 6.2: $P(w_i$ = "that" $\mid t_i)$ for two different values of t_i (from BNC)

t_i	**Complementiser**	**Determiner**
t_{i-1} = verb	0.0234	0.0296
t_{i-1} = start	0.0003	0.0652

Table 6.3: $P(t_i \mid t_{i-1})$ for different values of t_i and t_{i-1} (from BNC)

It is the interaction of these probabilities that leads to the SLCM behaviour. The word given tag probabilities alone (table 6.2) are far higher for the complementiser

reading than the determiner one.[49] The tag co-occurrence data (table 6.3) follows the regular pattern in the language – complementisers are more frequent following a verb than sentence initially, and determiners are more frequent sentence initially than after a verb. However, if we just look at the tag co-occurrence probabilities, the *most* probable tag in both contexts is 'determiner'; alone, this statistic does not account for Juliano and Tanenhaus' reported results. The interaction of the two probabilities ($P(w_i = \text{"that"} \mid t_i)P(t_t \mid t_{i-1})$) is shown in table 6.4.

t_i	**Complementiser**	**Determiner**
t_{i-1} = verb	0.0234	0.0051
t_{i-1} = start	0.0003	0.0111

Table 6.4: $P(w_i = \text{"that"} \mid t_i)P(t_i \mid t_{i-1})$ for different values of t_i and t_{i-1}

When "that" follows a verb, the very weak bigram probability bias in favour of the determiner reading is easily overcome by the stronger word given tag bias towards a complementiser reading. Thus the initial decision of the tag bigram SLCM when faced with "that" following a verb is in favour of a complementiser reading. This agrees with Juliano and Tanenhaus' results.

In the sentence initial case, the strong bigram bias in favour of a determiner is not overridden by the word given tag bias. Thus the initial decision of the SLCM when faced with "that" sentence initially is in favour of a determiner reading. Again, this agrees with Juliano and Tanenhaus' results.

So the predictions of the tag bigram SLCM match Juliano and Tanenhaus' results, provided we assume that the pronoun and determiner readings of "that" are assigned identical lexical categories. As the SUSANNE corpus only contains 238 occurrences of "that" as either a pronoun or a determiner, it is possible to manually annotate them; doing so reveals that 129 of them are pronouns and 109 are determiners. The exact maths then depends on whether we assume that "that" as a pronoun forms a category on its own, is grouped with all demonstrative pronouns, or is grouped with all pronouns of any sort. However, in all cases we find a slight preference for the pronoun reading sentence initially, with the determiner reading coming a close

[49] This slightly unintuitive result arises from the fact that the probability depends inversely on the overall frequency of the tag.

second. After a verb, the sentence complementiser reading is still preferred.

Under these assumptions, it seems that the SLCM therefore makes an incorrect prediction, suggesting that the pronoun analysis is initially preferred at the start of a sentence. This prediction does not appear to tally with Juliano and Tanenhaus' results. However, as we shall see in chapter 7, the tag bigram SLCM may initiate reanalysis of an earlier decision internally if it leads to the later supposition of a particularly improbable tag sequence. It so happens that the sequence pronoun–adjective is rare enough for this to happen (no matter how we group pronouns). The SLCM therefore changes its decision to prefer a determiner reading for "that" immediately the following adjective is read.

The tag bigram SLCM therefore, in common with Juliano and Tanenhaus' model, predicts an initial decision that is not transparently supported by the data. However, the SLCM model also leads to the prediction that reanalysis will occur on the following word. Juliano and Tanenhaus do not provide sufficient detail of their experimental results to either substantiate or refute this prediction;[50] it is therefore compatible with, but not directly supported by, their experiment.

Because we predict reanalysis on the second word, we do predict that, in the sentence initial case, the preferred analysis of the SLCM will be in favour of a determiner by the time the disambiguating region is encountered. The expected reading time pattern during the disambiguating region is therefore identical to that reported by Juliano and Tanenhaus and their results can be seen as support for the SLCM model. However, further work is required to determine whether the predicted reanalysis on the second word does actually occur.

6.3.3 Alternative Explanations

The Garden Path theory predicts a preference both sentence initially and following a transitive verb for the determiner or pronoun reading of "that", as this reading involves the construction of the fewest new nodes (MA). However, a variant of the theory in which subcategorisation information is available early (see section 7.4) makes different predictions. Following a verb that may only take a sentence complement, this variant of the Garden Path theory favours the sentence

[50] In fact, the experimental design would not allow this hypothesis to be tested.

complementiser reading, as this reading involves the immediate construction of fewer new nodes than supposing that "that" is a determiner introducing an embedded subject NP. Therefore, variants of the Garden Path theory in which subcategorisation preferences guide initial decisions could account for Juliano and Tanenhaus' data, provided that the majority of their verbs can only take a sentence complement. Unfortunately, Juliano and Tanenhaus do not give detailed information about their experimental materials, so it is unclear whether the account offered by the Garden Path theory is tenable.

Other non-statistical heuristic models, such as Construal and Generalised Theta Attachment, may also be able to account for Juliano and Tanenhaus' data, given sufficient information about their experimental materials. Therefore, contrary to the authors claims, the data as presented does not demonstrate that the HSPM must use statistical decision processes. We also suggested that MacDonald's (1993) data may be compatible with non-statistical models; we have not, as yet, proved that lexical category decisions are made on a statistical basis. We present our own evidence in chapter 8.

In contrast, Referential Support does not offer any predictions for this ambiguity. A coarse-grained variant of Tuning (which, we assume, ignores subcategorisation information) also appears incompatible with this data, given our intuitive assumption that verbs are more often followed by object noun phrases than by sentence complements; however, this assumption may be wrong, so further study is required.

Finally, interactive models may account for Juliano and Tanenhaus' data. However, whether they *do* or not will depend on the exact settings of a large number of different parameters. We would therefore argue that such models could never *predict* this data.

6.3.4 Summary

Juliano and Tanenhaus (1993) demonstrated that the HSPM makes a different initial decision when faced with the ambiguous word "that" sentence initially and after a verb. In the former condition, a determiner reading is preferred. In the latter, the HSPM initially assigns a complementiser reading to the ambiguous word.

In this section we showed that the predictions that can be inferred from Juliano and Tanenhaus' proposed model do not entirely match their results, as they should predict an initial decision favouring the pronoun reading in the sentence initial conditions. The unigram SLCM was also inadequate unless we supposed that an initial SLCM decision that does not agree with the results is rapidly revised by a statistical parser. Given the lack of evidence to support such a complex proposal, we concluded that a unigram SLCM is not a viable model if we wish to explain the empirical data. In section 4.5 we argued that word and combined bigram SLCMs, as well as models employing trigram and longer contexts, should be dispreferred on rationalist grounds. We are therefore left with the tag bigram SLCM as our preferred model, and we assume this variant in the remainder of this thesis.

The tag bigram SLCM did prove capable of explaining Juliano and Tanenhaus' results. However, if a distinction is made between pronoun and determiner readings of "that", then the tag bigram SLCM predicts an initial decision in favour of the pronoun reading, apparently contrary to the experimental data. We argued that SLCM internal reanalysis (explored in chapter 7) would in fact force early reanalysis, before the disambiguating material, to favour the determiner reading. The SLCM therefore predicts the results Juliano and Tanenhaus obtained.

Finally, we argued that other non-statistical models might also predict Juliano and Tanenhaus' results. The evidence so far does not conclusively show that initial lexical category decisions are made on a statistical basis. We present further evidence in chapter 8.

6.4 Conclusions

In this chapter we have tested our proposed SLCM against the existing evidence concerning the initial decisions of the HSPM in the face of lexical category ambiguities. The tag bigram variant of the SLCM has proved the simplest version that is capable of explaining the data.

Other proposed models rely on the supposition of syntactic complexity effects to explain unexpected results (MacDonald, 1993) or do not explain why or how a prediction that does not agree with the data is transformed into an initial decision that does (Juliano and Tanenhaus, 1993). MacDonald *et al.*'s (1994) model (see section

2.5) relies on fine-grained and complex statistical mechanisms; it is interesting to note that the coarse-grained and extremely simple tag bigram SLCM can fully explain data which their model fails to completely capture.

However, it is possible that other models (including the Garden Path theory) may be able to explain the data presented in this chapter, particularly if they assume rapid reanalysis when initial decisions lead to unlikely structures at some higher representational level. While the evidence presented in this chapter suggests that lexical category decisions are guided by statistical knowledge, this evidence is not conclusive. In chapter 8 we report the results of experiments that demonstrate that frequency effects do play a role in such decisions.

The tag bigram SLCM has proved the most viable SLCM variant, both on the grounds of its simplicity (compared to the word and combined bigram versions) and on the grounds of its ability to explain existing data (compared to the unigram SLCM). The bigram SLCM variants also display repair characteristics when they initially assign an incorrect tag; in the next chapter we demonstrate that the repair characteristics of the tag bigram SLCM may offer novel explanations for existing experimental results.

7: Existing Evidence – Internal Reanalysis

7.1 Introduction

In chapter 6 we considered the breakdown characteristics of the SLCM, and concluded that a tag bigram variant is sufficiently powerful to explain existing data, but has fewer parameters, and is therefore simpler, than many alternative models. In explaining MacDonald's (1993) results concerning noun–verb ambiguities, we suggested that reanalysis occurs when an initial decision of the SLCM is rejected by a higher level of processing; we will call such externally imposed reprocessing 'external reanalysis'.

However, bigram SLCM variants also display 'internal reanalysis' or 'repair' characteristics. That is, the SLCM may make an initial decision when faced with an ambiguous word, and then change that decision in the light of following context, without any external influence. In this chapter we consider the internal reanalysis characteristics of the SLCM and determine whether they can account for existing experimental evidence.

In section 7.2 we look at how repair works within the SLCM. Repair can only occur in a very limited set of circumstances; we define these circumstances in this section and then go on to consider a real world example where SLCM internal reanalysis might explain human behaviour.

In section 7.3 we consider experimental evidence about reduced relative constructions (from MacDonald, 1994) which appears to support an interactive model of sentence processing. However, SLCM internal reanalysis provides a simple modular account of the data. We then go on, in section 7.5, to consider Mitchell's (1987) data, which apparently demonstrates the late availability of verb subcategorisation information in human parsing; again, SLCM internal reanalysis offers an alternative explanation for this data.

Finally, we consider whether the behaviour we describe in this chapter really

originates from the SLCM or is just a crude simulation of syntax. By way of example, we consider Trueswell, Tanenhaus and Kello's (1993) results concerning "that"-preference.

7.2 SLCM Internal Reanalysis

In chapter 4 we considered the behaviour of the SLCM when faced with example 4.27, reproduced here as 7.1.

(7.1) That old man cries.

In our exposition, we assumed that the initial decisions of the SLCM were always correct. However, as we saw in chapter 6, both SLCM variants make incorrect initial decisions in a variety of circumstances. For instance, consider example 7.2.

(7.2) The old man the oars.

In this case, we would expect the SLCM decisions for the first three words to be similar to those made for example 7.1; the preferred sequence would be det–adj–noun. This lexical category assignment is not compatible with the following lexical context. In cases like this, the SLCM may sometimes revise earlier decisions without the need for external interference from later modules. Importantly, such revision is not an extension of the model but follows directly from its Bayesian underpinnings: the 'conclusion' about what is the best tag sequence can (and should) be revised on the basis of new 'evidence'.

7.2.1 When and Why

When would such a revision occur? In chapter 4 we presented a search tree (figure 4.3) for an HMM (figure 4.2) for example 7.1; figure 7.1 represents the search tree associated with an HMM for example 7.2. We simplify this diagram by assuming that "the" and "oars" are lexically unambiguous, so there are only four possible tag sequences that could be assigned to the sentence. The correct analysis of the sentence is represented by the search path numbered 4 in the figure, but we would expect an SLCM initial decision to favour path 1 when "man" is first encountered.

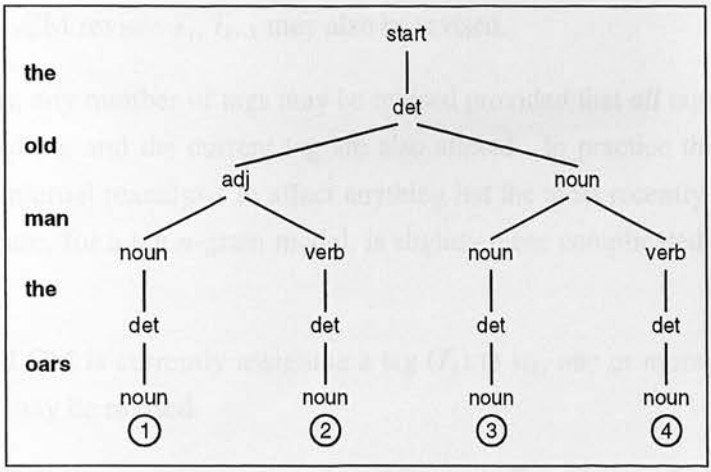

Figure 7.1: Search space for HMM for example 7.2

Internal reanalysis occurs when the SLCM switches to a different tag path without intervention from other modules. There are strict limitations on when this can occur. Consider the case of a tag bigram SLCM: following figure 7.1, a decision is made between adjective and noun readings for "old", or between tag paths 1-2 and 3-4. When "man" is processed, this decision may be changed; if the SLCM originally chose paths 1-2, it is *possible* that it would choose path 3 or 4 at this point. However, knowledge of co-occurrence frequencies in English suggests that the SLCM would actually prefer path 1. The SLCM may again change decision when faced with the next word "the". If it initially preferred path 1, it is *possible* that it would switch to path 2 or path 4. It is *not possible* for the preferred tag sequence to change to that represented by path 3.

In terms of the search tree, the SLCM preferred analysis may switch to an alternative branch, thereby revising an earlier decision, provided that branch does not share any common lexical categories with the initially preferred branch below the point where they split. Path 1 and path 3 both assign the tag 'noun' to "man", but split at the preceding word – it is therefore impossible for the SLCM to subsequently make a revision from the former path to the latter since the latter could not have a higher probability. An alternative way of stating this (for the tag bigram variant) is that the SLCM is constrained by the following two rules when revising previous decisions:

- If the SLCM is currently assigning a tag (T_k) to w_k, T_{k-1} may be revised.

- If the SLCM revises T_i, T_{i-1} may also be revised.

In other words, any number of tags may be revised provided that *all* tags between the earliest revised tag and the current tag are also altered. In practice this means it is very rare for internal reanalysis to affect anything but the most recently assigned tag. The general case, for a tag n-gram model, is slightly more complicated; the rules are as follows:

- If the SLCM is currently assigning a tag (T_k) to w_k, one or more of T_{k-n+1} to T_{k-1} may be revised.
- If the SLCM revises T_i, one or more of T_{i-n+1} to T_{i-1} may also be revised.

In chapter 4 we discussed the Viterbi algorithm (Viterbi, 1967), which relies on these restrictions on internal reanalysis. We have therefore already briefly considered the reason why these restrictions hold. To recap, and perhaps clarify, consider again example 7.2. In this example, internal reanalysis may occur on the third or fourth words. Given that the comparative probabilities for each of the four tag paths up to the third word ("man") are $P_1 \ldots P_4$ respectively, we can calculate the comparative probability of each of the tag paths for the first four words as shown in table 7.1 (based on equation 4.13).

Tag Path	Comparative Probability
1. det–adj–noun–det	$P_1 \times P(t_4 = \text{det} \mid t_3 = \text{noun}) P(w_4 = \text{the} \mid t_4 = \text{det})$
2. det–adj–verb–det	$P_2 \times P(t_4 = \text{det} \mid t_3 = \text{verb}) P(w_4 = \text{the} \mid t_4 = \text{det})$
3. det–noun–noun–det	$P_3 \times P(t_4 = \text{det} \mid t_3 = \text{noun}) P(w_4 = \text{the} \mid t_4 = \text{det})$
4. det–noun–verb–det	$P_4 \times P(t_4 = \text{det} \mid t_3 = \text{verb}) P(w_4 = \text{the} \mid t_4 = \text{det})$

Table 7.1: Comparative Probabilities of different tag paths

Suppose that the initial decision of the SLCM (after processing the third word) was in favour of tag path 1 (P_1 is greater than P_2, P_3 and P_4). If the probability of a determiner following a verb is greater than that of a determiner following a noun, it is possible that the preferred tag path after the fourth word will be whichever was previously more probable of 2 or 4; in this case, internal reanalysis would have occurred. However, when calculating the probability of tag paths 1 and 3 up to word 4, we multiply P_1 and P_3 respectively by the same quantity. It is therefore impossible

that, if P_1 is greater than P_3, the comparative probability of tag path 3 up to word 4 is greater than that of tag path 1. In other words, as the same completions are possible for tag paths 1 and 3 (once word 3 has been processed), the initially less preferred tag path can never come to have a greater probability than the more preferred alternative. If only internal reanalysis is considered, the SLCM will never make such a revision.

In the remainder of this chapter we examine psycholinguistic data that may be explained by SLCM internal reanalysis. Internal reanalysis differs from externally imposed reanalysis in that it occurs 'bottom up'. External reanalysis involves a higher level module rejecting the analysis of a previous module, and thereby imposing 'top-down' constraints on the analysis. The SLCM, and any other internally parallel statistical module, may also revise its initially preferred analyses internally on the evidence of following context; such revision forces higher level modules to also alter their analysis based on 'bottom up' information. This involves one less processing step than external reanalysis; we therefore postulate that such reanalysis results in comparatively small processing difficulty.

7.2.2 A Real World Example

In order to make this more concrete, we will look at a real internal reanalysis prediction of the SLCM. This concerns the word "her", which is ambiguous between accusative personal pronoun (cf. "him") and possessive (cf. "his") readings. Example 7.3a exemplifies the personal pronoun reading, and 7.3b the possessive reading.

(7.3) a. Without her the contributions were lost.
b. Without her contributions he was lost.

We can track the probability assigned to each of the two relevant tag paths using our SLCM simulation, described in chapter 5. We trained this on the comparatively small SUSANNE corpus; however, this corpus contains sufficient lexical category co-occurrence statistics for our purpose, and all the words in the example are of fairly high frequency. We would therefore expect the results to be representative of those we would obtain with a far larger training set. Figure 7.2 shows the comparative probability assigned to each tag path for each of the sentences in 7.3.[51]

[51] To make the graph clearer, we scale the probabilities at each word to add up to one. We make the

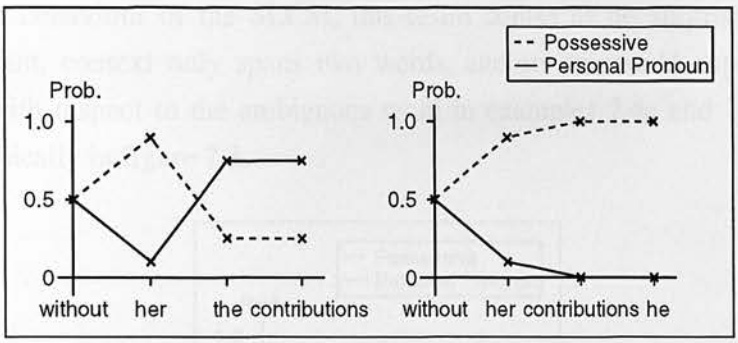

Figure 7.2: SLCM predictions for "her"

In both sentences, the initial decision of the SLCM when faced with the ambiguous word "her" is to prefer the possessive reading. However, in 7.3a the next word is a determiner, and the bigram sequence possessive–determiner occurs infrequently in the training corpus; it is therefore judged to be highly improbable. In contrast, the sequence personal pronoun–determiner is not so uncommon and is therefore considered far more probable. The difference in these two probabilities overcomes the initial preference for the possessive reading, and so the SLCM switches preferred tag path, thereby altering a decision made about a previous word.

In contrast, the initial preference for a possessive reading for sentence 7.3b is reinforced by the fact that "her" is followed by a noun; the sequence possessive–noun is more frequent than personal pronoun–noun. The initial decision is therefore maintained and, in this example, there is no internal reanalysis.

In both of these cases, the SLCM eventually settles on the correct reading for the sentence via purely internal reanalysis. This appears compatible with our intuitions; neither of these sentences lead to conscious processing difficulty. However, consider example 7.4a (adapted from Pritchett, 1992).

(7.4) a. Without her contributions were lost.
b. Without him contributions were lost.

This example is syntactically identical to 7.4b, which is unambiguous. It is therefore grammatical; however, according to Pritchett, respondents report that they experience conscious processing difficulty when trying to parse similar sentences. If we

same adjustment in all other figures showing the results of SLCM simulation.

consider the behaviour of the SLCM, this result comes as no surprise; in the tag bigram variant, context only spans two words, and so we would expect identical behaviour with respect to the ambiguous word in examples 7.4a and 7.3b. This is shown graphically in figure 7.3.

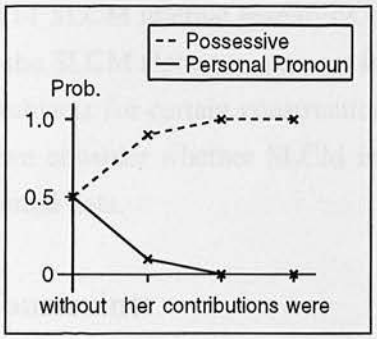

Figure 7.3: SLCM predictions for example 7.4a.

In this case, the initially incorrect decision supporting a possessive reading is not overridden by later context. No SLCM internal reanalysis occurs and we therefore predict greater processing difficulty than in either 7.3a or 7.3b; whether the sentence will be understood depends entirely on the behaviour of later modules. This prediction agrees with Pritchett's report that this sentence causes conscious processing difficulty. However, as Pritchett's data is based on his subjects' intuitions, rather than more objective experimentation, the accuracy of this prediction is still open to question. We know of no published experimental evidence about "her" ambiguities against which to test the SLCM predictions.

7.2.3 Summary

The SLCM may exhibit internal reanalysis. This occurs when a decision made about a previous word is revised in the face of following context; we show that this may happen in a very limited set of circumstances, normally only affecting the word preceding the one currently being processed.

Such 'bottom up' reanalysis differs fundamentally from more traditional 'top down' reanalysis in that it is instigated by modules earlier in the processing chain, but affects those later. In other words, it is unidirectional. We would therefore expect that such reanalysis is comparatively low cost; that is, the processing difficulty

associated with internal reanalysis is less than that for the external version, in which information flows in both directions through the module chain. Internal reanalysis therefore appears to offer benefits to the processing system, both in the early detection of errors and their rapid repair.

We considered an example of SLCM internal reanalysis – ambiguities involving the word "her". We show that the SLCM alone can account for the conscious processing difficulty experienced by subjects for certain constructions using this word. In the remainder of this chapter we consider whether SLCM internal reanalysis may also account for online experimental data.

7.3 Post-Ambiguity Constraints

MacDonald (1994) investigated a number of contextual manipulations which can reduce the processing difficulty associated with misanalysis of main verb/reduced relative ambiguities (see example 2.9). Among these are what MacDonald terms 'post-ambiguity constraints', where context following the ambiguous word reduces the processing difficulty experienced by experimental subjects. These appear to fit well with SLCM internal reanalysis, in which lexical category co-occurrence statistics dominate the decision process. In this section we consider whether the SLCM may account for MacDonald's data.

7.3.1 The Data and MacDonald's Account

Consider examples 7.5 and 7.6:

(7.5) a. The sleek greyhound *raced at the track* won four trophies.
b. The sleek greyhound *admired at the track* won four trophies.
c. The sleek greyhound shown at the track won four trophies.

(7.6) The sleek greyhound *raced at the track* all day long.

In sentences 7.5a and 7.5b, two possible syntactic structures may be assigned to the four italicised words (the ambiguous region). Either the first verb is the main verb of the sentence and the following prepositional phrase modifies it, as in example 7.6, or the region is a reduced relative clause, as in the unambiguous 7.5c. The disambiguating region ("won four trophies") in both 7.5a and 7.5b forces the latter reading. However, 7.5a and 7.5b differ in that, in the latter sentence, the ambiguous

verb is strongly biased towards being transitive (in its main verb reading), whereas the verb in 7.5a is more likely to be intransitive. In the unambiguous case (7.5c), the first verb can only be a past participle and so a main verb reading is ruled out on syntactic grounds.

MacDonald found a significant processing delay in the disambiguating region ("won four trophies") in the intransitive condition (7.5a), compared to unambiguous examples (7.5c), but no such delay in the transitive condition (7.5b). However, the ambiguous region in the transitive condition sentences took significantly longer to read than for the analogous region of the intransitive condition items.

MacDonald argues that the initial decision of the HSPM favours the incorrect main verb reading in both cases. However, transitive verbs are normally immediately followed by noun phrases; when this is not the case, a strong constraint is violated. This 'post-ambiguity' constraint results in immediate revision of the preferred analysis, before the end of the ambiguous region; hence the processing delay MacDonald found in the ambiguous region for the transitive condition. Assuming that the HSPM now favours the correct reduced relative reading, MacDonald predicts no processing delay in the disambiguating region for the transitive condition items; again, her results confirm this prediction.

In contrast, no constraint is violated when an intransitive verb is followed by a prepositional phrase. MacDonald therefore predicts no processing delay in the ambiguous region of sentence 7.5a. However, as the incorrect main verb analysis will still be preferred when disambiguation is reached, she does predict a processing delay, or perhaps even a conscious garden path, in the disambiguating region. Her reading time evidence again agrees with her predictions.

So far, we have considered only what MacDonald termed 'good constraints' – when the word following the verb unambiguously signals that the next phrase is not a noun phrase. MacDonald also experimented with 'poor constraints', illustrated by example 7.7:

(7.7) The sleek greyhound *admired all day long* won four trophies.

In this case, the constraining material following the ambiguous word ("all day long")

is also temporarily ambiguous, and could be analysed as a noun phrase until the third word ("long") is read. Such poor constraints proved less helpful to subjects – accuracy in answering comprehension questions following experimental items was lower for the poor constraint conditions than for the good constraint ones (exp. 3) and, reading time in the disambiguating region was greater in conditions where the post-ambiguity constraint was poor. This suggests that the reanalysis behaviour MacDonald observed is based on lexical, rather than syntactic, co-occurrence.

Such behaviour appears to fit well with a model in which lexical co-occurrence statistics dominate, such as the SLCM. However, as we shall see in the next section, the nature of the tags used in our current SLCM variant renders it insufficient to explain this reanalysis data. Our search for an explanation leads us to reexamine the concept of a lexical category.

7.3.2 Reconsidering Lexical Categories

If we assume that the lexical categories assigned by the SLCM are simply parts of speech, then SLCM internal reanalysis is transparently incapable of explaining MacDonald's (1994) data, as her materials involve no ambiguity at this level of granularity. However, it is in no way clear that lexical categories should be so coarse-grained. In a number of constraint-based, or 'lexicalist', models (e.g. MacDonald *et al.*, 1994), words are assigned very detailed lexical representations, including syntactic and semantic information and even partial parse trees.

Within a modular model, early determination of so detailed lexical representations does not make sense. Augmenting the output of the SLCM to include syntactic and semantic constructs without also giving it access to these representations appears counter-productive. Since we would expect that the semantic component of the HSPM would be best placed to make decisions in the face of purely semantic ambiguity, an architecture in which the SLCM makes such decisions would be unlikely to operate efficiently.

On the other hand, it does make sense to propose an SLCM that assigns lexical categories augmented with syntactic features that can (normally) be predicted from lexical co-occurrence. That is, where an SLCM architecture can assign syntactic features accurately, it seems plausible that it should do so. One example is verb

transitivity – whether a verb is transitive or not depends largely on the individual word and its immediate syntactic environment; however, the lexical context is extremely predictive of the immediate syntactic environment, so we would expect high accuracy from an SLCM that assigns tags that encode both part-of-speech and, in the case of verbs, transitivity. Moving the burden of determining whether a verb is transitive from the syntactic module to the SLCM reduces the burden on the syntactic processor, without greatly increasing the inaccuracy of initial decisions. We therefore posit that the SLCM does assign transitivity information, and reconsider MacDonald's (1994) results.

7.3.3 Training Data for Transitivity

Unfortunately, it is not immediately apparent how a tagger that encodes transitivity information may be trained; in no corpus (that we know of) do the lexical category tags include information pertaining to the transitivity of a particular verb token.[52] The SUSANNE corpus, which has extremely detailed tags, includes information on whether a verb type is (always) transitive or intransitive, but such information is only useful in the case of unambiguous verbs, which we would not expect to be problematic anyway. At first glance, it would appear that the only way to obtain training data for our model SLCM is to manually mark all verbs in a corpus for their transitivity. In chapter 5, we observed that the SUSANNE corpus was just large enough to provide representative lexical category co-occurrence data; it contains 23,545 verb tokens. Manually assigning transitivity information would therefore be prohibitively time consuming.

However, the SUSANNE corpus does have extremely detailed syntactic mark up, including both surface and logical structures. Using this, we may automatically assign transitivity information to verbs. The algorithm used is described below; it distinguishes three types of verbs – transitive verbs, which have a noun phrase or sentential complement, attributive verbs, which occur in an adjectival role (e.g. "the *running* man"), and intransitive verbs.

Informal study suggests that this algorithm is extremely accurate in assigning a correct category to a verb type. However, no objective precision statistics have been

[52] In subsequent work for Sharp Laboratories of Europe, the author has created a version of the Brown Corpus in which such information is included in the part-of-speech tags.

collated; we presume that, for the current work, it is accurate enough.

To determine the transitivity of a verb:

- A verb is transitive if it is preceded or followed by a noun or clause, in the same sentence, which is the logical or surface object of a verb, and:
  - the noun or clause is not contained within a verb phrase that does not include the verb.
  - the verb is not contained within a verb phrase or non-finite clause that does not include the noun or clause.
- A verb is attributive if it is not transitive and it occurs within a noun phrase but not within a verb phrase or non-finite clause dominated by that noun phrase ("the *running* man").
- Otherwise, the verb is intransitive.

Algorithm to determine the transitivity of a verb

7.3.4 Modelling MacDonald's Results

Our retagged SUSANNE corpus allows our SLCM simulation to be trained to assign tags that include transitivity information. This new version should assign distinct analyses to the sentences in each of MacDonald's experimental conditions; it is therefore possible that internal reanalysis within such a tag bigram SLCM might account for MacDonald's results. If predictions derived from this model match the post-ambiguity constraint data, then the SLCM provides a simpler and more predictive model than that proposed by MacDonald – it does not include arbitrary constraints, but is nevertheless powerful enough to explain complex reanalysis data.

The best way to discover whether SLCM predictions might explain MacDonald's data would be to determine the behaviour of the SLCM simulation when faced with each of MacDonald's experimental items. Unfortunately, the small size of the SUSANNE corpus renders such an approach impossible – while the corpus provides sufficient lexical category co-occurrence statistics, many of the words used in the experiment are either rare or absent in the corpus. None of MacDonald's contrasting triples of sentences is constructed purely of words that occur with high frequency. We have therefore invented our own triple, which mirrors MacDonald's both in syntactic structure and in the relative frequency of the different verb usages, but is constructed entirely from comparatively high frequency words; example 7.8 contains

our sentences.

(7.8) a. The man *fought at the police* station fainted.
b. The man *held at the police* station fainted.
c. The man shown at the police station fainted.

In sentence 7.8c, the main verb is unambiguous in tense and transitivity. While there is temporary ambiguity in the logical structure of the sentence (cf. example 7.9), there is no relevant ambiguity at the lexical category level. We would therefore (trivially) expect an initial SLCM decision favouring a past participle reading and no internal reanalysis; this prediction tallies with MacDonald's results.

(7.9) The man shown the knife at the police station fainted.

However, both 7.8a and 7.8b do exhibit lexical category ambiguity. The decisions of the SLCM therefore depend on the relative frequencies of different readings – figure 7.4 shows the probabilities assigned to each reading by our model SLCM as each word is processed.

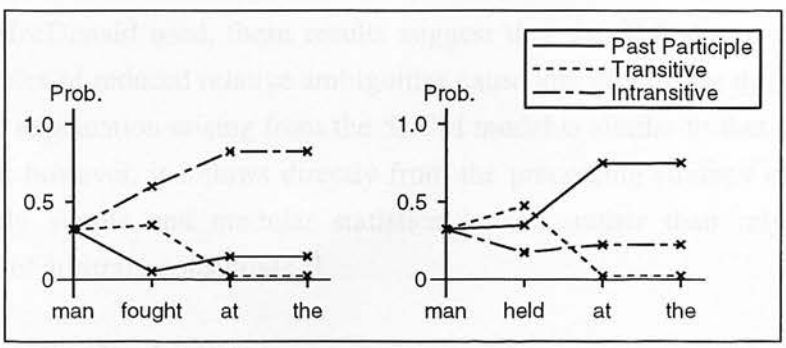

Figure 7.4: SLCM predictions for examples 7.8a and 7.8b

When processing sentence 7.8a, the initial decision of the SLCM is in favour of an intransitive reading for the ambiguous verb; the transitive reading comes second, and the past participle reading is deemed highly implausible. The following word ("at") is unambiguously a preposition and the sequence intransitive verb–preposition is not uncommon. There is therefore no internal reanalysis; the SLCM maintains its initial decision and later modules in the HSPM only force reanalysis when the disambiguating material ("fainted") is read. We would therefore predict that reduced relative constructions containing an intransitively biased verb may lead to human

processing difficulty when the disambiguating region is encountered; this prediction agrees with MacDonald's results.

It is interesting to note that the transitive verb reading, initially second choice, is assigned a far lower probability when the following preposition is read. This occurs because the sequence transitive verb–preposition is rare. Exactly the same effect occurs when processing sentence 7.8b – the initially preferred transitive analysis is assigned a lower probability when the following preposition is read, due to the infrequency with which these lexical categories co-occur. In this case, internal reanalysis does occur. The past participle reading, initially second choice, is fully compatible with a following preposition and therefore becomes the preferred analysis. We would predict that this reanalysis would lead to a slightly increased reading time for the ambiguous region. However, the preferred reading is now compatible with the disambiguation, and so we would also predict no processing difficulty when the disambiguating material is read. These predictions exactly coincide with MacDonald's experimental results.

While we have not been able to simulate the behaviour of the SLCM on the exact sentences MacDonald used, these results suggest that the SLCM can explain why some examples of reduced relative ambiguities cause less processing difficulties than others. The explanation arising from the SLCM model is similar to that proposed by MacDonald; however, it follows directly from the processing strategy employed by an extremely simple and modular statistical model, rather than relying on the supposition of arbitrary constraints.[53]

7.3.5 Alternative Explanations

In MacDonald's (1994) experiment, reanalysis takes place before syntactic or semantic disambiguation; it is syntactically permissible for a transitive verb to be immediately followed by a prepositional phrase. The standard assumption in the sentence processing literature is that reanalysis only occurs when absolute disambiguation is reached; however, this does not contradict the non-statistical

[53] We have not explicitly covered poor post-ambiguity constraints. However, a model such as the one we are proposing does lead to the prediction that poor post-ambiguity constraints will not be nearly so likely to lead to early reanalysis (due to the constraints on reanalysis outlined in section 7.2); this prediction agrees with MacDonald's results.

heuristic models outlined in section 2.3, as none of these models make any explicit claims about what triggers HSPM reanalysis (see Fodor & Ferreira, in press, for models that do). The initial decisions suggested by MacDonald's study are compatible with all these models; an explanation of the subsequent reanalysis effect is beyond their scope.

Statistical modular models, such as Tuning, may offer a fuller explanation; however, this would only be the case for a comparatively fine-grained variant of Tuning in which subcategorisation preferences guided parsing. While the Tuning authors make no claims about reanalysis, we can envisage an extension of the Tuning model in which reanalysis occurs when the probability of the current analysis goes below a predetermined threshold. Such a model may offer a realistic account of MacDonald's data.

Finally, interactive models can, of course, account for MacDonald's results. As argued in chapter 3, this is not surprising. What is in question is whether they can predict such data, and how many constraints must be stipulated for such models to afford an explanation.

7.3.6 Summary

MacDonald (1994) presented results that show that certain reduced relative ambiguities are easier to process than others. In particular, she demonstrated that when the ambiguous verb is biased towards a transitive reading, but is not immediately followed by a noun phrase, subjects appear to switch to the reduced relative reading prior to syntactic disambiguation.

While these results do not involve any ambiguity of grammatical class, we postulate that lexical categories may include more fine-grained information relating to the syntactic properties of an individual word. Within our framework, it makes little sense to include properties which cannot be predicted from the immediate lexical context; however, simple transitivity information does seem a natural candidate for augmented lexical categories. We can gain frequency information for lexical categories augmented with transitivity mark-up automatically from the SUSANNE corpus.

With such augmented tags, we find that the SLCM model provides a simple account of MacDonald's results, in terms of internal reanalysis. In order to explain this data, we do not need to augment the architecture of the SLCM, or propose arbitrary processing constraints.

7.4 Late Subcategorisation Information?

In section 7.3 we suggested that the SLCM might assign lexical categories that include transitivity information. However, as mentioned in section 3.5, Mitchell (1987) produced evidence suggesting that subcategorisation preferences are not available during initial syntactic structure building. This is clearly inconsistent with our proposed SLCM, in which transitivity is determined prior to parsing. In this section we reexamine Mitchell's evidence and show how internal reanalysis within a tag bigram SLCM which assigns tags that *do* include transitivity information may provide a novel account for this data. We also show that the SLCM offers an explanation for why later researchers (see Adams, Clifton and Mitchell, submitted) have failed to replicate Mitchell's result.

7.4.1 Mitchell's Data and Conclusions

Following on from earlier results suggesting that the initial decisions of the parser do make use of subcategorisation information (Mitchell and Holmes, 1985), Mitchell (1987) presented subjects with sentences in which obligatorily intransitive verbs are immediately followed by noun phrases, separated by an unmarked clause boundary. 7.10a exemplifies such a sentence; 7.10b is a control sentence in which the relevant verb may be transitive.

(7.10) a. After the child had sneezed the doctor prescribed a course of injections.
b. After the child had visited *the doctor* prescribed a course of injections.

In 7.10b, "visited the doctor" may be initially analysed as a transitive verb and noun phrase object. However, this analysis is not consistent with the following verb phrase. It would therefore seem likely that subjects experience processing difficulty when reading the disambiguating region.

In contrast, example 7.10a is unambiguous – the main verb must be intransitive (though see section 7.4.2). If verb subcategorisation information is used *at all* in

sentence processing, we would expect no processing difficulty to occur when reading the disambiguating region of this sentence.

Using a self-paced reading paradigm, Mitchell verified these predictions. Subjects did take significantly longer to read the disambiguating region of sentences such as 7.10b, compared to 7.10a. This suggests that subcategorisation preferences do affect sentence processing. However, Mitchell also discovered that the ambiguous region ("the doctor") takes longer to read in 7.10a than 7.10b. Surprisingly, it seems that some processing difficulty may occur in the unambiguous condition.[54]

If we equate these reading time increases with reanalysis, then Mitchell's data suggests that reanalysis occurs when reading the noun phrase. We know that after the noun phrase is processed, the HSPM has fixed on the correct (intransitive) analysis. Reanalysis during processing of the noun phrase therefore suggests that the initial attachment of the noun phrase is incorrect; the only likely candidate is that it is attached as the object of the preceding verb. However, this analysis should be ruled out by verb subcategorisation information.

Mitchell's hypothesis is that subcategorisation information is not available during structure building. Instead, there is a later 'checking' module that rejects analyses incompatible with such lexical preferences. In the example sentences, the parser initially proposes that the noun phrase is the object of the preceding verb. However, in the case of 7.10a, the checking module determines that such an analysis is inconsistent with the verb's subcategorisation feature, and so the parser is forced to offer an alternative analysis. This reanalysis is the cause of the reading time increase.

7.4.2 Criticisms of Mitchell's Results

Evidence supporting the late availability of subcategorisation information is incompatible with a number of proposed models of the HSPM, particularly interactive models in which all information sources are assumed to be available when making initial decisions. There have therefore been a number of attempts to offer alternative explanations for Mitchell's data.

[54] For further evidence of difficulty during the processing of syntactically unambiguous sentences see chapter 8.

It is not relevant to consider all criticisms here – instead we propose our own interpretation in section 7.4.3. However, a number of authors have been concerned that Mitchell's results might have been influenced by the way that the sentences were segmented for presentation. In order to address this and other criticisms, Adams, Clifton and Mitchell (submitted) attempted to replicate Mitchell's (1987) results using an eye-tracking paradigm. They failed – no significant reading time increase was found during processing of the post-verbal noun phrase. We discuss reasons why this might be the case in section 7.4.4.

Another criticism levelled at Mitchell's experiment is that many of his 'obligatorily intransitive' verbs may, in fact, occur in transitive constructions, as in 7.11:

(7.11) a. The child sneezed a big sneeze.
b. The diner sneezed the napkin right off the table.

Almost all 'intransitive' verbs can be used transitively. While such constructions are very rare, they suggest that the lexical entry for such verbs may include a transitive subcategorisation frame. In this case, an initial attachment of the following noun phrase as a verb object may be made even if subcategorisation information is available early.

7.4.3 SLCM Predictions

In section 7.3 we proposed that the SLCM assigns lexical categories that include transitivity information. However, such a model does not appear compatible with Mitchell's result, which supports the late availability of transitivity preferences.

It is clear that if we assume that the SLCM can only assign an intransitive reading to the verbs used in Mitchell's experiment, then we would predict no processing difficulty arising from lexical category decisions for sentences like 7.10a. However, in section 7.4.2 we noted that these verbs may occasionally be used transitively. Our modified SUSANNE corpus is not large enough to determine the frequency with which these verbs occur transitively; however, we can enter fake statistics into the SLCM simulation and discover its behaviour if these verbs occur transitively with a range of different frequencies. Figure 7.5 shows the SLCM behaviour if the frequency with which "sneezed" occurs as a transitive past participle is 1% of the

frequency with which it occurs intransitively.

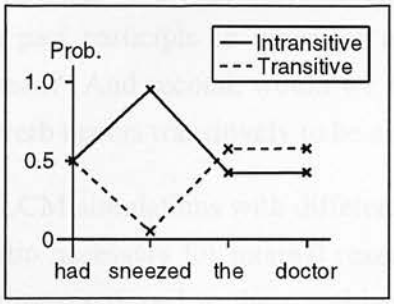

Figure 7.5: SLCM predictions for 7.10a (with 1% trans./intrans. ratio)

In this case, the initial decision of the SLCM favours the intransitive participle reading of "sneezed". However, when the following word is read, internal reanalysis occurs and the transitive reading is preferred. This is because of the high frequency of the sequence transitive verb–determiner, and the corresponding low frequency of the sequence intransitive verb–determiner; it is exactly analogous to the post-ambiguity constraint considered in section 7.3.

If we assume that an initial decision favouring the (correct) intransitive reading is discarded in favour of the incorrect transitive reading when the following determiner is read, then we can explain the processing difficulty observed by Mitchell when subjects read the ambiguous noun phrase. However, we would also predict a processing delay when the disambiguating material is read, as subjects should now favour the incorrect reading; Mitchell observed no such delay. We postulate that this is because subjects in fact perform two reanalyses when reading the noun phrase. First, an SLCM internal reanalysis switches from the intransitive to the transitive reading. However, once the head noun of the 'object' noun phrase is read, this analysis becomes semantically implausible, and so a second reanalysis is forced by later processing modules. At this point the preferred analysis switches back to the (correct) intransitive reading; as both reanalyses occur before the disambiguating material is read, we predict a significant reading time increase during processing of the ambiguous noun phrase, but no delay when integrating the disambiguating material.[55] These predictions exactly match Mitchell's results.

[55] The analysis presented here relies on the assumption that the delay caused by the second reanalysis occurs primarily during the processing of the disambiguating region. While this assumption seems reasonable, it neither entailed nor implied by our model.

Having shown that the SLCM may offer an account of Mitchell's data, two questions remain. Firstly, what is the minimum necessary frequency ratio of transitive and intransitive readings for a past participle in order for reanalysis to occur when a following noun phrase is read? And second, would we really expect the frequency with which an intransitive verb occurs transitively to be even this large?

By running a number of SLCM simulations with different word frequencies, we can determine the minimum ratio necessary for internal reanalysis to occur. Figure 7.6 shows the relative probabilities assigned to the transitive and intransitive tag paths when the determiner following the verb is read, plotted against the frequency ratio. From this data, it is easy to calculate that the frequency of the transitive usage must be at least 0.72% that of the intransitive usage in order for internal reanalysis to occur.

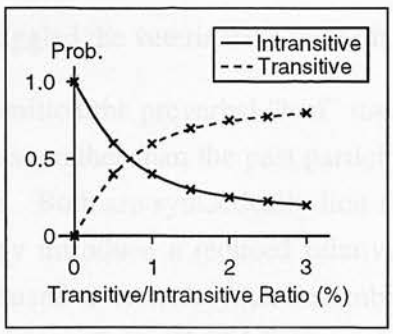

Figure 7.6: Relative probs of tag paths when det. is read (past participle)

Would we expect the transitive usage of a verb like "sneezed" to occur with even 0.72% the frequency of the intransitive usage? As we have no large corpora marked with transitivity information, there is no available evidence from which to determine an objective answer. Based on intuition, it seems likely that the answer is 'no'. However, this does not invalidate our account. So far, we have been assuming that the SLCM assigns tags without meaning; that is, the SLCM simulation does not 'know' that an intransitive verb is more like a transitive verb than it is like a determiner. This is an idealised situation; in a real (biological) model, we would expect some sort of excitatory links between similar lexical categories. The high activation gained by the intransitive verb category when processing sentence 7.10a may lead to an increased activation for the transitive verb category. In this case, we would expect the SLCM prediction outlined above to hold even if the frequency of

the transitive reading of Mitchell's verbs is (on average) somewhat less than 0.72% of their intransitive frequency.

7.4.4 Contrary Data

Adams, Clifton and Mitchell (submitted) attempted to replicate Mitchell's (1987) results using an eye-tracking paradigm. However, they found no significant processing delay when subjects read the noun phrase following the ambiguous verb. This new evidence suggests that Mitchell's original results may have been due to an artefact in the experimental design; however, it is still not clear what this artefact might have been. The SLCM offers an alternative explanation, in which the results obtained in both experiments are valid. This is possible as there was one crucial change in the experimental materials for the later study; these materials are exemplified in 7.12:

(7.12) After the dog struggled the veterinarian took off the muzzle.

Adams *et al.* (submitted) omitted the preverbal "had" used in the earlier study. This means that it is the past tense, rather than the past participle, form of the verb that is used in the correct analysis. Both are syntactically licit following a noun, as a post-nominal past participle may introduce a reduced relative clause. The omission of "had" has therefore introduced a lexical-syntactic ambiguity. However, we may assume that the initial decision of the SLCM favours the (far more common) past tense reading, and this ambiguity is therefore unproblematic.

What is more interesting is that, according to the statistics gained from the modified SUSANNE corpus, an intransitive past participle occurs far more frequently than a transitive one; there is no such difference for past tense verbs. This is shown in table 7.2.

Lexical Category	Frequency
Intransitive Past Participle	2794
Transitive Past Participle	490
Intransitive Past Tense	2012
Transitive Past Tense	1894

Table 7.2: Frequency of verb forms in SUSANNE corpus

Reconsider equation 5.3, reproduced below as 7.1, which is used to estimate word–tag co-occurrence frequencies:

$$P(W_i = w^x \mid T_i = t^y) \stackrel{\text{def}}{=} \frac{|\, W_i = w^x, T_i = t^y \,|}{|\, T_i = t^y \,|} \qquad (7.1)$$

As the divisor is the frequency of occurrence of a particular tag, this equation is biased towards infrequent tags. Initially, this appears counter intuitive. However, the equation is correct; if a tag is infrequent, then we would also expect a low frequency for the co-occurrence of any word and this tag. A small amount of evidence supporting a word and a frequent tag co-occurring may be treated as noise, but evidence for an infrequent tag should be given more weight.

As the transitive past participle occurs far less frequently than the intransitive, and we manipulated the *frequency* of the verb "sneezed" co-occurring with a particular tag, we would expect that the word–tag *probability* estimated for the transitive usage to be greater than that expected from its frequency. In contrast, the transitive past tense is not infrequent, and so the probability of "sneezed" being tagged as such gains no bonus. We would therefore expect that the a priori frequency of the transitive past tense usage of "sneezed" would have to be far greater in order for SLCM internal reanalysis to occur in Adams *et al.*'s experiment. This is confirmed by simulation; the equivalent of figure 7.6, this time using the sentence in 7.12, is shown in figure 7.7. The transitive past tense usage of "sneezed" would need to occur with 7.51% the frequency of the intransitive version in order for tagger internal reanalysis to occur.

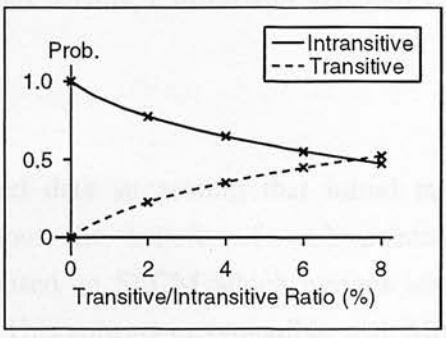

Figure 7.7: Relative probs of tag paths when det. is read (past tense)

It could be argued that the difference between figures 7.6 and 7.7 is due to the small

size of the corpus used, or to some inaccuracy in our automatic addition of transitivity marking to the SUSANNE corpus. The first is unlikely; as table 7.2 shows, all verbs forms were fairly frequent. In order to counter the second criticism, further evaluation of our automatic transitivity marking would be necessary; visual inspection suggests that it is unlikely that it is inaccurate enough to skew the data to the extent necessary to produce these results.

7.4.5 Alternative Explanations

As discussed in section 7.4.1, Mitchell's (1987) results are compatible with any model in which subcategorisation information is not used at the structure building stage. This may include the Garden Path theory and does include a coarse-grained variant of the Tuning Hypothesis. It does not include Construal or Generalised Theta Attachment; in both these models, thematic and subcategorisation preferences play a central role in argument attachment. Referential Support is unpredictive with respect to this ambiguity.

Interactive models may be able to explain Mitchell's (1987) data, by proposing, as we have for the SLCM, that the preferred analysis changes twice. However, they, in common with the Garden Path theory and the Tuning Hypothesis, would appear to predict the same results for Adams, Clifton and Mitchell's materials. The fact that the SLCM prediction differs for these materials arises from the way lexical category probabilities are calculated; the fact that the data appears to support this prediction offers strong support for the proposed SLCM, or some other statistical model that uses similar probabilities. However, this support depends on whether we believe that both Mitchell's (1987) and Adams, Clifton and Mitchell's (submitted) data are free from artefactual effects.

7.4.6 Summary

Mitchell (1987) produced data suggesting that initial structural decisions in the HSPM are made without the benefit of verb subcategorisation information. However, we have proposed an SLCM which assigns lexical categories including transitivity information. This appears incompatible with Mitchell's results.

In this section we showed that this quandary may be explained by SLCM internal reanalysis. This is interesting because, in this case, the SLCM revises an initially

correct decision to prefer an incorrect one; such behaviour does not occur in many processing models and is not often considered in the literature.

Adams, Clifton and Mitchell (submitted) tried to replicate Mitchell's (1987) data, and failed. We showed that a small difference between the experimental materials means that the SLCM does not predict internal reanalysis in the later study; the SLCM predictions therefore match the experimental data for both studies. No other model can explain both sets of data.

7.5 Are we Simulating Syntax?

In this chapter, we have suggested that the SLCM may assign verb tags that include transitivity information. We made this extension because the SLCM offers a low cost and accurate mechanism for doing this. The fact that an SLCM that assigns such augmented tags explains a range of experimental results both simply and elegantly is a powerful argument in favour of both the SLCM and the inclusion of transitivity information in its remit. However, there is another possibility. It could be that the reanalysis behaviour we have attributed to the SLCM actually occurs within a statistical parser module; the close relation between sequence and structure allows the SLCM to capture some of the regularities inherent in syntactic frequency information and therefore to predict some of the effects we might expect from a statistical parser. In other words, the SLCM simply offers an approximation of the behaviour of such a parser.

In chapter 8, we present experimental results that suggest that the SLCM is distinct from the parser; however, this does not demonstrate that effects that rely on syntactic features (such as transitivity) do not emanate from the parser. Our problem is therefore one of grain – what granularity of lexical representation is assigned by the SLCM and what features are best left to later modules. We had a stab at answering this question in section 7.3.2; in this section we discuss some data which might be explained if we augmented the SLCM output further, and use this data to exemplify where the line should be drawn.

7.5.1 "That"-Preference Data

Trueswell, Tanenhaus and Kello (1993) investigated the behaviour of the HSPM

when faced with verbs that are strongly biased towards taking a sentence complement (S-bias verbs). They demonstrated that readers may experience difficulty when such verbs are followed by a reduced sentence complement, as in example 7.13a, but not when they are followed by a full sentence complement, as in 7.13b.

(7.13) a. The student implied the book was stolen.
b. The student implied that the book was stolen.

This result appears similar to Mitchell's (1987) finding that verb subcategorisation information is apparently ignored in making initial attachment decisions (see section 7.4). If verb subcategorisation information is ignored, then we would expect initial attachment of the post-verbal NP in 7.13a as a direct object; the processing difficulty experienced by readers would then result from reanalysis when subcategorisation information becomes available.

However, Trueswell *et al.* (1993) also discovered that the magnitude of the difficulty readers have with noun phrases following S-bias verbs depends on the "that"-preference of the individual verb; that is, there is a correlation between the reading time increase at the noun phrase and the frequency with which the verb occurs with a following complementiser, as measured by a sentence completion study. The "that"-preference was in turn shown to be correlated to the frequency of the verb; Juliano and Tanenhaus (1993, experiment 3) replicated Trueswell *at al.*'s result and showed a direct correlation between the reading time increase and the verb frequency (see also Juliano and Tanenhaus, 1994).

7.5.2 SLCM Predictions

Can this result be explained by the tag bigram SLCM? Not as it stands, since no distinction is made between verbs taking a sentence complement (S-verbs) and verbs taking a direct object noun phrase (NP-verbs); the SLCM is compatible with the "that"-preference results, but it is up to a later parsing module to offer an explanation for why they occur.

However, if we determine that the SLCM assigns a different tag for S and NP verbs, then the SLCM predictions differ. We would expect an initial decision favouring an

S-verb tag when an S-bias verb such as "implied" is read. However, S-verbs are less frequently followed by determiners than NP-verbs are; SLCM reanalysis may therefore occur, but only if the evidence supporting the initial decision is not particularly strong. In the case of an infrequent S-verb, this evidence will be weak, and the SLCM is quite likely to revise its analysis to favour the incorrect NP-verb tag. We would therefore expect some processing delay when reading the material immediately following the verb. In contrast, when the verb is frequent, the initial decision was made on the basis of strong evidence; we would therefore expect no SLCM internal reanalysis and no processing delay. These predictions agree with the results reported by Trueswell *et al.* (1993) and Juliano and Tanenhaus (1993).

7.5.3 Going too far?

Having shown that the SLCM could account for "that"-preference effects, we argue that these effects should, in fact, be attributed to a statistical parser. While it is useful to allow the SLCM to differentiate between S-verbs and NP-verbs in order to explain this data, augmenting the SLCM in this way may be going too far. It appears that it is syntactic rather than lexical context that best arbitrates when there is ambiguity between these two verb types. Consider, for example, the sentences in 7.14:

(7.14) a. The man knew the stupid old rugby player well.
b. The man knew the stupid old rugby player was fast.

In 7.14a, "knew" is disambiguated as an NP-verb, whereas in 7.14b it is a S-verb. However, there is nothing in the immediate lexical context of "knew" that helps to disambiguate. In section 4.2, we characterised lexical ambiguities as tending "to be disambiguated locally, normally within the phrase in which they occur". The ambiguity between S-verbs and NP-verbs appears to pattern better with syntactic ambiguities, which "regularly span phrasal nodes". We are therefore reluctant to suggest that they are the province of the SLCM.

A similar argument may be made for verb transistivity information; this is a matter for future research. However, the simple transitive/intransitive distinction is often (if not almost always) disambiguated extremely locally; we would therefore expect the SLCM to be fairly reliable when determining transitivity.[56] Given that such a simple

[56] Subsequent work by the author while at Sharp Laboratories of Europe suggests that an HMM tagger

model can explain the range of data put forward in this chapter, it is appealing to suggest that transitivity is the province of the SLCM.

7.5.4 Summary and Conclusions

Many of the ambiguities we have attributed to the breakdown and repair characteristics of the SLCM may also be explained by the operation of a statistical parser. In this case, our SLCM results could just be a crude simulation of the behaviour of this parser. However, the simplicity and elegance of the account offered by the SLCM makes it an appealing model; in chapter 8 we present some novel experimental evidence which suggests that there is a division between lexical category disambiguation and syntax. However, it could still be that by allowing the SLCM to assign comparatively fine-grained tags, we are encroaching on the role of the parser.

We highlighted the danger in this section by showing that the SLCM could explain Trueswell, Tanenhaus and Kello's (1993) results concerning the "that"-preference of S-bias verbs, provided we allow the tags assigned by the SLCM to include information about whether the verb takes a sentence or a noun phrase complement. However, such ambiguities are best resolved at the syntactic level; they do not pattern with the other lexical category ambiguities we have considered in that they are not normally resolved locally. To include them would undermine our reasons, stated in section 4.2, for suggesting independent treatment for lexical category ambiguities. There is, then, a danger of making our lexical categories too fine-grained.

The comparatively coarse-grained transitivity information we have been considering throughout this chapter is another matter. The simplicity with which the SLCM can explain some established results when its output is augmented with transitivity information is appealing, and verb transitivity is often disambiguated extremely locally. The inclusion of this information is therefore warranted until the question is decided by further evidence; however, it should be treated with a degree of caution.

---

may assign verb transitivity information with reasonable accuracy, but does not do a good job of distinguishing verbs taking an NP-complement and those taking an S-complement.

7.6 Conclusions

SLCM internal reanalysis (or 'repair') is both predictive and accurate when tested against existing experimental data. However, the particular predictions depend on our definition of lexical category. In this chapter we have suggested that the lexical categories assigned by the SLCM might include transitivity information; with this augmentation, the SLCM offers simple explanations for data previously thought to rely on the interaction of complex constraints.

However, while we have shown that the SLCM offers an attractive model for predicting the behaviour of the HSPM when faced with lexical category ambiguities, none of the evidence so far presented has conclusively shown that the initial decision of the HSPM when faced with such ambiguities are dependent on experience-based statistics, nor that lexical category decisions are made independent of syntax. All the existing evidence we have examined has proved compatible with non-statistical models and combined models. In the next chapter, we present novel experimental evidence that suggests that a separate, statistical, lexical category decision module does exist.

8: Experimental Evidence

8.1 Introduction

There are a number of related proposals put forward in this thesis. At the most general level we hypothesize that the HSPM is both modular and statistical; this position is captured by the Modular Statistical Hypothesis (MSH), presented in chapter 3. In section 4.2 we argued for the existence of a modular statistical process which arbitrates lexical category ambiguities; by definition, any model including such a process falls within the scope of the MSH. Finally, in section 4.5, we introduced a detailed architecture for this module, which we call the SLCM.

In the previous two chapters, we have been concerned with the final, most detailed, hypothesis. We have presupposed the existence of a separate SLCM and presented evidence supporting the architecture we have outlined. The simplicity of the model, its predictiveness, and its explanatory power have proved strong reasons to suppose both that this architecture is largely correct, and that such a module may exist within the HSPM. However, the data presented is also compatible with other models. Therefore, the analysis in the previous chapters does not directly address the mid-level proposal – we have not yet provided direct evidence that a separate and statistical lexical category disambiguation module exists. In this chapter, we step back from the details of the model and experimentally investigate the existence of such a module. By implication, evidence for such a module would also support our most fundamental hypothesis – the MSH.

In order to investigate the existence of the SLCM, we postulate two hypotheses:

The Statistical Lexical Category Hypothesis (SLCH):

Initial lexical category decisions are made on the basis of frequency-based statistics.

The Modular Lexical Category Hypothesis (MLCH):

Lexical category decisions are made by a pre-syntactic module.

Experiment 1 (presented in section 8.2) investigates the SLCH and provides strong evidence that there is a statistical basis to lexical category decisions. Experiment 2 (presented in section 8.3) explores the MLCH, and provides initial evidence supporting a model in which lexical category disambiguation occurs in a pre-syntactic module.

8.2 Experiment 1

In order to investigate the SLCH, we must consider a lexical category ambiguity in which a statistical measure may provide a strong bias in favour of a particular reading. If such a bias can only arise from a frequency-based account, then evidence demonstrating that this bias affects initial decisions would strongly support the SLCH and be hard to reconcile with a non-statistical model. In contrast, if the bias is strong and no conflicting biases may affect the decision procedure, then failure to find such evidence would render unlikely any statistical model in which this measure is dominant.

We have already considered an ambiguity which, according to our SLCM model, is dominated by a single statistical measure. This is the noun–verb ambiguity, first considered by Frazier and Rayner (1987) and then explored by MacDonald (1993). In sections 4.4 and 6.2 we gave reasons to doubt the former's empirical data and the latter's explanation. Further exploring this ambiguity may provide not only evidence for or against the SLCH, but also shed light on the correct analysis of MacDonald's experiment.

We noted that all MacDonald's ambiguous words were biased towards a noun reading. We can test the SLCH by comparing the HSPM behaviour on words biased towards both noun and verb readings, with disambiguations favouring either reading. If the initial decision of the HSPM is determined by the statistical bias, then we would expect processing delay when the disambiguating region is inconsistent with the bias, and no such delay when bias and disambiguation are in agreement. In contrast, no non-statistical model could account for an initial decision apparently determined by frequency-based bias.[57]

[57] A non-statistical model might explain such data if a correlation could be shown between some lexical feature and bias for the experimental materials. While the verb readings of the ambiguous

8.2.1 Method

Subjects

32 students at the University of Edinburgh were paid three pounds each for their participation in this experiment and a (following) unrelated questionnaire study. The online part of the experiment lasted for about 25 mins. All subjects were native speakers of English.

Materials

12 pairs of ambiguous words were selected; one of each pair occurred more frequently as a noun and the other as a verb (according to data collated from the BNC). A two word noun compound (the critical region – c_1 and c_2) was invented for each of the 24 ambiguous words.

Disambiguating and introductory material were then constructed for each pair. For each item, two disambiguations favouring the noun reading and two favouring the verb reading were generated. Noun disambiguations always began with "are" or "were", as these are unambiguously tensed active verbs, and therefore incompatible with an analysis in which the previous word is also a tensed active verb. Verb disambiguations begun with "the", which unambiguously introduces a noun phrase and is therefore only compatible with the verb reading of the ambiguous word.[58]

In order to form semantically viable sentences, noun and verb disambiguations were not identical across pairs; however, at least the first two words of the disambiguating region (d_1 and d_2) were the same for each pair. Introductory material was introduced to ensure that the subject had a 'run up' to the critical region. This also varied across a pair, though it always had a similar syntactic structure and never contained any words that were semantically related to any meaning of the ambiguous word. With hindsight, it would have been ideal to keep introductory material identical across a pair; however, this is unlikely to affect the outcome of the experiment.

---

words we used do not all have the same subcategorisation frame, they can all be used transitively. This rules out a Generalised Theta Attachment account, as considered in section 4.3.1.

[58] An alternative reading in which "the" introduces the subject of a reduced relative is possible but extremely rare. It therefore seems unlikely that the sentence processor would prefer such an analysis, and it can safely be ignored.

A sample sentence from each condition is given as 8.1. Full experimental materials can be found in appendix A.

(8.1) a. The woman said that the german *makes* the beer she likes best.
b. The woman said that the german *makes* are cheaper than the rest.
c. The foreman knows that the warehouse *prices* the beer very modestly.
d. The foreman knows that the warehouse *prices* are cheaper than the others.

Sentences 8.1a and 8.1b represent the verb bias condition (as "makes" occurs more frequently as a verb than a noun). Sentences 8.1c and 8.1d represent the noun bias condition. In the verb disambiguation conditions (8.1a and 8.1c), d_1 and d_2 are identical ("the beer"), so comparison of reading times between these two conditions is certainly valid. The same is true for the noun disambiguation conditions (8.1b and 8.1d).

As two different sets of four materials were constructed for each pair of ambiguous words, a total of 96 experimental items were created. These items were divided into four lists, in such a way that each list contained exactly one occurrence of each ambiguous word, and only one item from each set of four materials. In other words, no list contained items with the same ambiguous word or related disambiguating material.

80 filler items were also created. These fillers were syntactically unrelated to the experimental materials, and none of the fillers contained any of the ambiguous words. However, the fillers were of a similar length to the experimental items, and had a similar 'style' of content. In other words, it was unlikely that a linguistically naïve subject would be able to differentiate between experimental and filler materials.

Finally, 8 comprehension questions were invented for each list of experimental materials, and a further 26 for the filler items. These questions required an answer of "yes" or "no" and referred to some aspect of the accompanying sentence. Exactly half the questions in each set required the answer "yes", and the other half "no". Example questions for the sentence in 8.1 are given in 8.2.

(8.2) a. Does the woman like beer?
b. Does the woman think that the german products are expensive?
c. Is the beer expensive?
d. Is the warehouse overpriced?

Each subject saw the experimental materials in one list together with all the fillers and associated questions. The order of the sentences and fillers was random; each subject saw a different ordering.

Plausibility and Frequency

In order to avoid using noun compounds that could be considered single lexical items (see section 6.2.2), all compounds chosen for this experiment occurred with very low frequency in the BNC. Noun and verb biased examples were balanced for frequency of noun compound reading. These precautions also avoid the danger that, if word co-occurrence is also dominant (as in the combined bigram SLCM), it might affect the decision process. In other words, we would not expect that any difference in the early decision of the HSPM could be attributed to the frequency with which the noun compound occurred in subjects' previous linguistic experience; 'supportive' and 'unsupportive' bias (MacDonald, 1993; see section 6.2) would not be expected to affect the decision procedure in this study.

If plausibility and frequency are highly correlated, then we would expect that our materials are also balanced for plausibility of the noun compound reading. In this case, any results of our experiment are unlikely to arise from plausibility effects. If the relationship between plausibility and frequency is more complex, then we should perform a pretest using naïve subjects to estimate the plausibility of each noun compound. However, subjects' judgements in such a pretest would be influenced by their initial preference when assigning a reading to the ambiguous word (c_2). In other words, if we accept a frequency-based model of initial lexical category or syntactic decisions, we would not expect the results of a pretest to reflect plausibility independent of frequency bias. There is no clear way to determine the plausibility of these examples, or any clear definition of plausibility, divorced from frequency.

We therefore assume that plausibility and word co-occurrence frequency can be equated, and so our materials are also balanced for noun compound plausibility. As a

further check, two experts looked over the materials and determined that all the noun compounds could be considered plausible but unusual.[59]

Finally, it proved impossible within the constraints of the experimental design to also balance the items for bias. This was due to the comparative rarity of verb bias ambiguous words. The noun bias items were, on average, more strongly biased than the verb bias materials; strongly verb biased ambiguous words are comparatively rare. As our predictions (below) depend only on the direction and not the strength of the bias, this was considered unimportant.

Procedure

Each subject was tested individually. The subjects read sentences in a moving window display (Just, Carpenter & Woolley, 1982). Each non-whitespace character in a sentence initially appeared as a dash on the computer screen. Subjects were instructed that they should press a key to see each word of the sentence in turn. The first keypress revealed the first word; subsequent keypresses revealed the next word in sequence, and reverted the previous word to dashes. Following the keypress after the last word in each sentence, the comprehension question was displayed, if there was one associated with the sentence. The subject pressed one of two keys to indicate an answer of "yes" or "no"; no feedback was given. If there wasn't a comprehension question associated with the sentence, "No Question" was displayed in place of the question.

After completion of the experiment, subjects were given a short interview in which they were encouraged to expound on whether they had had conscious difficulty with any part of the experiment. The purpose of the experiment was then explained to them.

A moving window display paradigm was chosen for a number of reasons. Firstly, we required reading times for individual words. Other self-paced paradigms can only provide reading times for multi-word segments of the sentence and we make predictions (below) for effects on the ambiguous word itself. While alternative methodologies such as eye-tracking do allow reading times to be calculated for individual words, such procedures require expensive equipment and are time

[59] Thank you to Matt Crocker and Chuck Clifton for their assistance in this capacity.

consuming. Finally, MacDonald's (1993) results were obtained using a moving window paradigm, and she observed effects similar to that those we anticipate; it therefore seems likely that the methodology is sensitive enough for our purposes.

The software package used in the experiment, running on a PC under MS-DOS, was written by Chuck Clifton. Without his generous donation of software and helpful advice, it is unlikely that this experiment would have been successfully completed.

Predictions and Analyses

If the model we proposed in chapter 4 (and therefore the SLCH) is correct, then we would expect the initial decisions of the HSPM to be determined solely by the frequency-based bias of the ambiguous word (there is very little contextual bias as the noun–noun and noun–verb bigrams occur with similar frequency). In all experimental conditions, this initial decision appears both syntactically and semantically congruent when the word is read. Reanalysis will only occur in the disambiguating region, when the initial decision is not compatible with the disambiguating material. We therefore predict a reading time increase on d_1 in the noun bias verb disambiguation (NV) and verb bias noun disambiguation (VN) conditions, compared to the NN and VV conditions. As we would not anticipate internal or external reanalysis or complexity effects while reading the ambiguous word (c_2), we expect no difference in reading times across all four conditions for this word.

MacDonald's (1993) analysis of her experimental results also rests on frequency effects. We therefore presume that she would also predict reading time increases at the start of the disambiguating region in the NV and VN conditions. However, she also postulated a 'reverse ambiguity' effect to explain the reading time increase she observed on c_2 in three of her four conditions; she suggested that this increased reading time could be attributed to the syntactic complexity involved in verb phrase formation. In the current study, verb phrase structure will be created in the VN and VV conditions; MacDonald's account therefore entails a reading time increase on the ambiguous word in these conditions.

Any model where decisions are motivated purely by syntactic structure should predict no difference in initial decision across the four experimental conditions. In

our own analysis of the Garden Path theory (section 4.3.1) we suggested that it predicts an initial decision favouring the noun reading. We would therefore expect reading time increase in the NV and VV conditions. Frazier and Rayner (1987) proposed a slightly different analysis (see section 4.3.1); they would anticipate reading time increase in the NN and VN conditions. Whatever the correct analysis, a main effect of reading time dependent purely on disambiguation, and no effect of bias, would be strong evidence for a model of lexical category disambiguation in which statistical mechanisms do not play a central role. In contrast, any effect of bias on the initial decision can be seen as strong evidence against a non-statistical model.

Finally, the delay strategy suggests yet another prediction. We would expect reduced reading time for the ambiguous word (compared to some unambiguous equivalent) across all four conditions, while syntactic processing is suspended. As all reading times are equally reduced, we would anticipate no reading time difference across conditions. In the disambiguating region we would expect increased reading time (compared to the mythical unambiguous control) while the processor caught up – again, across all four conditions. The delay strategy therefore leads to the prediction of no significant reading time differences between conditions in either the ambiguous or disambiguating regions.

8.2.2 Results

All sentences were included in reading time analysis, ignoring whether the comprehension question following the sentence was answered correctly; it was assumed that sentences which had not been correctly understood constituted random noise. The reading times were adjusted for word length, following a procedure described by Ferreira and Clifton (1986): for each subject, a linear regression equation was computed to predict reading time from word length, based on reading times recorded for all words in the experiment.[60] Length-adjusted reading times for each word were then calculated by subtracting the expected time from the actual time. In this way, it was possible to compare reading times for words of different lengths.

[60] Subjects tended to be very slow to press a key after the last word of each sentence; reading time for this word was therefore omitted in calculating the regression equation.

Length-adjusted average reading times for both words in the critical region (c_1 and c_2) and for the first two words in the disambiguating region (d_1 and d_2) are depicted in figure 8.1.

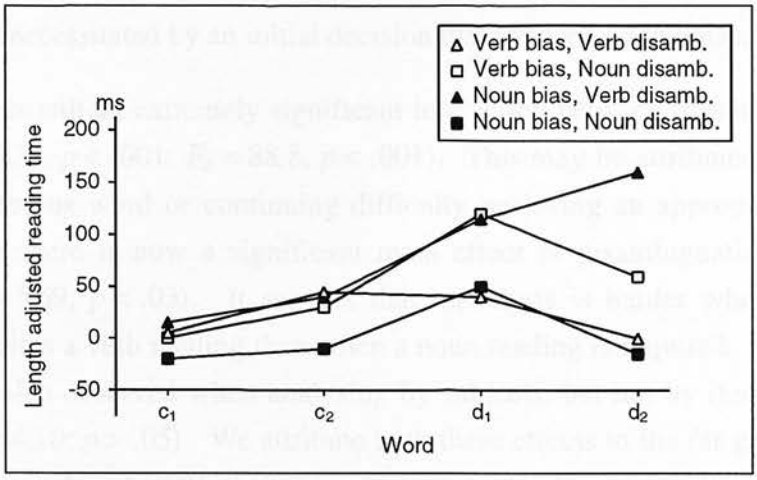

Figure 8.1: Length-adjusted reading times for experiment 1

On d_1, we found no main effect of either bias or disambiguation. However, there is a highly significant interaction between bias and disambiguation, when analysed both by subjects ($F_1 = 8.05$, $p < .01$) and by materials ($F_2 = 27.99$, $p < .001$). A planned comparison of means for d_1 in the verb disambiguation conditions (NV and VV) revealed a very significant difference in reading times ($F_1 = 8.27$, $p < .01$; $F_2 = 10.86$, $p < .01$) and a similar comparison between the noun disambiguation conditions (NN and VN) also revealed a significant difference ($F_1 = 4.72$, $p < .05$; $F_2 = 7.46$, $p < .02$). This effect can be attributed to the comparatively large reading times for the NV and VN conditions. The reading times for d_1 are shown on their own in figure 8.2.

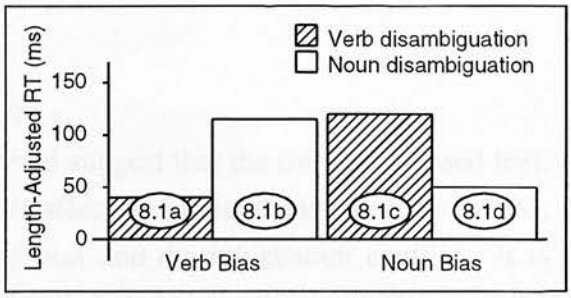

Figure 8.2: Length-adjusted reading times for d_1

These results suggest that the initial decision of the HSPM is affected by the frequency-based lexical category bias of an ambiguous word. Subjects appear to experience processing difficulty when the disambiguation is in conflict with the frequency bias, but not when they agree. The processing difficulty can be attributed to reanalysis necessitated by an initial decision that is based on the bias.

On d_2, there is still an extremely significant interaction between disambiguation and bias ($F_1 = 30.37$, $p < .001$; $F_2 = 88.8$, $p < .001$). This may be attributed to spill over from the previous word or continuing difficulty retrieving an appropriate analysis. Interestingly, there is now a significant main effect of disambiguation ($F_1 = 6.46$, $p < .02$; $F_2 = 5.59$, $p < .03$). It appears that reanalysis is harder when the correct analysis requires a verb reading than when a noun reading is required. A main effect of bias was also observed when analysing by subjects, but not by items ($F_1 = 6.46$, $p < .02$; $F_2 = 4.10$; $p > .05$). We attribute both these effects to the far greater reading time for this word in the NV condition. These two results combined suggest that for (at least) some sentences, subjects had more difficulty recovering when their initial decision favoured a noun reading, but disambiguation required a verb reading. We discuss why such an effect may occur in the next section.

Finally, on the ambiguous word itself (c_2) we found no significant effects. We would not expect disambiguation to influence the decisions of the sentence processor at this point, as the disambiguating material has not yet been read. However, an effect of bias is entailed by MacDonald's (1993) account; we found no such effect ($F_1 = 1.62$, $p > .21$; $F_2 = 1.81$, $p > .19$). Such a negative result does not prove that bias does not influence reading times for this word, but it suggests that we should prefer explanations that do not involve postulating such an unsupported effect.

8.2.3 Discussion

Support for the SLCH

The results we observed suggest that the frequency-based lexical category bias of an ambiguous word does affect the initial decision of the HSPM. The average reading time is greater when bias and disambiguation conflict. It is highly plausible that these increases can be attributed to reanalysis; in this case, the initial decision of the HSPM favours the most frequent category for an ambiguous word. It appears that

lexical category decisions *are* guided by frequency information.

The main effects observed on d_2 can be attributed to greater average bias of the noun bias items. We might expect that where the bias is greater, it would be more difficult to recover the less frequent alternative;[61] in this case recovering the verb reading causes particular processing difficulty when the noun reading is much more frequent. The effects observed on d_2 can therefore be seen as circumstantial evidence for a model in which decisions are guided by true frequency information, rather than just frequency rankings – in the latter case, the *size* of the bias towards a particular reading would not be expected to affect processing.

These results constitute strong and direct evidence for the SLCH; they are also consistent with the detailed SLCM architecture we explored in chapters 4, 6 and 7. Moreover, the MSH gains support from this evidence – in order to explain this data, any modular model must be at least partially statistical, and therefore fall within the province of the MSH.

Constraint-Based Models

These results are also compatible with constraint-based models, which similarly possess statistical decision procedures. However, MacDonald's (1993) explanation for the reading time increases she discovered on the ambiguous word entails greater reading times for the ambiguous word in the verb bias conditions of the current study. We found no such effect. While such a null result does not constitute strong evidence (and may be attributed to the fact that we used fewer subjects), it still suggests that our model, which makes no unsupported prediction, should be preferred to MacDonald's. In experiment 2 we discover that in cases where the SLCM does predict processing difficulty on the ambiguous word we do get a robust effect even with comparatively few subjects.

In general, it seems that constraint-based models may offer an account of the results we have presented here, provided they do not postulate complexity effects on the ambiguous word. However, the only published constraint-based explanation of MacDonald's (1993) results does postulate just such an effect. While different

[61] This behaviour might arise from the manner in which the model is implemented in a real (biological) system; however, the more abstract mathematical model we have presented does not predict such behaviour.

constraint-based models can individually account for our results and for MacDonald's, there is (as yet) no uniform account for both. It is necessary to tweak the parameters of the model in order to get the correct 'predictions' in each case. In contrast, the SLCM predicts both MacDonald's (1993) results and the data we have presented here without the need for any adjustment. It is also a far simpler, more precise, and more predictive model.

Non-Statistical Models

Non-statistical models do not offer an explanation for the results reported here. Those outlined in chapter 2 are either not predictive for this ambiguity (Generalised Theta Attachment) or predict uniform initial decisions irrespective of bias (Garden Path theory and Construal). In the former case the model is not falsified by the data, but it is also not sufficient. The latter class of models predict a main effect of disambiguation on d_1 and no interaction between bias and disambiguation. We found exactly the opposite results; these models are therefore inappropriate for modelling lexical category disambiguation.

The delay strategy is also incompatible with our results. It does not predict any effects in the ambiguous region, and we found none. However, it also fails to predict the reading time differences we observed when the disambiguating material was read. The data presented here therefore constitutes further evidence (see MacDonald, 1993) for the untenability of the delay strategy.

An Alternative Explanation

It may be suggested that the ambiguous words used in our experimental sentences are too strongly biased, and the HSPM treats them as unambiguous. In the case of the noun bias items, the bias is strong. We might suggest that the HSPM is initially unaware of the less frequent alternative reading; it is only when the disambiguating material is read that either conscious or special purpose processes recover the less frequent analysis. If the dispreferred reading is very infrequent, we might suggest that it is treated as a neologism.

It seems unlikely that conscious processes are involved. Abnormally long reading times occurred no more frequently at the start of the disambiguating region than at other positions in the sentence. On average, subjects answered over 7 out of the 8

comprehension questions following experimental items correctly. During the post experiment interview, none reported finding the sentences in any way difficult to read. When the ambiguity was explained to them, none recalled that they had seen sentences of this type. If subjects did suffer from conscious difficulty, they tended to recover the correct analysis extremely quickly and had no memory of such processing problems. Unconscious reanalysis appears a more plausible explanation.

A special purpose unconscious process must be able to recover both the correct part-of-speech and the meaning of the ambiguous word. While it is possible to create a new verb from a noun or visa versa, such neologisms tend to be morphologically marked, as in 8.3:

(8.3) He verbized the noun.

Further, none of the words used in the experiment was extremely infrequent in either usage – it is likely that the subject would have encountered this usage frequently in their prior linguistic experience. It is therefore probable that the less frequent usage is retrieved rather than constructed using special purpose procedures for interpreting neologisms. If this is the case, then frequency information is used to determine which stored representations are available for initial syntactic processing. While differing reanalysis strategies may be suggested, a model in which a special purpose process recovers less frequent alternatives is consistent with the SLCH, and does not significantly differ from the proposed SLCM.

8.2.4 Summary

The results of this experiment strongly suggest that the mechanisms responsible for initial lexical category decisions are (at least partially) guided by frequency-based mechanisms. In other words, the observed behaviour of the HSPM strongly supports the Statistical Lexical Category Hypothesis, presented in section 8.1.

These results are also consistent with the detailed SLCM architecture we proposed in section 4.5 and explored in chapters 6 and 7. While a number of other statistical models are compatible with this experiment, the SLCM is a very simple model that accurately predicted the effects observed. The experiment also renders unlikely MacDonald's (1993) analysis of her experimental results, and therefore gives

credence to our alternative proposal. The fact that we have found support for our SLCM based analysis of the earlier result provides further evidence for the proposed SLCM architecture.

Finally, these results offer some evidence for the MSH in that they suggest that any modular model must be partially statistical. In experiment 2 we attempt to determine whether lexical category disambiguation is a process in its own right; evidence supporting such a position would also further strengthen the case for the MSH.

8.3 Experiment 2

Experiment 1 provided strong evidence for the SLCH. However, this evidence is compatible with models in which lexical category decisions are made by a statistical parser, as well as with our proposed SLCM. In other words, we have not yet produced any evidence for or against the MLCH.

In order to determine whether statistical lexical category decisions are made prior to parsing, we must look at HSPM behaviour when syntactic constraints determine the outcome of a lexical category decision as soon as the word is read[62], but a pre-syntactic lexical category module would make an initial decision favouring the syntactically illicit reading. In experiment 1, we determined the HSPM behaviour when faced with noun–verb ambiguities. It therefore seems sensible to adjust the sentences used in experiment 1 to create syntactically unambiguous versions, and observe the HSPM behaviour when these are read.

MacDonald (1993) already observed the behaviour of the HSPM for this ambiguity, using noun biased ambiguous words that must, to obey syntactic constraints, be interpreted as verbs. We could complete this experiment by looking at the HSPM behaviour with verb biased ambiguous words; however, the effects we would then predict on the ambiguous word would not differ from those entailed by MacDonald's account, with the exception of the ambiguous verb bias condition which we have already explored in experiment 1. However, if our unambiguous sentences are only compatible with a noun reading, we would predict different effects on c_2 from those entailed by MacDonald's account. We therefore constructed such materials for this experiment.

[62] In chapter 4 we called these NSC ambiguities.

8.3.1 Method

Subjects

32 members of the University of Edinburgh community were paid three pounds each for their participation in this experiment and a (following) unrelated questionnaire study. The online part of the experiment lasted for about 25 mins. All subjects were native speakers of English and none had taken part in experiment 1.

Materials

The materials were based on those constructed for experiment 1. Only the noun disambiguation materials were used; for each of these materials a new unambiguous version was constructed. This was achieved by removing the final "-s" from the ambiguous word and changing the disambiguating region to agree with a singular noun. Some other small changes also proved necessary.

A sample sentence for each condition is shown in 8.4. Full experimental materials can be found in appendix B.

(8.4) a. The woman said that the german *makes* are cheaper than the rest.
b. The woman said that the german *make* is cheaper than the rest.
c. The foreman knows that the warehouse *prices* are cheaper than the others.
d. The foreman knows that the warehouse *price* is cheaper than the others.

8.4a is identical to 8.1b. The sentence is temporarily ambiguous until the disambiguating region is encountered; c_2 is biased towards the verb reading but the disambiguation is only compatible with the noun reading (AV condition). In contrast, 8.4b is unambiguous (UV condition). In this case, "make" may not be a verb as it would then disagree, in number, with the potential subject "german"; however, there is no disagreement if a noun reading is chosen and so this is the only permissible reading (at the syntactic level).

Sentence 8.4c is identical to 8.1d; it is temporarily ambiguous and both biased and disambiguated towards the noun reading (AN condition). The unambiguous version (8.4d) is again constrained by number agreement to be compatible only with the noun reading (UN condition).

These items were again divided into four lists, in such a way that no list contained items with the same ambiguous word or related disambiguating material. The same 80 fillers were used as for experiment 1. While subjects saw no items similar to the experimental items but disambiguated the other way, the large amount of fillers and random presentation means that they are unlikely to have developed strategies for interpreting the experimental items. If they had developed such strategies, the outcome of the experiment would be less likely to support the SLCM predictions.

Finally, the comprehension questions from experiment 1 were modified to fit the new materials. Each subject saw the experimental materials in one list together with all the fillers and associated questions. The order of the sentences and fillers was random; each subject saw a different ordering.

Procedure

The procedure used was identical to that for experiment 1.

Predictions and Analyses

If the MLCH is correct, lexical category decisions are made prior to syntactic structure building. The lexical category disambiguation module does not know that the verb reading is not syntactically licit in the 'unambiguous' conditions. Our results from experiment 1 suggest that its decisions will be based exclusively on frequency information. However, the outcome may be affected by the granularity of lexical category over which co-occurrence statistics are collated. To understand this, we must look at the behaviour of the tag bigram SLCM.

If the SLCM assigns tags that include number information, then we would expect the SLCM to always assign a noun tag to the ambiguous word in the unambiguous conditions. This is because a singular noun is rarely followed by a plural verb; the probability of the verb reading should therefore be low. However, we gave reasons (in sections 4.2.4 and 7.5) why we not expect the SLCM to assign highly detailed lexical category tags. Number is often used (and determined) according to dependencies between non-adjacent words; such information is best excluded from SLCM assigned lexical categories.

In this case, we would expect the initial decisions of the SLCM to mirror the bias of

the ambiguous word, as in experiment 1. In both noun bias conditions, the initial decision favours a noun reading. This decision is compatible with both the syntactic constraints and the disambiguation. We would therefore not predict any reading time increase during the processing of these sentences.

In both the verb bias conditions, a verb reading will be initially preferred. In the AV condition (8.4a), this decision is incompatible with the disambiguating material. We would therefore predict a reading time increase at d_1. In the UV condition, a plural verb reading cannot be integrated into the current syntactic structure[63]; we would therefore expect the parser to reject the SLCM initial decision and force reanalysis as soon as the ambiguous word is read. This should be reflected by a reading time increase on the ambiguous word (c_2); slow reading times may also be observed on the following material, resulting from spill over.

In common with many constraint-based models, MacDonald, Pearlmutter and Seidenberg's (1994) proposal makes use of extremely detailed lexical category representations. They would therefore expect co-occurrence information, including noun and verb number, and syntactic constraints to affect initial decisions. In this case the parser should make an early decision favouring the noun reading on all but the AV condition; in these conditions, the model suggests no reading time increase in the disambiguating region. In the AV condition, the early decision would favour the incorrect verb reading, and we would therefore expect a reading time increase in the disambiguating region.

The 'reverse ambiguity' effect (MacDonald, 1993) can also be applied to this experiment. As the AV condition is the only one in which a verb phrase is actually constructed, this is the only condition in which MacDonald might predict a reading time increase on the ambiguous word. This contrasts with the SLCM prediction of a reading time increase only in the UV condition. However, see section 8.3.3 for an alternative 'prediction' that might be licensed by a constraint-based model.

In section 8.2.3 we demonstrated that the results of experiment 1 could not be explained by the non-statistical models outlined in chapter 2. The models that are predictive all make similar predictions for this experiment. In the two ambiguous

[63] See section 8.3.4 for further discussion.

conditions, the initial preference of the HSPM should always be the same, regardless of the bias of the ambiguous word. We would therefore expect no reading time difference during disambiguation between these two conditions. In the unambiguous conditions, the noun reading should be initially preferred (though see section 8.3.4); again, reading times for the disambiguating region in these two conditions should be similar.

Finally, the delay strategy predicts reduced reading times on the ambiguous word in the ambiguous conditions, compared to the unambiguous conditions, as processing is temporarily suspended. At the start of the disambiguating region, the delay model predicts increased reading times on both ambiguous conditions, again compared to the unambiguous controls, as the processor catches up.

8.3.2 Results

All sentence were included in reading time analysis. Word length adjustment was performed on the reading times, as for experiment 1. The average reading times for each word in the critical region, and for the first two words in the disambiguating region, are depicted in figure 8.3.

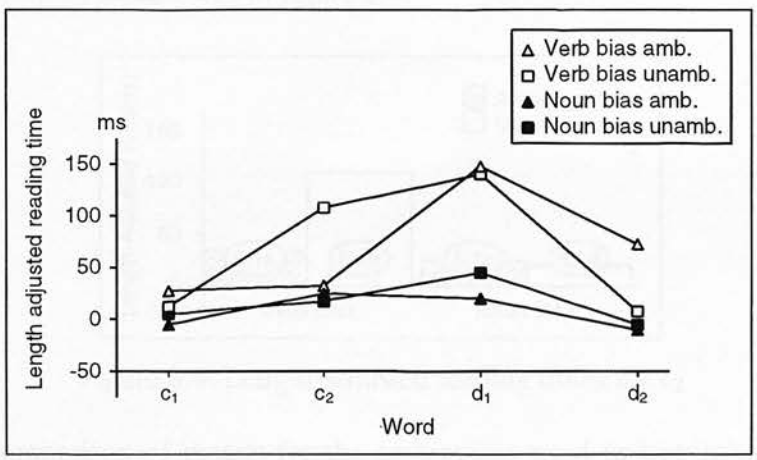

Figure 8.3: Length-adjusted reading times for experiment 2

On d_1, we found a main effect of bias ($F_1 = 20.1$, $p < .001$; $F_2 = 18.68$, $p < .001$). This effect results from the comparatively large reading times for the verb bias materials. There was no effect of ambiguity ($F_1 = 0.26$, $p > .6$; $F_2 = 0.16$, $p > .6$) and no interaction between bias and ambiguity.

From this result, it would appear that an early decision is made based on the bias of the word and ignoring the syntactic illegality of this reading in the UV condition. However, the results on word d_2 suggest a slightly different story. Here we find a significant effect of both bias ($F_1 = 8.91$, $p < .01$; $F_2 = 8.36$, $p < .01$) and ambiguity ($F_1 = 4.67$, $p < .05$; $F_2 = 4.73$, $p < .05$), and an interaction between bias and ambiguity ($F_1 = 8.79$, $p < .01$; $F_2 = 5.36$, $p = .03$). These effects can be attributed to increased reading time in the AV condition compared to the other three conditions.

If we take these two results together, they suggest that it is more difficult to recover a correct analysis in the AV condition than in the UV condition. By d_2, there is little if any residual processing difficulty in the UV condition (compared to the noun bias conditions). In contrast, processing is still significantly slower in the AV condition. This effect might occur if reanalysis was triggered earlier in the unambiguous condition; the earlier reanalysis is triggered, the sooner we would expect completion. According the SLCM predictions, reanalysis should be triggered on the ambiguous word (c_2) in the UV condition but only on d_1 in the AV condition. The results on d_2 are consistent with this prediction; however, we must examine the reading times for the ambiguous word itself to determine if there is direct evidence for early reanalysis. These are shown by themselves in figure 8.4.

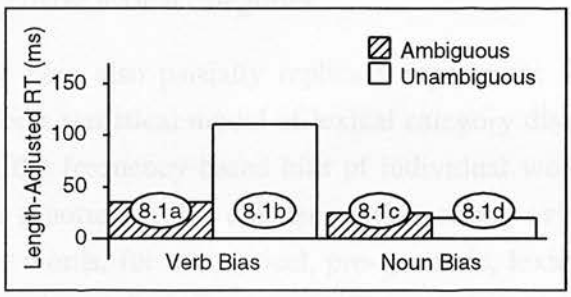

Figure 8.4: Length-adjusted reading times for c_2

A planned comparison of means for the ambiguous word (c_2) reveals a significant difference in reading time between the AV and UV conditions ($F_1 = 5.24$, $p < .03$; $F_2 = 7.16$, $p < .015$) but not between the AN and UN conditions ($F_1 = 0.12$, $p > .7$; $F_2 = 0.10$, $p > .75$). Subjects took longer to read unambiguous verb bias items. If we attribute this processing delay to reanalysis, then it suggests that an initial decision was made in favour of the verb reading in the verb bias unambiguous condition. The

syntactic incongruency of such a reading was ignored. However, syntax does play a role, and reanalysis occurs – or at least begins – while reading this word. In other words, an initial decision is made irrespective of syntactic number disagreement, but this disagreement forces rapid reanalysis.

8.3.3 Discussion

Supporting the MLCH

The results of this experiment are exactly as predicted by our proposed SLCM and fully support the MLCH. They suggest that initial lexical category decisions are made without reference to syntactic constraints – in other words, lexical category decisions are made pre-syntactically. In particular, we predicted processing delay on the ambiguous word in just the UV condition. This is exactly what we observed. This difficulty spills over onto the next word, but is resolved by d_2.

Interestingly, these results also support an SLCM variant that does not assign highly detailed lexical category tags. If number agreement information was included in these tags, then we would not expect the initial decisions of the SLCM to favour the verb reading in the UV condition, as the sequence singular noun followed by plural verb is rare. This experiment therefore provides some evidence for the early assignment of fairly coarse lexical categories.

The results reported here also partially replicate experiment 1. In particular, we again find support for a statistical model of lexical category disambiguation – initial decisions do reflect the frequency-based bias of individual words. Taken together, the two experiments reported in this chapter provide strong evidence for the SLCH and MLCH; in other words, for a statistical, pre-syntactic, lexical category decision module. The fact that the predictions of the SLCM exactly matched the experimental results suggests that the particular architecture we have proposed is on the right track. However, there are other possible *post hoc* accounts for the behaviour observed in experiment 2; we consider these below. Before that, we turn our attention to constraint-based models and MacDonald's 'reverse ambiguity' effect.

The 'Reverse Ambiguity' Effect and Constraint-Based Models

In the discussion following experiment 1 (section 8.2.3), we observed that MacDonald's (1993) account entails reading time differences on the ambiguous word

between our experimental conditions, but we observed none. While such a null effect suggests that MacDonald's syntactic complexity account is not valid, it does not prove the point. For experiment 2, MacDonald's hypothesis entails a reading time increase on the ambiguous word just in the AV condition (the only one in which a verb phrase is actually constructed) whereas our SLCM model predicts a reading time increase only in the UV condition. The latter prediction is supported by the observed behaviour. The fact that our reanalysis account makes correct predictions for novel experiments, whereas MacDonald's hypothesis has found no support, suggests that our analysis of her data (see section 6.2) should be preferred to her own (MacDonald, 1993). As our account arises from the SLCM model, this strengthens our argument that MacDonald's data supports our model.

Interactive models presuppose the immediate availability of all levels of information, including syntactic. We would therefore expect no processing difficulty on or after the ambiguous word in either of the unambiguous conditions. However, we did discover reading time increases on and after the ambiguous word in the verb bias unambiguous condition. Our results are therefore at odds with the 'natural' predictions of any interactive approach.

Unfortunately, while our results appear inconsistent with an interactive view, there is a possible account for them within a constraint-based approach. It could be argued that conflicting constraints (word bias and number agreement) apply in exactly those cases in which a reading time increase is observed on the ambiguous word. When constraints conflict (provided they are roughly equally weighted), we might expect that the sentence processor would take longer to come to an initial decision. A similar account may be offered for MacDonald's (1993) data. The problem with such an account is that it can only ever be proffered after the fact – unless all relevant constraints and the associated weights are predetermined. In other words, constraint-based accounts can explain (almost) anything, but getting stable predictions from them is far harder. We argued this point in chapter 3. Using our model, in contrast, we were able to predict exactly the results we obtained.

Non-Statistical Models and The Delay Strategy

The non-statistical models we considered in chapter 2 are either unpredictive for this ambiguity, or predict effects that do not tally with our results. In the discussion

following experiment 1 (section 8.2.3) we argued that such models are not sufficient to account for lexical category ambiguities; the results of this experiment confirm this analysis.

The results we obtained are also not compatible with the expectations of the delay strategy, which predicts increased reading times at the ambiguous word for both unambiguous conditions. We discovered increased reading time on the ambiguous word in just the verb bias unambiguous condition. The reading times in the disambiguating region also fail to match delay strategy predictions; these results can therefore be seen as a further nail in the delay strategy's coffin.

An Alternative Explanation

This and the previous experiment have allowed us to rule out an account in which a non-statistical parser is responsible for lexical category disambiguation. This experiment has provided evidence that such ambiguity is also not the province of a statistical parser. However, there is a possible alternative account.

Number agreement was used in this experiment to create unambiguous versions of the materials. However, it could be that initial structural decisions ignore number agreement. A modular parser creates a syntactic structure, and then a later module checks that such constraints as number agreement hold; if they do not, a new analysis is requested from the parser.

Such a model is similar to that proposed by Mitchell (1987) to explain another result in which processing delays occur when reading an unambiguous sentence (see section 7.4). A modular model in which a statistical parser makes initial attachment decisions without checking number agreement and then a later module makes this check is compatible with the results reported in this section. Such a model still falls within the Modular Statistical Hypothesis; our results therefore provide strong evidence for this hypothesis, whichever interpretation is placed on them.

However, it is not clear whether a 'late number agreement' account offers an explanation of MacDonald's (1993) results, whereas the SLCM offers a unified explanation for both MacDonald's data and the present experiment. Further, the fact that the SLCM model exactly predicted the results we obtained cannot be discounted – while alternative *post hoc* analyses may also account for the data, predictions of the

SLCM model have been upheld.

8.3.4 Collective Nouns

A revue of our experimental materials reveals that they suffer from a defect. In a few cases, c_1 can be used as a collective noun. This makes no difference to the results for experiment 1. In experiment 2, unambiguous materials were constructed by manipulating the number of the ambiguous word; as c_1 is always singular, it is not compatible with a reading in which c_2 is a plural verb. However, a collective noun is morphologically singular but syntactically plural. Therefore, when c_1 may be a collective noun, the 'unambiguous condition' materials are not truly syntactically unambiguous.

Consider, for example, 8.4a, which is one of our 'unambiguous' experimental materials:

(8.4) a. The council are proud that the village boast is the talk of the county.
b. The council are proud that the village boast the best wine in France.
c. The council are proud that the villagers boast the best wine in France.

The word "village" is singular, whereas "boast", as a verb, must be plural; they disagree in number, and therefore the verb reading of "boast" should be ruled out. However, as shown in example 8.4b, it is possible to construct a grammatical disambiguation for this sentence which forces the verb reading. While the collective usage of "village" is unusual – 8.4c would be a far more normal way to express this proposition – it suggests that some of our 'unambiguous' sentences are, in fact, ambiguous.

In the UN condition, this will make little difference. In experiment 1 we observed that the HSPM initially favours a noun reading in the noun biased condition. It is therefore highly unlikely that the existence of a less preferred verb reading would affect the behaviour of the sentence processor. We can therefore ignore the collective reading of some of the unambiguous noun biased materials.

In the UV condition, things are somewhat different. We have interpreted experiment 2 as demonstrating that the HSPM makes initial category decisions even when syntactic constraints render such a reading illegal (in chapter 4 we termed these NSC

ambiguities). This evidence supports a modular model in which (some) syntactic information is not available to the module that makes lexical category decisions. However, if some of our sentences were not syntactically unambiguous, then it could be argued that the reanalysis effect we observed on the ambiguous word is due to semantic rather than syntactic incongruency.

In order to determine whether the presence of collective nouns is affecting the outcome of the experiment, we considered which of the words used at c_1 in the verb bias condition have a possible collective reading. We discovered two – "church" and "village". We removed all experimental items containing these words, and all noun biased materials that were paired with these items; in total, 16 of the 96 sentences were rejected. Having filtered all sentences that could adversely affect the result of the experiment, we reanalysed the data associated with the remaining items.

Our results followed exactly the same pattern as that reported in section 8.3.2. In particular, we observed a reading time difference on the ambiguous word between the UV and AV conditions ($F_1 = 7.31, p < .015$; $F_2 = 7.50, p < .015$) but not between the UN and AN conditions ($F_1 = 0.68; p > .4$; $F_2 = 0.87, p > .3$). This result indicates that the presence of collective nouns had no significant effect on the outcome of experiment 2; this experiment still provides strong evidence for the independence of lexical category disambiguation.

8.3.5 Summary

In experiment 2 we tested the hypothesis that lexical category disambiguation occurs as a separate modular process, prior to syntactic parsing. The results of this experiment strongly supported this hypothesis (the MLCH). We also found further support for the SLCH, formally tested in experiment 1.

These results are compatible with the detailed SLCM model proposed in chapter 4. They do not appear compatible with MacDonald's (1993) complexity account of reading time increases on the ambiguous word in noun–verb ambiguities, but are entailed by our reanalysis account. Our own analysis of MacDonald's data (section 6.2) and therefore the SLCM on which it is based therefore gain further support from this experiment.

Unfortunately, there is an alternative explanation of the results we have observed – number agreement constraints may only apply after syntactic structure building. However, it is not clear that such an account could also explain MacDonald's data, and the fact that SLCM predictions made prior to experimentation exactly matched observed results is compelling evidence for our model.

The MSH also gains strong support from these results. We have already shown (in experiment 1) that lexical category decisions are controlled by a frequency-based process. This experiment strongly suggests that this process occurs prior to syntactic parsing, and therefore that the HSPM is modular.

8.4 Conclusions

Our experiments strongly support the existence of a modular statistical lexical category disambiguation module. The existence of such a module is one of the core proposals of this thesis. Analysis of existing evidence in chapters 6 and 7 has provided support for the detailed proposals concerning the architecture of this module presented in chapter 4. More generally, these experiments also provide support for the most basic proposal of this thesis, the MSH, as any model that incorporates a statistical lexical category disambiguation module automatically falls within this hypothesis.

However, it would be wrong to pretend that the debate is won. While the results presented here are exactly as predicted by the SLCM, they are also open to alternative interpretations. Some of these have been covered in this chapter. Our evidence also fails to prove that interactive or constraint-based accounts do not hold the correct explanation for this data; as argued in chapter 3, it is less than clear whether this proposition could ever be proven.

In conclusion, these results strongly support the SLCM. However, they also raise many questions that can only be answered by future experimentation. In the final chapter we suggest what form such experimentation might take.

9: Conclusions

If you've got something left to say
You'd better say it now
(The Cure, "Bare", from the album "Wild Mood Swings")

9.1 Achievements

In chapter 1, we set out the aims of this thesis as follows:

- To argue that the correct architecture of the HSPM is both modular and statistical – the Modular Statistical Hypothesis, introduced in chapter 3.
- To propose and provide empirical support for a position in which human lexical category disambiguation occurs within a modular process, distinct from syntactic parsing and guided by a statistical decision process.

We addressed the first of these directly in chapters 2 and 3. In chapter 2 we reviewed a number of proposals concerning the architecture of the HSPM and, in particular, examined statistical positions and the evidence that supports them. This evidence is inconclusive; however, there are some results (summarised in Cuetos *et al.*, 1996 and Corley, 1995) that are difficult to explain without making appeal to statistical mechanisms. Our own experimental results, presented in chapter 8, strongly support a position in which at least part of the HSPM makes decisions based on statistical regularities in the individual's prior linguistic experience.

In chapter 3 we concerned ourselves with modularity, arguing that a *modular* statistical position is both empirically plausible and rationally preferable. Such a position places strong constraints on the possible behaviours of the sentence processor and is therefore open to empirical investigation. However, we did not provide direct evidence for or against the Modular Statistical Hypothesis – instead, we chose to consider one possible position encompassed by the MSH, and demonstrate the validity of the wider hypothesis by providing empirical support for

the narrower one.

We introduced this hypothesis in chapter 4, where we both investigated the notion of statistical lexical category disambiguation and proposed a simple and fully specified mathematical model of how human lexical category disambiguation occurs; we called this model the SLCM. Chapters 6 and 7 reinterpret existing experimental studies in the light of this model. We conclude that the SLCM is compatible with previously presented results; in fact, in some cases it offers an apparently more plausible explanation than that proposed by the original researchers. The evaluation in chapters 6 and 7 also allowed us to rule out a number possible SLCM variants and commit to a model in which tag bigram probabilities play a central role in lexical category decisions.

Finally, in chapter 8, we empirically tested the wider claim that lexical category decisions are made by a modular and statistical process; we called the two hypotheses embodied by this claim the Modular Lexical Category Hypothesis (MLCH) and the Statistical Lexical Category Hypothesis (SLCH). Our experimental findings coincided exactly with the predictions derived from the SLCM and provided strong support for both hypotheses. Evidence for the MLCH clearly constitutes support for a modular theory of human sentence processing; likewise, a statistical theory of human sentence processing is supported by evidence for the SLCH. These results therefore strongly favour our most basic claim, the Modular Statistical Hypothesis.

In conclusion, the evidence presented in this thesis strongly supports a position in which lexical category decisions are made on the basis of frequency information derived from an individual's previous linguistic experience, and in which relevant syntactic information is ignored when making initial lexical category decisions. By implication, this evidence also strongly supports a position in which the HSPM is both modular and at least partially statistical.

9.2 Limitations

While the evidence we have considered in this thesis strongly indicates that the HSPM is modular and statistical, and that there is a distinct statistical lexical category disambiguation module, the SLCM proposed in chapter 4 has certain

limitations which suggest it is a simplification of the human lexical category decision process. In particular, we have not considered how unknown words are handled. Further, while we have shown that much existing evidence can be explained by a model in which very simple statistical measures, such as tag bigram probabilities, play a central role, it is by no means clear that more fine-grained statistics are never used. Finally, the model does not, in itself, define lexical category. In this section we consider each of these limitations and propose possible additions to the SLCM.

9.2.1 Unknown Words

Language learners, and even fluent speakers, frequently encounter unknown words. It is often possible to discover the meaning of such words from contextual clues. Even when many of the words in a sentence are unknown, it is sometimes possible to assign a structure to the sentence;[64] it must therefore also be possible to determine the grammatical category of an unknown word. However, the proposed SLCM does not contain any mechanism for handling unknown words.

If we were to extend the model in order to cope with unknown words, the simplest option would be to assume that the estimated probability of a word given a tag ($P(w_i \mid t_i)$) is equal for all tags for any unknown word. The SLCM decision then depends purely on the tag co-occurrence probability ($P(t_i \mid t_{i-1})$). That is, lexical category decisions in the face of unknown words depend purely on the lexical context in which the word occurs.

However, the orthography (or phonetic realisation) of the word itself may give clues as to its part of speech, and it seems plausible that the SLCM could make use of such information. For example, if a word ends in '-ly' it is highly likely that it is an adverb. When several words are joined by hyphens, the resultant unit is almost always used as an adjective. Capitalised words that are not sentence initial are frequently proper nouns. Clues of this sort might be used by an SLCM that is far more sophisticated than the simple model we have proposed. Determining human behaviour in the face of unknown words and neologisms would be an interesting project that may throw further light on the issues discussed in this thesis.

[64] Obvious examples include nonsense poetry, such as Lewis Carroll's 'Jabberwocky', and cyberpunk fiction.

9.2.2 Fine-grained Statistics?

We have argued throughout this thesis that comparatively coarse-grained frequency information about lexical category co-occurrence is sufficient to explain a wide range of data about human language understanding. However, it is not clear that more fine-grained statistical knowledge is never used by the SLCM. For example, word co-occurrence (rather than tag co-occurrence) statistics may be available for very common word pairs only; in fact, the SLCM need only maintain statistics for such pairs when they tend to have an anomalous tag sequence. However, most such occurrences are more plausibly construed as single lexical items – such as many of the experimental items in MacDonald's (1993) study (see section 6.2.2) and the second occurrence of "has been" in example 9.1:

(9.1) He has been a has been all his adult life.

Again, the granularity of the information used by the SLCM would be an interesting area for further study. However, as we have argued throughout this thesis, proposing that the HSPM makes use of fine-grained information is often unnecessary and leads to a more complex and less predictive model. By Occam's razor, we should prefer models of human sentence processing that only make use of coarse-grained information provided they offer a plausible and simple account of the available empirical evidence.

9.2.3 Lexical Category Tags

Finer-grained statistical information may also be incorporated into the SLCM without any change in processing architecture, by simply proposing a more detailed set of lexical category tags. In chapter 7 we proposed that verb tags may include limited transitivity information. It is clear that there are a number of other distinctions that could be made without greatly reducing the initial accuracy of the SLCM. The examples in 9.2 to 9.4 exemplify some of these.

(9.2) a. The *running* man passed the bus stop.
b. The man was *running* past the bus stop.

(9.3) a. The *angry* man shouted at us.
b. The man who shouted at us was *angry*.

(9.4) a. The dragon flies *of* South America are huge.

b. The dragon flies *over* the village at night.

In 9.2a, the present participle "running" is used attributively, whereas in 9.2b it is predicative. This is a distinction we made in chapter 7 when marking the SUSANNE Corpus for transitivity; however, we never made use of this distinction. A similar distinction can be made for adjectives, as in example 9.3.

The preposition "of" almost never modifies a verb, whereas most other prepositions do. This is exemplified in 9.5; it is our intuition that in 9.5a the noun reading of "flies" is preferred as soon as the following "of" is read; in contrast, 9.5b supports a verb reading of "flies". Such an effect could be captured by the SLCM if we assumed that "of" has a different lexical category from other prepositions. Again, we might suggest that the SLCM assigns more fine-grained categories.[65]

In common with many psycholinguistic models, the predictions of the SLCM therefore depend on the linguistic theory on which it is based. Determining the set of lexical categories assigned by the SLCM is of paramount importance if we wish to establish firm predictions on which the model stands or falls. However, as we suggest in the next section, it isn't that easy.

9.3 Future Directions

There are two clear directions in which the research presented in this thesis should continue. Firstly, more evidence is required showing whether syntactic processes affect initial lexical category decisions. While the evidence presented in chapter 8 strongly supports a model in which it does not, we have only examined one particular lexical category ambiguity. Empirical evidence showing that syntactic processes do not affect lexical category decisions (but lexical co-occurrence does) across a range of syntactic constructs would be far more compelling.

The second area to be addressed in future research are the limitations presented in section 9.2. The most important of these is the determination of the granularity of lexical categories assigned by the SLCM. Experimental study should reveal what granularity is necessary to explain the data. However, it is important that we separate

[65] The distinctions suggested here, along with a number of others, have all been added to the Brown Corpus by the author in recent work for Sharp Laboratories of Europe. A parts-of-speech tagger trained on this data proved highly effective.

work on determining granularity from work on further establishing the model itself; otherwise, we too fall prey to the criticism we have levelled at interactive models – that adjustments to the model can be made to explain almost any data, and so the model itself is unfalsifiable.

References

Adams, B.C., Clifton, C. and Mitchell, D.C. (submitted). *Lexical Guidance in Sentence Processing.*

Altmann, G.T.M. (1988). Ambiguity, Parsing Strategies and Computational Models. *Language and Cognitive Processes*, **3**, 73–97.

Altmann, G.T.M. and Steedman, M. (1988). Interaction with Context during Human Sentence Processing. *Cognition*, **30**, 191–238.

Bever, T.G. (1970). The Cognitive Basis for Linguistic Structures. In J.R. Hayes (ed.) *Cognition and the Development of Language*. Chapter 9, pp. 279–362. Wiley, New York.

Branigan, H.P. (1995). *Language Processing and the Mental Representation of Syntactic Structure*. PhD thesis, University of Edinburgh.

Branigan, H.P., Sturt, P. and Matsumoto Sturt, Y. (1996). *Left Branching Attachment and Thematic Domains*. Poster presented at the AMLaP-96 Conference, Trento, Italy.

Brill, E. (1992). A Simple Rule-Based Parts of Speech Tagger. In *Proceedings of the Third Conference on Applied Natural Language Processing*. Trento, Italy. ACL.

Britt, M.A. (1994). The Interaction of Referential Ambiguity and Argument Structure in the Parsing of Prepositional Phrases. *Journal of Memory and Language*, **33**, 251–283.

Britt, M.A., Perfetti, C.A., Garrod, S. and Rayner, K. (1992). Parsing in Discourse: Content Effects and their Limits. *Journal of Memory and Language*, **31**, 293–314.

Brysbaert, M. and Mitchell, D.C. (1996a). Modifier Attachment in Sentence Parsing: Evidence from Dutch. *Quarterly Journal of Experimental Psychology*, **49A**.

Brysbaert, M. and Mitchell, D.C. (1996b). *Modifier Attachment in Dutch: Deciding between Garden-Path, Construal and Statistical Tuning Accounts of Parsing.* Paper presented at the Workshop on Computational Models of Human Syntactic Processing, Wassenaar, Holland.

Carreiras, M. and Clifton, C. (1993). Relative Clause Interpretation Preferences in

Spanish and English. *Language and Speech*, **36**, 353–372.

Charniak, E. (1993). *Statistical Language Learning*. MIT Press, Cambridge, MA.

Charniak, E., Hendrickson, C., Jacobson, N. and Perkowitz, M. (1993). Equations for Part-of-Speech Tagging. In *Proceedings of the Eleventh National Conference on Artificial Intelligence*. pp. 784–789, Washington, DC. AAAI Press/MIT Press.

Chomsky, N. (1957). *Syntactic Structures*. Mouton, The Hague, NL.

Chomsky, N. (1981). *Lectures on Government and Binding*. Foris Publications, Dordrecht, NL.

Christ, O. (1993). *The Xkwic User Manual*. IMS, Universität Stuttgart.

Christ, O. (1994). A Modular and Flexible Architecture for an Integrated Corpus Query System. In *Proceedings of COMPLEX '94: 3rd Conference on Computational Lexicography and Text Research*. pp. 23–32, Budapest, Hungary.

Church, K.W. (1988). A Stochastic Parts Program and Noun Phrase Parser for Unrestricted Text. In *Proceedings of the Second Conference on Applied Natural Language Processing*. pp. 136–143, Austin, Texas. ACL.

Churchland, P.M. (1988). Perceptual Plasticity and Theoretical Neutrality: A Reply to Jerry Fodor. *Philosophy of Science*, **55**, 167–187.

Clifton, C. (1994). *Restrictions on Late Closure: Appearance and Reality*. Paper presented at the 6th Australian Language and Speech Conference, Sydney.

Clifton, C. and Ferreira, F. (1989). Ambiguity in Context. *Language and Cognitive Processes*, **4**, SI77–104.

Clifton, C., Frazier, L., Rapoport, T. and Radó, J. (submitted). *Adjunct Predication: Attachment or Construal.*

Clifton, C., Speer, S. and Abney, S.P. (1991). Parsing Arguments: Phrase Structure and Argument Structure as Determinants of Initial Parsing Decisions. *Journal of Memory and Language*, **30**, 251–271.

Coltheart, M. (1981). The MRC Psycholinguistic Database. *Quarterly Journal of Experimental Psychology*, **33A**, 497–505.

Corley, M.M.B. (1995).*The Rôle of Statistics in Human Sentence Processing*. PhD thesis, University of Exeter, Exeter, UK.

Corley, M.M.B. and Caldwell, C. (1996). *Towards a Simple Statistical Model of the HSPM*. Poster presented at the AMLaP-96 Conference, Turin, Italy.

Corley, M.M.B. and Corley, S. (1995). *Cross-linguistic and Intra-Linguistic Evidence for the use of Statistics in Human Sentence Processing*. Unpublished manuscript, University of Exeter.

Corley, M.M.B., Mitchell, D.C., Brysbaert, M., Cuetos, F. and Corley, S. (1995). Exploring the Rôle of Statistics in Human Natural Language Processing. In *Proceedings of the 4th International Conference on the Cognitive Science of Natural Language Processing*. Dublin, Ireland.

Corley, S. (1996). *CORSET: A Corpus Search Toolkit. User Manual.* Department of Artificial Intelligence, University of Edinburgh.

Crain, S. and Steedman, M.J. (1985). On not being led up the Garden Path: The Use of Context by the Psychological Parser. In D.R. Dowty, L. Karttunen and A.M. Zwicky (eds.) *Natural Language Parsing: Psychological, Computational, and Theoretical Perspectives*. Chapter 10, pp. 320–358. Cambridge University Press, Cambridge, UK.

Crocker, M.W. (1991). *A Logical Model of Competence and Performance in the Human Sentence Processor*. PhD thesis, University of Edinburgh, Edinburgh, UK.

Crocker, M.W. (1996). *Computational Psycholinguistics: An Interdisciplinary Approach to the Study of Language*. Kluwer Academic Publishers.

Cuetos, F. and Mitchell, D.C. (1988). Cross-Linguistic Differences in Parsing: Restrictions on the use of the Late Closure Strategy in Spanish. *Cognition*, **30**, 73–105.

Cuetos, F., Mitchell, D.C. and Corley, M.M.B. (1996). Parsing in Different Languages. In M. Carreiras, J. Garcia-Albea and N. Sabastian-Galles (eds.) *Language Processing in Spanish*. Chapter 6, pp. 145–187. Erlbaum, Hillsdale, NJ.

Cutting, D., Kuppiec, J., Pedersen, J. and Sibun, P. (1992). A Practical Parts of Speech Tagger. In *Proceedings of the Third Conference on Applied Natural Language Processing*. Trento, Italy. ACL. Also available as Xerox PARC technical report SSL-92-01.

DeRose, S.J. (1988). Grammatical Category Disambiguation by Statistical Optimization. *Computational Linguistics*, **14**, 31–39.

Ferreira, F. and Clifton, C. (1986). The Independence of Syntactic Processing. *Journal of Memory and Language*, **25**, 348–368.

Ferreira, F. and Henderson, J.M. (1990). The Use of Verb Information in Syntactic

Parsing: Evidence from Eye Movements and Word-by-Word Self-Paced Reading. *Journal of Experimental Psychology: Learning, Memory and Cognition*, **16**, 555–568.

Ferreira, F. and Henderson, J.M. (1991). Recovery from Misanalysis of Garden-Path Sentences. *Journal of Memory and Language*, **30**, 725–745.

Fodor, J.A. (1983). *The Modularity of Mind.* MIT Press, Cambridge, MA.

Fodor, J.A. (1987). Modules, Frames, Fridgeons, Sleeping Dogs, and the Music of the Spheres. In J.L. Garfield (ed.) *Modularity in Knowledge Representation and Natural-Language Understanding*. Chapter 1, pp. 25–36. MIT Press, Cambridge, MA.

Fodor, J.A. and Bever, T.G. (1965). The Psychological Reality of Linguistic Segments. *Journal of Verbal Learning and Verbal Behaviour*, **4**, 414–420.

Fodor, J.A., Garret, M.F. and Bever, T.G. (1968). Some Syntactic Determinants of Sentential Complexity, II: Verb Structure. *Perception and Psychophysics*, **3**, 453–461.

Fodor, J.D. and Ferreira, F., eds. (in press). *Reanalysis in Sentence Processing*. Kluwer Academic Publishers.

Ford, M., Bresnan, J. and Kaplan, R.M. (1982). A Competence-Based Theory of Syntactic Closure. In J. Bresnan (ed.) *The Mental Representation of Grammatical Relations*. Chapter 11, pp. 727–796. MIT Press, Cambridge, MA.

Forster, K.I. (1976). Accessing the Mental Lexicon. In R.J. Wales and E. Walker (eds.) *New Approaches to Language Mechanisms*. pp. 257–287. North Holland, Amsterdam, NL.

Forster, K.I. (1979). Levels of Processing and the Structure of the Language Processor. In W.E. Cooper and E.C.T. Walker (eds.) *Sentence Processing: Psycholinguistic Studies presented to Merrill Garrett.* Erlbaum, Hillsdale, NJ.

Francis, W.N. and Kucera, H. (1982). *Frequency Analysis of English Usage: Lexicon and Grammar*. Houghton Mifflin, Boston, MA.

Frazier, L. (1979). *On Comprehending Sentences: Syntactic Parsing Strategies.* PhD thesis, University of Massachusetts.

Frazier, L. (1985). Modularity and the Representational Hypothesis. In *Proceedings of the Northeastern Linguistics Society, November 1984*. Providence, RI. Brown University.

Frazier, L. (1987a). Sentence Processing: A Tutorial Review. In M. Coltheart (ed.) *Attention and Performance XII: The Psychology of Reading*. pp. 554–586. Erlbaum, Hillsdale, NJ.

Frazier, L. (1987b). Theories of Sentence Processing. In J.L. Garfield (ed.) *Modularity in Knowledge Representation and Natural-Language Understanding*. Chapter 15, pp. 291–307. MIT Press, Cambridge, MA.

Frazier, L. (1990). Exploring the Architecture of the Language-Processing System. In G.T.M. Altmann (ed.) *Cognitive Models of Speech Processing: Psychological and Computational Perspectives*. Chapter 19, pp. 409–433. MIT Press, Cambridge, MA.

Frazier, L. and Clifton, C. (1996). *Construal*. MIT Press, Cambridge, MA.

Frazier, L. and Fodor, J. (1978). The Sausage Machine: A New Two-Stage Parsing Model. *Cognition*, **6**, 291–325.

Frazier, L. and Rayner, K. (1982). Making and Correcting Errors during Sentence Comprehension: Eye Movements in the Analysis of Structurally Ambiguous Sentences. *Cognitive Psychology*, **14**, 178–210.

Frazier, L. and Rayner, K. (1987). Resolution of Syntactic Category Ambiguities: Eye Movements in Parsing Lexically Ambiguous Sentences. *Journal of Memory and Language*, **26**, 505–526.

Garside, R. (1987). The CLAWS Word-Tagging System. In R. Garside, G. Leech and G. Sampson (eds.) *The Computational Analysis of English: a corpus-based approach*. Chapter 3, pp. 30–41. Longman.

Gernsbacher, M.A., Hargreaves, D.J. and Beeman, M. (1989). Building and Accessing Clausal Representations: the Advantage of First Mention versus the Advantage of Clause Recency. *Journal of Memory and Language*, **28**, 735–755.

Greene, B.B. and Rubin, G.M. (1971). *Automated Grammatical Tagging of English*. Department of Linguistics, Brown University, Providence, Rhode Island.

Gibson, E. (1991). *A Computational Theory of Human Linguistic Processing: Memory Limitations and Breakdown*. PhD thesis, Carnegie Mellon University, Pittsburgh.

Gibson, E., Pearlmutter, N., Canseco-Gonzalez, E. and Hickok, G. (1996). Recency Preference in the Human Sentence Processing Mechanism. *Cognition*, **59**, 23–59.

Gibson, E. and Schütze, C.T. (1996). *The Relationship between the Frequency and the Perceived Complexity of Conjunction Attachments: On-line Evidence*. Poster

presented at the 9th Annual CUNY Conference on Human Sentence Processing, New York.

Gibson, E., Schütze, C.T. and Salomon, A. (1996). The Relationship Between the Frequency and the Processing Complexity of Linguistic Structure. *Journal of Psycholinguistic Research*, **25**, 59–92.

Gilboy, E., Sopena, J.M., Clifton, C. and Frazier, L. (1995). Argument Structure and Association Preferences in Spanish and English complex NPs. *Cognition*, **54**, 131–167.

Hertz, J., Krogh, A. and Palmer, R.G. (1991). *Introduction to the Theory of Neural Computation*. Addison-Wesley, Redwood City, CA.

Holmes, V.M., Kennedy, A. and Murray, W.S. (1987). Syntactic Structures and the Garden Path. *Quarterly Journal of Experimental Psychology*, **39A**, 277–293.

Johansson, S., Atwell, E., Garside, R. and Leech, G. (1986). *The Tagged LOB Corpus: Users' Manual.*

Juliano, C. and Tanenhaus, M.K. (1993). Contingent Frequency Effects in Syntactic Ambiguity Resolution. In *Proceedings of the Fifteenth Annual Conference of the Cognitive Science Society*. pp. 593–598. Lawrence Erlbaum Associates.

Juliano, C. and Tanenhaus, M.K. (1994). A Constraint-Based Lexicalist Account of the Subject/Object Attachment Preference. *Journal of Psycholinguistic Research*, **23**, 459–471.

Just, M.A. and Carpenter, P.A. (1980). A Theory of Reading: From Eye Fixations to Comprehension. *Psychological Review*, **87**, 329–354.

Just, M.A., Carpenter, P.A. and Woolley, J.D. (1982). Paradigms and Processes in Reading Comprehension. *Journal of Experimental Psychology: General*, **111**, 228–238.

Kamide, Y. and Mitchell, D.C. (1996). *Relative Clause Attachment: Evidence from Japanese*. Poster presented at the 9th Annual CUNY Conference on Human Sentence Processing, New York.

Katz, S.M. (1987). Estimation of Probabilities from Sparse Data from the Language Model of a Speech Recogniser. *IEEE Transactions on Acoustics, Speech and Signal Processing*, **35**, 400–401.

Kennedy, A., Murray, W.S., Jennings, F. and Reid, C. (1989). Parsing Complements: Comments on the Generality of the Principle of Minimal Attachment. *Language and Cognitive Processes*, **4**, SI57–76.

Kimball, J. (1973). Seven Principles of Surface Structure Parsing in Natural Language. *Cognition*, **2**, 15–47.

Klein, S. and Simmons, R.F. (1963). A Computational Approach to Grammatical Coding of English Words. *Journal of the Association for Computing Machinery*, **10**, 334–347.

Kucera, H. and Francis, W.N. (1967). *Computational Analysis of Present-Day American English*. Brown University Press, Providence, RI.

Levelt, W.J.M. (1978). A Survey of Studies in Sentence Perception: 1970-1976. In W.J.M. Levelt and G.B. Flores d'Arcais (eds.) *Studies in the Perception of Language*. Chapter 1, pp. 1–74. Wiley, Chichester, UK.

MacDonald, M.C. (1993). The Interaction of Lexical and Syntactic Ambiguity. *Journal of Memory and Language*, **32**, 692–715.

MacDonald, M.C. (1994). Probabilistic Constraints and Syntactic Ambiguity Resolution. *Language and Cognitive Processes*, **9**, 157–201.

MacDonald, M.C., Just, M.A. and Carpenter, P.A. (1992). Working Memory Constraints on the Processing of Syntactic Ambiguity. *Cognitive Psychology*, **24**, 56–98.

MacDonald, M.C., Pearlmutter, N.J. and Seidenberg, M.S. (1994). Lexical Nature of Syntactic Ambiguity Resolution. *Psychological Review*, **101**, 676–703.

Marcus, M.P., Hindle, D. and Fleck, M. (1983). D-theory: Talking about Talking about Trees. In *Proceedings of the 21st Annual Meeting of the Association for Computational Linguistics*. pp. 129–136, Cambridge, MA.

Marcus, M.P., Kim, G., Marcinkiewicz, M.A., MacIntyre, R., Bies, A., Ferguson, M., Katz, K. and Schasberger, B. (1994). *The Penn Treebank: Annotating Predicate Argument Structure*. Presented at the ARPA Human Language Technology Workshop.

Marcus, M.P., Santorini, B. and Marcinkiewicz, M.A. (1993). Building a Large Annotated Corpus of English: the Penn Treebank. *Computational Linguistics*, **19**, 313–330.

Marslen-Wilson, W.D. (1975). Sentence Perception as an Interactive Parallel Process. *Science*, **189**, 226–228.

Marslen-Wilson, W. and Tyler, L.K. (1987). Against Modularity. In J.L. Garfield (ed.) *Modularity in Knowledge Representation and Natural-Language Understanding*. Chapter 2, pp. 37–62. MIT Press, Cambridge, MA.

McDermott, D. (1986). We've Been Framed: Or, Why AI is Innocent of the Frame Problem. In Z. Pylyshyn (ed.) *The Robot's Dilemna: The Frame Problem in Artificial Intelligence*. Ablex.

Merlo, P. (1992). *On Modularity and Compilation in a Government-Binding Parser*. PhD thesis, University of Maryland, College Park, MD.

Miller, G.A. and Isard, S. (1963). Some Perceptual Consequences of Linguistic Rules. *Journal of Verbal Learning and Verbal Behaviour*, **2**, 217–228.

Mitchell, D.C. (1987). Lexical Guidance in Human Parsing: Locus and Processing Characteristics. In M. Coltheart (ed.) *Attention and Performance XII: The Psychology of Reading*. Chapter 27, pp. 601–618. Erlbaum, Hillsdale, NJ.

Mitchell, D.C. (1994). Sentence Parsing. In M.A. Gernsbacher (ed.) *HandBook of Psycholinguistics*. Chapter 11, pp. 375–409. Academic Press, San Diego, CA.

Mitchell, D.C., Corley, M.M.B. and Garnham, A. (1992). Effects of Context in Human Sentence Parsing: Evidence Against a Discourse-Based Proposal Mechanism. *Journal of Experimental Psychology: Learning, Memory and Cognition*, **18**, 69–88.

Mitchell, D.C. and Cuetos, F. (1991). The Origins of Parsing Strategies. In C. Smith (ed.) *Current Issues in Natural Language Processing*. University of Austin, Austin, TX.

Mitchell, D.C., Cuetos, F., Corley, M.M.B. and Brysbaert, M. (1995). Exposure-Based Models of Human Parsing: Evidence for the Use of Coarse-Grained (Nonlexical) Statistical Records. *Journal of Psycholinguistic Research*, **24**, 469–488.

Mitchell, D.C. and Holmes, V.M. (1985). The Role of Specific Information about the Verb in Parsing Sentences with Local Structural Ambiguity. *Journal of Memory and Language*, **24**, 542–559.

Morton, J. (1969). Interaction of Information in Word Recognition. *Psychological Review*, **76**, 165–178.

Pritchett, B.L. (1992). *Grammatical Competence and Parsing Performance*. University of Chicago Press, Chicago, IL.

Raphael, B. (1971). The Frame Problem in Problem-Solving Systems. In N.V. Findler and B. Meltzer (eds.) *Artificial Intelligence and Heuristic Programming*. pp. 159–169. Edinburgh University Press, Edinburgh, UK.

Rayner, K., Carlson, M. and 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 Behaviour*, **22**, 358–374.

Rayner, K. and Duffy, S.A. (1986). Lexical Complexity and Fixation Times in Reading: Effects of Word Frequency, Verb Complexity, and Lexical Ambiguity. *Memory and Cognition*, **14**, 191–201.

Rayner, K. and Pollatsek, A. (1989). *The Psychology of Reading*. Prentice Hall.

Ritchie, G.D. (1983). Semantics in Parsing. In M. King (ed.) *Parsing Natural Language*. Chapter 10, pp. 199–217. Academic Press, London, UK.

Sampson, G. (1995). *English for the Computer: the SUSANNE Corpus and Analytic Scheme*. Clarendon Press, Oxford, UK.

Schank, R.C. (1972). Conceptual Dependency: A Theory of Natural Language Understanding. *Cognitive Psychology*, **3**, 552–631.

Schulze, B.M. and Heid, U. (1994). State-of-the-Art Survey of Corpus Query Tools. Deliverable of DECIDE (MLAP-Project 93-19) D-1b.

Seidenberg, M.S. and McClelland, J.L. (1989). A Distributed, Developmental Model of Word Recognition and Naming. *Psychological Review*, **96**, 523–568.

Seidenberg, M.S., Tanenhaus, M.K., Leiman, J.M. and Bienkowski, M. (1982). Automatic Access of the Meanings of Ambiguous Words in Context: Some Limitations on Knowledge-Based Processing. *Cognitive Psychology*, **14**, 489–537.

Sells, P. (1985). *Lectures on Contemporary Syntactic Theory*. CSLI, Menlo Park, CA.

Spivey-Knowlton, M. and Eberhard, K. (1996). The Future of Modularity. In *Proceedings of the Eighteenth Annual Conference of the Cognitive Science Society*. pp. 39––40, San Diego, CA. Lawrence Erlbaum Associates.

Spivey-Knowlton, M. and Sedivy, J.C. (1995). Resolving Attachment Ambiguities with Multiple Constraints. *Cognition*, **55**, 227–267.

St. John, M.F. and McClelland, J.L. (1990). Learning and Applying Contextual Constraints in Sentence Comprehension. *Artificial Intelligence*, **46**, 217–257.

Stowe, L.A. (1986). Parsing WH-constructions: evidence for on-line gap location. *Language and Cognitive Processes*, **1**, 227–245.

Sturt, P. and Crocker, M.W. (1996). Monotonic Syntactic Processing: A Cross-

Linguistic Study of Attachment and Reanalysis. *Language and Cognitive Processes*, **11**, 449–494.

Swinney, D.A. (1979). Lexical Access during Sentence Comprehension: (Re)Consideration of Context Effects. *Journal of Verbal Learning and Verbal Behaviour*, **18**, 645–659.

Tanenhaus, M.K. and Donnenworth-Nolan, S. (1984). Syntactic Context and Lexical Access. *Quarterly Journal of Experimental Psychology*, **36A**, 649–661.

Tanenhaus, M.K., Leiman, J.M. and Seidenberg, M.S. (1979). Evidence for Multiple Stages in the Processing of Ambiguous Words in Syntactic Contexts. *Journal of Verbal Learning and Verbal Behaviour*, **18**, 427–440.

Tanenhaus, M.K., Spivey-Knowlton, M.J., Eberhard, K.M. and Sedivy, J.C. (1995). Integration of Visual and Linguistic Information in Spoken Language Comprehension. *Science*, **268**, 1632–1634.

Tanenhaus, M.K., Spivey-Knowlton, M.J. and Hanna, J.E. (forthcoming). Modelling Discourse Context Effects: A Multiple Constraints Approach. In M.W. Crocker, M.J. Pickering and C. Clifton (eds.) *Architectures and Mechanisms for Language Processing*. Cambridge University Press, Cambridge, UK.

Taraban, R. and McClelland, J.L. (1988). Constituent Attachment and Thematic Role Assignment in Sentence Processing: Influences of Content-Based Expectations. *Journal of Memory and Language*, **27**, 597–632.

Traxler, M. and Pickering, M.J. (forthcoming). Ambiguity Resolution in Sentence Processing: Simplicity or Likelihood? In M.W. Crocker, M.J. Pickering and C. Clifton (eds.) *Architectures and Mechanisms for Language Processing*. Cambridge University Press, Cambridge, UK.

Trueswell, J.C. and Tanenhaus, M.K. (1994). Toward a Lexicalist Framework for Constraint-Based Syntactic Ambiguity Resolution. In C. Clifton, Jr., L. Frazier and K. Rayner (eds.) *Perspectives on Sentence Processing*. Chapter 7, pp. 155–179. Erlbaum, Hillsdale, NJ.

Trueswell, J.C., Tanenhaus, M.K. and Kello, C. (1993). Verb-Specific Constraints in Sentence Processing: Separating Effects of Lexical Preference from Garden-Paths. *Journal of Experimental Psychology: Learning, Memory and Cognition*, **19**, 528–553.

de Vincenzi, M. and Job, R. (1993). Some Observations on the Universality of the Late-Closure Strategy. *Journal of Psycholinguistic Research*, **22**, 189–206.

de Vincenzi, M. and Job, R. (1995). An Investigation of Late Closure: The Role of

Syntax, Thematic Structure, and Pragmatics in Initial and Final Interpretation. *Journal of Experimental Psychology: Learning, Memory and Cognition*, **21**, 1303–1321.

Viterbi, A.J. (1967). Error Bounds for Convolution Codes and an Asymptotically Optimal Decoding Algorithm. *IEEE Transactions on Information Theory*, **13**, 260–269.

Vonk, W. (1984). Eye Movements during the Comprehension of Pronouns. In A. Gale and F. Johnson (eds.) *Theoretical and Applied Aspects of Eye Movement Research*. MIT Press, Cambridge, MA.

Weinberg, A. (1993). Parameters in the Theory of Sentence Processing: Minimal Commitment Theory Goes East. *Journal of Psycholinguistic Research*, **22**, 339–364.

A: Materials for Experiment 1

Condition 1: Verb Bias, Verb Disambiguation

The student thinks that the stone keeps the doors from opening.
The woman said that the german makes the beer she likes best.
The attendant discovered that the museum finds the public a nuisance.
The architect mentioned that the door pulls the building out of balance.
The teacher found out that the schoolboy dares the girls to skip class.
The curate reported that the church draws the money from its funds.
The student thinks that the stone keeps the cloth from blowing away.
The woman said that the german makes the jewellery in his cellar.
The attendant discovered that the museum finds the staff hard working.
The architect mentioned that the door pulls the carpet when it closes.
The teacher found out that the schoolboy dares the children to swear at her.
The curate reported that the church draws the funds from its reserves.
The tourist board reckons that the winter lets the hoteliers have a break.
The experimentalist thinks that the sample means the results will be rubbish.
The enthusiast says that the pub meets the requirements of his group.
The locals are convinced that the lion kills the sheep at night.
The coach thinks that the tennis serves the community by training the children.
The council are proud that the village boasts the best wine in Devon.
The tourist board reckons that the winter lets the seaside towns recover from the pollution.
The experimentalist thinks that the sample means the cattle are BSE free.
The enthusiast says that the pub meets the needs of his association.
The locals are convinced that the lion kills the young cattle for food.
The coach thinks that the tennis serves the children as their main exercise.
The council are proud that the village boasts the largest church hall in the county.

Condition 2: Verb Bias, Noun Disambiguation

The student thinks that the stone keeps were built by the Romans.
The woman said that the german makes are better value than most.
The attendant discovered that the museum finds were destroyed last night.
The architect mentioned that the door pulls are too expensive to use.
The teacher found out that the schoolboy dares are getting out of hand.
The curate reported that the church draws are making lots of money.
The student thinks that the stone keeps are situated in Scotland.
The woman said that the german makes are cheaper than the rest.

The attendant discovered that the museum finds are checked every day.
The architect mentioned that the door pulls are necessary to the design.
The teacher found out that the schoolboy dares were becoming more dangerous.
The curate reported that the church draws are helping the community.
The tourist board reckons that the winter lets are the best available bargain.
The experimentalist thinks that the sample means are larger than predicted.
The enthusiast says that the pub meets are the biggest he has attended.
The locals are convinced that the lion kills are the best means of population control.
The coach thinks that the tennis serves are getting much better.
The council are proud that the village boasts are published in the Times.
The tourist board reckons that the winter lets are scarcer than ever before.
The experimentalist thinks that the sample means are surprisingly low.
The enthusiast says that the pub meets are very popular with his group.
The locals are convinced that the lion kills are making the gods angry.
The coach thinks that the tennis serves are often extremely fast.
The council are proud that the village boasts are the talk of the county.

Condition 3: Noun Bias, Verb Disambiguation

The man told us that the factory machines the doors too roughly.
The foreman knows that the warehouse prices the beer very modestly.
The spokesman stated that the restaurant bills the public far too much.
The director mentioned that the company services the building too frequently.
The man argued that the community schools the girls in the wrong subjects.
The people believe that the student pieces the money together somehow.
The man told us that the factory machines the cloth very quickly.
The foreman knows that the warehouse prices the jewellery very cheaply.
The spokesman stated that the restaurant bills the staff at cost.
The director mentioned that the company services the carpet cleaner regularly.
The man argued that the community schools the children for too long.
The people believe that the student pieces the funds together from thin air.
The rep informed us that the government funds the hoteliers to attend the conference.
The scientists stated that the radiation effects the results of their experiment.
The programmer explained that the system structures the requirements of the users.
The English lord is annoyed that the clan groups the sheep close at hand.
The employee is certain that the office conditions the community to be afraid.
The secretary is keen that the club books the best wine for the conference.
The rep informed us that the government funds the seaside towns to encourage tourism.
The scientists stated that the radiation effects the cattle in the vicinity.
The programmer explained that the system structures the needs of processes hierarchically.
The English lord is annoyed that the clan groups the young cattle by age.

The employee is certain that the office conditions the children into obedience.
The secretary is keen that the club books the largest church in the vicinity.

Condition 4: Noun Bias, Noun Disambiguation

The man told us that the factory machines were built during the war.
The foreman knows that the warehouse prices are better value by far.
The spokesman stated that the restaurant bills were destroyed last night.
The director mentioned that the company services are too expensive at present.
The man argued that the community schools are getting far too full.
The people believe that the student pieces are making the biggest impact.
The man told us that the factory machines are situated near the exit.
The foreman knows that the warehouse prices are cheaper than the others.
The spokesman stated that the restaurant bills are checked by the manager.
The director mentioned that the company services are necessary for morale.
The man argued that the community schools were becoming very elitist.
The people believe that the student pieces are helping advertise the college.
The rep informed us that the government funds are the best available choice.
The people know that the radiation effects are larger than predicted.
The programmer explained that the system structures are the biggest he has worked on.
The English lord is annoyed that the clan groups are the best warriors in Scotland.
The employee is certain that the office conditions are getting much better.
The secretary is keen that the club books are published in the hardback.
The rep informed us that the government funds are scarcer than in previous years.
The people know that the radiation effects are surprisingly moderate.
The programmer explained that the system structures are very popular in the industry.
The English lord is annoyed that the clan groups are making the soldiers look stupid.
The employee is certain that the office conditions are often extremely bad.
The secretary is keen that the club books are the talk of the industry.

B: Materials for Experiment 2

Condition 1: Verb bias, Ambiguous

The student thinks that the stone keeps were built by the Romans.
The woman said that the german makes are better value than most.
The attendant discovered that the museum finds were destroyed last night.
The architect mentioned that the door pulls are too expensive to use.
The teacher found out that the schoolboy dares are getting out of hand.
The curate reported that the church draws are making lots of money.
The student thinks that the stone keeps are situated in Scotland.
The woman said that the german makes are cheaper than the rest.
The attendant discovered that the museum finds were checked every day.
The architect mentioned that the door pulls are necessary to the design.
The teacher found out that the schoolboy dares were becoming more dangerous.
The curate reported that the church draws are helping the community.
The tourist board reckons that the winter lets are the best available bargain.
The experimentalist thinks that the sample means are larger than predicted.
The enthusiast says that the pub meets are the biggest he has attended.
The locals are convinced that the lion kills have made their homes safer.
The coach thinks that the tennis serves have got much faster.
The council are proud that the village boasts are published in the Times.
The tourist board reckons that the winter lets are less popular than ever before.
The experimentalist thinks that the sample means are surprisingly low.
The enthusiast says that the pub meets are very popular with his group.
The locals are convinced that the lion kills have made the gods angry.
The coach thinks that the tennis serves have been ignored by the students.
The council are proud that the village boasts are the talk of the county.

Condition 2: Verb Bias, Unambiguous

The student thinks that the stone keep was built by the Romans.
The woman said that the german make is better value than most.
The attendant discovered that the museum find was destroyed last night.
The architect mentioned that the door pull is too expensive to use.
The teacher found out that the schoolboy dare is getting out of hand.
The curate reported that the church draw is making lots of money.
The student thinks that the stone keep is situated in Scotland.
The woman said that the german make is cheaper than the rest.
The attendant discovered that the museum find was checked every day.

The architect mentioned that the door pull is necessary to the design.
The teacher found out that the schoolboy dare was becoming more dangerous.
The curate reported that the church draw is helping the community.
The tourist board reckons that the winter let is the best available bargain.
The experimentalist thinks that the sample mean is larger than predicted.
The enthusiast says that the pub meet is the biggest he has attended.
The locals are convinced that the lion kill has made their homes safer.
The coach thinks that the tennis serve has got much faster.
The council are proud that the village boast is published in the Times.
The tourist board reckons that the winter let is less popular than ever before.
The experimentalist thinks that the sample mean is surprisingly low.
The enthusiast says that the pub meet is very popular with his group.
The locals are convinced that the lion kill has made the gods angry.
The coach thinks that the tennis serve has been ignored by the students.
The council are proud that the village boast is the talk of the county.

Condition 3: Noun Bias, Ambiguous

The man told us that the factory machines were built during the war.
The foreman knows that the warehouse prices are better value by far.
The spokesman stated that the restaurant bills were destroyed last night.
The director mentioned that the company services are too expensive at present.
The man argued that the community schools are getting far too full.
The people believe that the student pieces are making the biggest impact.
The man told us that the factory machines are situated near the exit.
The foreman knows that the warehouse prices are cheaper than the others.
The spokesman stated that the restaurant bills were checked by the manager.
The director mentioned that the company services are necessary for morale.
The man argued that the community schools were becoming very elitist.
The people believe that the student pieces are helping advertise the college.
The rep informed us that the government funds are the best available choice.
The people know that the radiation effects are larger than predicted.
The programmer explained that the system structures are the biggest he has worked on.
The English lord is annoyed that the clan groups have made their homes on his land.
The employee is certain that the office conditions have got much worse.
The secretary is keen that the club books are published in hardback.
The rep informed us that the government funds are less popular than in previous years.
The people know that the radiation effects are surprisingly moderate.
The programmer explained that the system structures are very popular in the industry.
The English lord is annoyed that the clan groups have made the soldiers look stupid.

The employee is certain that the office conditions have been ignored by the management.
The secretary is keen that the club books are the talk of the industry.

Condition 4: Noun Bias, Unambiguous

The man told us that the factory machine was built during the war.
The foreman knows that the warehouse price is better value by far.
The spokesman stated that the restaurant bill was destroyed last night.
The director mentioned that the company service is too expensive at present.
The man argued that the community school is getting far too full.
The people believe that the student piece is making the biggest impact.
The man told us that the factory machine is situated near the exit.
The foreman knows that the warehouse price is cheaper than the others.
The spokesman stated that the restaurant bill was checked by the manager.
The director mentioned that the company service is necessary for morale.
The man argued that the community school was becoming very elitist.
The people believe that the student piece is helping advertise the college.
The rep informed us that the government fund is the best available choice.
The people know that the radiation effect is larger than predicted.
The programmer explained that the system structure is the biggest he has worked on.
The English lord is annoyed that the clan group has made their homes on his land.
The employee is certain that the office condition has got much worse.
The secretary is keen that the club book is published in hardback.
The rep informed us that the government fund is less popular than in previous years.
The people know that the radiation effect is surprisingly moderate.
The programmer explained that the system structure is very popular in the industry.
The English lord is annoyed that the clan group has made the soldiers look stupid.
The employee is certain that the office condition has been ignored by the management.
The secretary is keen that the club book is the talk of the industry.