Good Reasons for Noting Bad Grammar: Empirical Investigations into the Parsing of Ungrammatical Written English

Jennifer Foster

A Thesis submitted for the Degree of Doctor of Philosophy

Department of Computer Science

University of Dublin

Trinity College

Vol. I

29th October 2004

Declaration

I hereby declare that this thesis is entirely my own work and that it has not been submitted as an exercise for a degree at any other university. I hereby agree that the library may lend or copy this thesis upon request.

________________________________ November 15, 2004

Jennifer Foster

Summary

This thesis is concerned with the parsing of ungrammatical written English sentences. Over a period of eighteen months, a 20,000 word corpus was developed which consists of ungrammatical sentences which were noticed while reading a variety of English texts. Each sentence in this corpus was corrected, producing a second corpus of grammatical sentences. This thesis argues that the compilation of such a corpus is a useful computational linguistic resource, describes the methodological decisions which were made in compiling the corpus, presents and discusses the results of a small questionnaire study which was used to investigate the reliability of the corpus data, analyses the differences between the ungrammatical and grammatical corpus, and makes use of the corpus in three separate parsing studies. The first is a parser which uses a bottom-up active chart parsing algorithm and an error grammar to parse ungrammatical sentences. The error grammar is derived from a conventional grammar and the differences between the sentences in the ungrammatical corpus and the corrected grammatical corpus are used to inform this derivation process. The second piece of parsing research is rooted in the linguistic framework of typed feature structures. An extended notion of a typed feature structure is presented which allows the inconsistent information contained in an agreement error to be stored. A form of relaxed unification is also defined which operates on these feature structures so that sentences containing an agreement error can be parsed. This idea was tested by modifying the parser in the Linguistic Knowledge Base, a widely-used natural language parser/generator which employs typed feature structures as linguistic objects. The modifications allowed sentences from the corpus which contain an agreement error to be parsed. The third contribution is a parser evaluation method which measures a parser's ability to parse ungrammatical sentences by comparing the parses it produces for the ungrammatical sentences from the corpus to the parses it produces for the equivalent grammatical sentences in the corrected corpus. This method is flexible enough to be applied to any type of parser, regardless of the linguistic framework used to encode analyses. The method was applied to two wide-coverage probabilistic parsers, and the results of the evaluation are discussed.

Acknowledgements

I would like to thank my supervisor Dr. Carl Vogel for introducing me to the topic of ungrammaticality and robust parsing, for giving me many helpful comments and much practical support over the years, and for never minding when our grammaticality judgements differed. Thanks to the Forbairt Basic Research Grant SC/97/623 and the TCD Broad Curriculum Postgraduate Teaching Fellowship for supporting this research financially. Thanks to those in the field who reviewed earlier versions of this work for academic papers and those who gave me feedback on this work at talks or poster sessions. Thanks to my family for all their help. Thanks to my colleagues in the Computational Linguistics group, especially to Ann, who finished at the same time as me and who provided much needed moral support (and green tea) in the last few months. Thanks to Fiona for being a good friend and squash partner. Thanks to Louise for keeping me sane throughout. Finally, thanks to Emmet for everything.

Contents

List of Figures

Chapter 1

Introduction

This chapter introduces the topic of the thesis, spells out the thesis goals and explains how the thesis is organized. The discussion of the thesis organization highlights the main results of the thesis research.

1.1 Thesis Topic

The theme of this thesis is the parsing of ungrammatical written sentences. In a system which attempts to process written sentences in some way, e.g. translate them into another language or summarize the information therein, it is necessary to know how the individual words in the sentence are related. It is the role of a parser to perform this kind of analysis. Parsing is the process of scanning a sequence of words from a particular language and grouping these words together in ways that reflect possible meanings of the word sequence. Karttunen and Zwicky (1985) discuss the origins of the term *parsing*, and look at the different characteristics of this term in the fields of traditional grammar, formal linguistics, formal language theory, computer science, artificial intelligence and psycholinguistics. Unless other specified, when I talk about parsing, I will be referring to the analysis of a natural language as opposed to an artificial language, by a machine as opposed to a human.

Designing a parser is a complicated task because a parser's ability to determine possible plausible structures of a sentence depends on its access to a large repository of linguistic information, whether in the form of hand-crafted rules which demarcate the boundaries of a language, or information learnt from a large collection of previously analysed sentences. This task is made slightly simpler by assuming that the sentences which will be encountered by the parser conform to the structural standards of the language in question. Given the human propensity to err and given the fact that many people are forced to produce sentences in a language other than their native one, the reasonableness of this assumption may be questioned. A parser must be able to produce meaningful representations for sentences which are deviant according to human standards, yet which are routinely interpreted correctly by the human brain. If a parser is embedded in a computer-aided language learning system, it might also be required to provide possible corrections for an ungrammatical sentence. In this thesis, attention is restricted to the parsing of ungrammatical English sentences but it is expected that the methods used and the results obtained will be applicable to the parsing of written ungrammatical sentences in other languages.

1.2 Thesis Aims

When I started working on this thesis, my initial aim was investigate the incorporation of inconsistent information into a feature structure system such as those described by Johnson (1988) and Carpenter (1992), with a view to applying the fruits of this labour in the parsing of ungrammatical sentences. After several months grappling with abstract notions of inconsistency, I became convinced that the only sensible way for me to approach the problem of parsing ungrammatical sentences was to look for instances of ungrammatical language and to let the data lead my investigations. As a result I started to record any ungrammatical sentences noticed in my reading material, and this evolved into the development of a corpus of ungrammatical sentences from English texts. This corpus forms the backbone of this thesis, and it is employed in three separate pieces of work. All three are concerned with the parsing of ungrammatical written sentences, two are parsing experiments and one is an evaluation procedure.

The aims of the thesis are given below. These aims evolved during the course of my research into the problem.

1. Demonstrate that collecting a corpus of ungrammatical written English is a useful thing to do in the context of an investigation into the problem of parsing ungrammatical written English.

2. Provide a methodology which can be used to compile a corpus of ungrammatical English sentences.

3. Investigate the kind of ungrammatical sentences that tend to be produced in the sentences of written English.

4. Provide a procedure for evaluating a parser's ability to handle ungrammatical sentences.

5. Investigate parsing approaches which are capable of finding analyses for the kind of ungrammatical sentences found in written English.

In Chapter 8, I will return to this list of thesis objectives, and discuss to what extent they have been fulfilled.

1.3 Thesis Structure

In Chapter 2, I argue that a parser which is attuned to the structure of ungrammatical language is better equipped to parse it accurately. Thus, I advocate a grammar for a parsing system which has two components: the first component describes the structure of the *grammatical* sentences which a parser of written language is likely to encounter, and the second component describes the structure of the *ungrammatical* sentences which the same parser might encounter. In order to test the second component of this grammar, linguistic evidence in the form of real ungrammatical written language is needed. However, in order to decide whether a sentence counts as ungrammatical it is necessary to appeal to the linguistic knowledge of speakers of the language. Therefore, Chapter 2 begins with a review of both kinds of linguistic evidence, namely, evidence from linguistic knowledge and evidence from language in use. The review of evidence from linguistic knowledge is mainly concerned with grammaticality judgements since evidence is often obtained in this form. This section is quite detailed because grammaticality judgements are a controversial form of linguistic evidence. In discussing them, I touch upon fundamental linguistic notions such as the nature of grammar, the role of context and the competence/performance distinction.

Chapter 2 highlights one motivation for the collection of ungrammatical sentences: they should be collected in order to inform the "ill-formed" component of a parser's grammar. In Chapter 3 I describe how I collected ungrammatical English sentences from a variety of written language sources over a period of eighteen months. This error corpus was compiled on the basis of grammaticality judgements I made on sentences in context. I provide a definition of the term *ungrammatical* used when making these judgements. I describe how a second corpus of grammatical sentences was also compiled, containing corrections of the sentences in the error corpus. I argue that context has an important role to play in judging grammaticality and in correcting ungrammatical sentences. I discuss the safeguards which were employed to ensure the data in the error corpus and the corrected corpus is reliable. A small questionnaire study is described which compares my opinions on some of the sentences in the error corpus with those of twenty other English speakers. The results of this questionnaire study lend support to an argument made in Chapters 2 and 3 that opinions on sentences presented in isolation will be more variable than those on sentences presented in a discourse context. Chapter 3 ends with a comparison of my corpus compilation method and the methods used in compiling other error corpora. The error corpus (*corpus1*) and the corrected corpus (*corpus2*) are contained in Appendix A. Some of the work presented in Chapters 2 and 3 is described in Foster and Vogel (2004a).

Chapter 4 is a literature review. Although this thesis is concerned with the problem of parsing ungrammatical written language, this review chapter covers the more general topic of robust parsing. In this chapter, a distinction is made between parsers that attempt to diagnose and provide feedback on errors in an ungrammatical sentence, and parsers which attempt to proceed normally despite the errors in the input sentence. The focus of some of the robust parsing approaches discussed in this chapter is not *ungrammatical* language per se, but *extragrammatical* language, or language which is simply not covered by a parser's grammar. While Chapter 2 contains a review of the grammaticality judgement and linguistic methodology literature within linguistics, the literature that is reviewed in Chapter 4 is of a technical rather than a theoretical nature. The aim of the chapter is to give an overview of the four main error-handling methods that have been used in parsing as opposed to producing a detailed account of any one parser. This chapter provides a background to the robust parsing accounts described in Chapters 5 and 7.

In Chapter 5 I analyse the corrections which were applied to the ungrammatical sentences in *corpus1* to arrive at the grammatical sentences in *corpus2*. These corrections are divided into four groups: substitution, insertion, deletion and composite, and within these four groups further classifications are made. The frequency of each correction type is also provided.

I then describe how some of the results of this analysis were brought to bear in deriving an *error grammar* from a conventional grammar of well-formed language. A chart parsing algorithm which processes this error grammar is presented and a parser which uses the error grammar and which implements the algorithm is evaluated. This parsing approach is compared to the error-handling approaches which were described in Chapter 4. Earlier versions of the work presented in this chapter are described in Foster (2003) and Foster and Vogel (2004b).

The work presented in Chapter 6 represents a contribution to the field of parser evaluation. An earlier version of this work is described in Foster (2004). A method is presented which can be used to evaluate how well a parser parses sentences which are ungrammatical. This method is based on the existence of an error corpus and a corrected corpus such as *corpus1* and *corpus2*. It considers a parse of an ungrammatical sentence to be a good parse if it is similar to the parse produced by the same parser for a corrected grammatical version of the ungrammatical sentence. Thus, this method is really measuring how robust to errors a parser is, as opposed to how good it is at reporting an error. The method is applied to two freely available probabilistic parsers (Collins 1999, Charniak 2000) which have been trained on English sentences from the Penn Treebank (Marcus *et al.* 1993). In this particular application of the evaluation method, the Parseval measures of labelled precision and recall are used to measure parse accuracy (Black *et al.* 1991) and the results of the evaluation are presented and discussed.

Chapter 7 describes an extension I made to the typed feature structure system formally described by Carpenter (1992) and implemented in the Linguistic Knowledge Base (LKB), a system which parses sentences to produce typed feature structure representations of the linguistic content of the sentence, and generates sentences from the semantic information in a typed feature structure (Copestake 2002). This extension allows a certain type of inconsistent information to be stored in a typed feature structure and its application is in the parsing of ungrammatical sentences containing agreement errors. In this chapter I also describe the modification to the LKB which I made to accommodate the extension. An earlier version of the work presented in this chapter can be found in Foster (2000). Although this is the last piece of research to be presented in this thesis, some of it actually predates the work described in previous chapters and is not inspired by the error corpus. However, in order to test the changes that were made to the LKB parser, the sentences from *corpus1* that contain an agreement error were used in conjunction with a large English Head-Driven Phrase Structure Grammar called the English Resource Grammar (Pollard and Sag 1994, Copestake and Flickinger 2000). The work presented in Chapter 7 is compared to other work on typed feature structures and robust parsing.

In Chapter 8, the final chapter of the thesis, I summarize the original contributions described in Chapters 3, 5, 6 and 7, and provide suggestions for how this research can be usefully extended.

Chapter 2

Grammars, Linguistic Evidence and Grammaticality

2.1 Introduction

In Chapter 1, I defined parsing as the process of scanning a sequence of words from a particular language and grouping these words together in ways that reflect possible meanings of the word sequence. In order to perform its task properly, a parser must have at its disposal the conventions according to which the language in question groups meaningful units. In constructing a grammar of a language what one is essentially doing is constructing a theory of these structural conventions. The following are two ways in which such a theory can be tested:

1. By comparing the structural conventions licensed by the theory with the opinions of speakers of the language in question

2. By comparing the structural conventions licensed by the theory with actual language in use

The first approach to testing a grammar, typically associated with methodology in the generative grammar tradition, emphasizes as linguistic evidence the linguistic knowledge of speakers of the language, whereas the second, aligned with corpus linguistics, emphasizes as evidence instances of actual language in use. In this chapter I will explore how linguistic evidence, evidence from linguistic knowledge and evidence from actual language in use, is used in grammar testing. The aim of the chapter is to motivate a type of grammar in which both evidence from linguistic intuition and evidence from actual language in use have important roles to play. I argue that this type of grammar should be used by a parser so that grammatical and ungrammatical sentences can be successfully analysed.

This chapter provides the background for the novel work described in Chapter 3. It is organized as follows: in Section 2.2, evidence from linguistic knowledge is discussed, followed in Section 2.3 by a discussion of evidence from actual language in use. In Section 2.4, I argue that grammars which form the basis of parsing systems should encode not just the structural conventions of the language but also the systematic ways in which these conventions are broken, so that ungrammatical language can be parsed. In order to encode this information, I argue that these grammars need to be based on evidence from linguistic knowledge *and* evidence from language in use: neither on its own is sufficient. The chapter ends with a brief summary in Section 2.5.

2.2 Evidence from Linguistic Knowledge

Laying the foundations of generative grammar, Chomsky claimed that

> The fundamental aim in the linguistic analysis of a language L is to separate the *grammatical* sequences which are the sentences of L from the *ungrammatical* sequences which are not sentences in L and to study the structure of the grammatical sequences. The grammar of L will thus be a device that generates all of the grammatical sequences of L and none of the ungrammatical ones. (Chomsky 1957, p.13)

This characterisation of a linguist's task brings to the fore the notion of *grammaticality* since this is the property distinguishing sentences from non-sentences. The emphasis placed by generative grammar on distinguishing the grammatical from the ungrammatical led to a de-emphasis of the corpus as a source of linguistic evidence. While a corpus can – depending on its scope – provide evidence that a particular string is part of the language it cannot confirm that a string is not, because, no matter how large the corpus, there remains the possibility that the string *is* a grammatical sentence in the language but has just not been recorded in this corpus. A type of evidence which does have the potential to provide negative evidence is that obtained by asking people to judge the grammaticality of a sentence. This type of evidence, which appeals to speakers' linguistic knowledge, came to form the basis of linguistic theories.

The way in which grammaticality judgements have typically been employed as linguistic evidence can be explained as follows: a grammar or linguistic theory accepts as grammatical certain structures and prohibits as ungrammatical others. The division between grammatical and ungrammatical word sequences made by the theory should reflect the division obtained by asking speakers their opinions about the grammaticality of various example word sequences. Consider a linguistic theory which predicts that extraction from possessor phrases is impossible in English. In order to test such a theory, the linguist might invent the following two example sentences, the first which the theory accepts and the second which the theory rejects:[1]

(2.1) *I met a friend of Leslie's.*

(2.2) *Leslie's I met a friend of.*

Speakers of English might then be asked to comment on the grammaticality of sentences (2.1) and (2.2). If they concur with the theory that the first sentence is grammatical, this counts as *positive* evidence for the theory's claims. If they judge the second sentence to be ungrammatical, this counts as *negative* evidence for the theory's claims.

Chomsky's assertion that a generative grammar should reflect a speaker's competence or knowledge of his native language (see, for example, Chomsky

[1]These example sentences are from Pollard and Sag (1994, p.176).

(1965, p.4)) rather than his performance or use of this language was, for those who accepted it, another reason to reject corpus data as a valid source of evidence. This is because observable language contains errors which can be seen to reflect how the language is produced, as opposed to reflecting knowledge of the language. Chomsky (1961, p.130-131) makes this point in the context of a discussion of child language acquisition:

> [Thus] it is absurd to attempt to construct a grammar that describes observed linguistic behaviour directly. The tape-recordings of this conference give a totally false picture of the conceptions of linguistic structure of the various speakers. Nor is this in the least bit surprising. The speaker has represented in his brain a grammar that gives an ideal account of the structure of the sentences of his language, but, when actually faced with the task of speaking or 'understanding', many other factors act upon his underlying linguistic competence to produce actual performance. He may be confused or have several things in mind, change his plans in midstream, etc. Since this is obviously the condition of most actual linguistic performance, a direct record – an actual corpus – is almost useless as it stands, for linguistic analysis of any but the most superficial kind.

This argument cannot, however, be used to *favour* evidence from grammaticality judgements over evidence from corpus data because it has been pointed out that a speaker's judgement is itself a form of performance data and cannot be seen as a direct reflection of the speaker's linguistic knowledge in the same way that utterances that have actually been produced cannot:

> While grammaticality judgements offer a *different* access path

> from language use to competence, they are themselves just another sort of performance, and as such are subject to at least as many confounding factors as production, and likely even more. (Schütze 1996, p.6)

Chomsky, himself, takes quite a pessimistic view of the relationship between grammaticality judgements and grammaticality:

> Any interesting generative grammar will be dealing, for the most part, with mental processes that are far beyond the level of actual or even potential consciousness; furthermore, it is quite apparent that a speaker's reports and viewpoints about his behavior and his competence may be in error. Thus a generative grammar attempts to specify what the speaker actually knows, not what he may report about his knowledge. (Chomsky 1965, p.8)

This view of grammaticality judgements as performance data was reflected in the adoption of the term *acceptability* to refer to the aspect of a sentence that speakers were capable of judging. It was assumed that the property of grammaticality was not amenable to intuition and the challenge then was to relate acceptability to grammaticality.[2]

The use of grammaticality judgements as linguistic data has always been controversial. A large body of literature exists which criticizes the reliability of such data. Much of this criticism, (see, for example, Labov (1972) and Derwing (1979)) dates back to the 1970s, although, more recently, Schütze (1996) has provided a comprehensive review of the issue. Schütze's work is a constructive criticism of the way grammaticality judgements have been

[2] To avoid confusion, I will continue to use the term *grammaticality judgement* throughout the rest of this thesis.

employed in generative linguistics. Instead of rejecting grammaticality judgements, he calls for a better understanding of how they work in the hope that they can be collected and interpreted properly and thus form a trustworthy body of evidence upon which a theory of grammar can be reliably based. Rather than summarizing all the issues which are covered in the grammaticality judgement literature, I discuss those which I consider to be the most pertinent, namely, the role of sentential context, the choice of informant and the significance of graded judgements.[3] I then provide a brief sketch of some metalinguistic tasks which have been used instead of explicit grammaticality judgements to access linguistic knowledge. The section ends with a brief summary.

2.2.1 Isolated Sentences

Chomsky's definition of a language as a set of sentences emphasized the sentence as an individual unit of investigation, and grammaticality judgements have usually been made on sentences appearing in isolation from a discourse context. Lakoff (1971), among others, argues that it is not possible to judge the grammaticality of sentences taken in isolation because a sentence can only be judged relative to a set of presuppositions supplied by the context in which the sentence occurs.

In order to explain why sentential context cannot be ignored, it is useful to consider the information being judged when judging a sentence's gram-

[3] Issues related to the design of judgement elicitation experiments, e.g. the type of scale used to measure judgements or the order or manner in which the items to be judged are presented to the experimental subjects, are not discussed in any detail. This is not because I do not think that these issues are relevant but because I think they impact less on the fundamentals of the thesis.

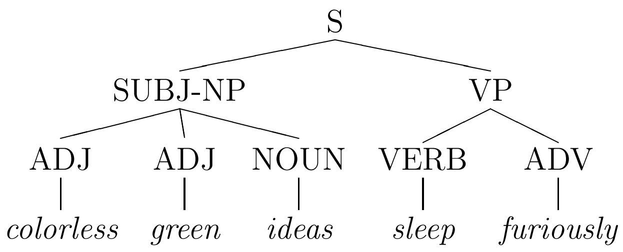


Figure 2.1: Analysis of Sentence (2.3)

maticality. The aspect of a sentence that is being judged is the sentence's structure. If the sentence can be found to have a structure which obeys the structural rules of the language in question (or rather, what the person making the judgement believes are the structural rules) then the sentence will be accepted. So, when a person is judging the grammaticality of a sentence (s)he is essentially carrying out the same task which is required of a computer parser, i.e. looking for a parse or sentential structure which fits in with what (s)he knows about language.

Consider, for example, Chomsky's famous example in (2.3):

(2.3) *Colorless green ideas sleep furiously.* (Chomsky 1957, p.15)

Without a sentential context it is difficult to know what this sentence means. Yet it isn't difficult to determine the structure in Fig. 2.1 as a possible reasonable sentential structure or parse and hence to conclude that the sentence is grammatical. What allows one to determine this reasonable structure? For this example, it is two kinds of linguistic knowledge, both of which fall under the remit of grammar:

1. A knowledge of part-of-speech categories of the words in the sentences, in this case that *colorless* and *green* can be adjectives, that *ideas* can be a noun, that *sleep* can be a verb and that *furiously* can be an adverb.

2. A knowledge of English word order, i.e. that adjectives appear before the noun they are modifying, that an adverb can appear after the verb it modifies and that a subject noun appears before a verb.

Chomsky (1957, p.15-16) argued that the above sentence was meaningless and he used the fact that it is undeniably grammatical to argue that meaning and grammar are two autonomous concepts. An alternative way of looking at this, is to see grammar as an aspect of meaning, the aspect of meaning termed an *episememe* by Bloomfield (1935, p.166). Thus, determining a structure for the sentence or parsing the sentence lends it some meaning. Finding a structure or a parse for a candidate sentence is the same thing as finding its episememe or grammatical meaning and the ability to do this will determine whether the sentence is judged to be grammatical.

In a sentence like (2.3), the grammatical meaning is the only part of the sentence's overall meaning which is transparent. For some sentences, the grammatical meaning is less obvious and only becomes clear when other aspects of the sentence's meaning are in place. An example of such a sentence is *He me* given by Van Dijk (1976). Appearing on its own this sentence may be judged ungrammatical but appearing as part of the discourse in (2.4), its meaning can be determined.

(2.4) A: *Did you hit him?*
B: *No. He me.*

Knowing its meaning implies knowing its grammatical meaning which is given in Fig. 2.2:[4] A different sentential context can lead to the rejection of the same utterance:

[4]The implied verb "hit" is displayed in the structural analysis even though it does not overtly appear in the sentence.

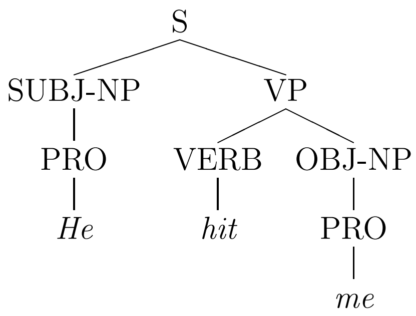


Figure 2.2: Grammatical Meaning of (2.4)

(2.5) A: *Where is he going?*

B: *I don't know. He me.*

In this case, it is very possible that the utterance made by A does not lead to an understanding of B's *He me* response. We could guess that what is meant is the sentence *He didn't tell me* but our linguistic knowledge tells us that, unlike *hit* in the previous context, *didn't tell* can't be omitted from the sentence in this context. Thus, a structure which fits in with the surrounding context *and* with our linguistic knowledge cannot be found, and the utterance must be rejected.

Chomsky (1964) and Ziff (1964) distinguish between those sentences which are interpretable out of context and those which need a context in order to be interpreted. Chomsky (1964) terms the former "grammatical" and the latter "semi-grammatical". This position is a difficult one to maintain, however, since there are many types of sentences which can be classified as semi-grammatical for a variety of different reasons. According to this theory, a sentence such as the final sentence in (2.6) is the same as a sentence such as *He me* appearing in the context given in (2.4), and repeated in (2.7) below.

(2.6) *Jane works in IT and Petra works in marketing. They are always com-*

plaining about their jobs. "I'm not going back to work this afternoon," ***the IT commented*** *over lunch in the park.*

(2.7) A: *Did you hit him?*

B: *No.* ***He me****.*

Both sentences are easy to understand in the context in which they are presented, even though in both sentences, a word has been omitted. However the two sentences are clearly different: in Sentence (2.6) the noun *girl* has been *accidently* left out after the word *IT* resulting in a "semi-grammatical" yet ill-formed sentence; in the sentence *He me* in the context (2.7), a verb has been *purposely* omitted in accordance with the rules of English which sanction ellipsis, resulting in a "semi-grammatical" yet well-formed sentence.

To summarize, the context in which a sentence occurs can *sometimes* be crucial to determining the sentence's structure which in turn is crucial to judging its grammaticality. If the sentential context isn't provided it must be imagined. Levelt *et al.* (1976) report that this is precisely what people do when asked to judge the grammaticality of an isolated sentence. However, since the ability to imagine a plausible sentential context is a skill which varies from individual to individual, the reliability of judgements made on isolated sentences does seem to be in doubt.[5] Crain and Steedman (1985) make a similar point when they argue that the use of isolated sentences does not prevent context from interfering with the judgement process:

> Much current research has adopted the strategy of supposedly eliminating contextual effects by the use of the so-called null context; that is, by presenting sentences in isolation. However, we

[5]Quite apart from natural imaginative ability, the *willingness* to make an effort to imagine a plausible context, is also variable.

> have argued that the context of sentence processing consists in a mental model that is actively constructed by the hearer, using inference on the basis of his or her knowledge of the way the world works. The fact that the experimental situation in question makes a null contribution to the context does not mean that the context is null. It is merely not under the experimenter's control. (Crain and Steedman 1985, p.338)

It seems obvious, therefore, that when a sentence's grammaticality is being judged, its sentential context (either in the form of prior sentences or a description of the scenario in which the sentence is supposed to have been uttered) must be provided. In recent years the issue of sentential context in the grammaticality judgement process seems to have been sidelined in favour of experimental issues such as the type of scale used to measure judgements. The grammaticality judgement experiments described in Bard *et al.* (1996) and Keller and Asudeh (2000), for example, use the sophisticated psychophysical technique of magnitude estimation (Stevens 1956), but still present subjects with artificial sentences without a discourse context. While the choice of a measurement scale is an important decision, the benefits accrued by using better scales will be diminished if the context factor is not controlled for.

2.2.2 The Informant Issue

Another reason to doubt the reliability of grammaticality judgements are the informants who make the judgements. These informants can be the actual linguists (who may or may not be responsible for devising the linguistic analysis being tested) or a group of *linguistically naive* informants. An informant is linguistically naive if (s)he has not studied linguistics.

Before discussing the problems associated with the type of informant, I will discuss a fundamental problem which affects both linguists and naive informants. This problem relates to the concept of grammaticality itself. There is no general consensus within the linguistics community on what exactly the terms *grammaticality*, *grammatical* and *ungrammatical* mean, because the concept of *grammar* and the kind of linguistic information contained therein is itself open to interpretation. Sampson (2001d) makes the point that Chomsky changed his mind between the publication of *Syntactic Structures* and *Aspects of the Theory of Syntax* over whether or not selectional restrictions on verbs should be included in a grammar of a language. In Sentence (2.8), a selectional restriction on the verb *disperse* has been violated by the use of the subject noun *book*. Chomsky originally felt that a sentence such as (2.8) should not be considered ungrammatical but later felt that it should.

(2.8) *The book dispersed.*

Modern lexicalist grammatical theories such as Head-Driven Phrase Structure Grammar (Pollard and Sag 1994) and Lexical Functional Grammar (Bresnan and Kaplan 1982) include selectional restrictions in the lexical entries for the verbs of the language, and, since the lexicon is considered to be part of the grammar, a sentence such as (2.8) would be ruled out as ungrammatical by such theories. For a linguist such as Sampson, this sentence is not ungrammatical because the reason why it is strange has, according to him, nothing to do with language and everything to do with the behaviour of books in the real world. The point I am making here is that if linguists are applying different concepts of grammaticality when making grammaticality judgements, the data they collect will be difficult to compare. This issue will also affect naive informants who are providing grammaticality judgements for linguists and who have been instructed by linguists on how to make a grammaticality judgement.

Linguists as Informants

One of the long-standing criticisms of generative grammar methodology is that linguists working within this tradition have tended to act as their own informants, resulting in unreliable evidence. For examples of such criticism see Labov (1972), Greenbaum (1976b), Greenbaum (1977), Lakoff (1976), Ringen (1977) and Snow and Meijer (1976). Two of the most frequently heard criticisms of this practice are as follows:

1. **Experimenter Bias**: it is possible that a linguist's introspective judgements may be influenced by the predictions made by the theory under his construction. The credibility of this kind of evidence is further undermined if the sentences whose grammaticality is judged by the linguist have been invented by the same linguist. If the linguist's judgements can be biased by the predictions of the theory, then the linguist's ability to generate suitable test sentences is surely also compromised. According to Greenbaum (1977), corpus linguistic evidence has often been criticized for not being a complete sample of the relevant grammatical constructions[6] but the same criticism can be levelled at the language sample dreamed up by a linguist.

2. **Clouded judgements due to lengthy exposure to example sentences**: If a linguist uses as linguistic evidence his own grammaticality judgements, then these judgements will not have been obtained in a controlled, experimental setting and thus there is no limit to the time

[6]The availability of very large corpora makes this criticism less valid now than it was during the sixties and seventies when the generative grammar methodology was predominant within linguistics

> the linguist can spend thinking about the grammaticality of a sentence. This potentially lengthy exposure to a sentence, particularly one with no anchoring context, can interfere with the introspective process, clouding the judgement.

Newmeyer (1983) has attempted to defend the "linguist as informant" model: according to him, it is quicker and easier if the linguist provides his own data rather than carrying out expensive and timely experiments, and the linguist is in a better position to make grammaticality judgements than the linguistically naive informant since the linguist has been specifically trained to make such judgements properly. Newmeyer's first defence cannot really be justified if the reliability of the data provided (however quickly and easily) by the linguist is in doubt. The second defence finds support in an experiment carried out by Snow and Meijer (1976) which ellicited grammaticality judgements from a group of linguists and a group of non-linguists and which found greater inter-subject agreement and intra-subject consistency among the linguist group. Snow and Meijer give two alternative interpretations of these results: either the linguists are better equipped than the non-linguists to ignore factors unrelated to the structure of the sentences which could influence judgements (this is the position taken by Newmeyer) or the linguists are being biased in their judgements by the predictions of the syntactic theory they are working with.[7] Since it is not clear which of these interpretations is the correct one, it is not possible to conclude on the basis of this experiment that linguists are a better choice than non-linguists as a source of grammaticality judgements.

[7] Snow and Meijer's article does not tell us whether the linguists were all working with the same linguistic theory, or whether their interests were related to the phenomena tested.

To summarize, it has been widely acknowledged that it is scientifically unsound for linguists to use their own linguistic intuitions to test their own grammatical theories.[8] However, the issue of bias no longer applies when a theory is tested using the intuition of linguists who are not responsible for devising the theory and whose work does not overlap significantly with the theory's claims. Presuming that such unbiased linguists can be found, the question remains of whether an unbiased linguist is more cabable of judging grammaticality than someone who doesn't know a subcategorization frame from a climbing one. In the next section, I turn to the issue of using such linguistically naive informants.

Non-linguists as Informants

A serious problem with using linguistically naive speakers to make grammaticality judgements is that it is difficult to instruct the informants on what exactly they are supposed to be judging such that *all* informants are judging the same thing, that is, the grammaticality of the example sentences. Earlier in Section 2.2.2, I touched upon the problem of disagreement amongst linguists concerning the term "grammaticality". However, I will assume in the present discussion that the linguist who is asking naive informants for their judgements has a clear idea of what (s)he means by the term.[9] The problem then is how to communicate this.

[8] Sampson (2001d) makes the important point that there is nothing scientifically suspect about linguists appealing to their intuitions when devising (as opposed to testing) their grammatical theories.

[9] In a keynote speech at the *International Conference on Linguistic Evidence, Tübingen, 2004*, Schütze made the depressing point that experimenters can sometimes contradict themselves or be so vague in their instructions to subjects that it is obvious that they themselves do not have a clear idea of what the subjects are supposed to be judging.

Using the word "grammatical" in the instructions to the informants brings with it the problem that the related word "grammar" conjures up for many informants prescriptive rules of writing style learnt at school, e.g. rules which proscribe against ending a sentence with a preposition or starting a sentence with *and*. The exact nature of these rules and how firmly entrenched they are in the mind of the informant will depend on when and where (s)he was educated. For some informants, e.g. those educated in the fifties, a sentence which ends in a preposition *is* ill-formed and something that they would genuinely avoid when writing. Those educated in the eighties might be aware of the rule but pay it little attention. The challenge, then, is to word the instructions so that such rules do not come into play in the judgement process.

Replacing the word "grammatical" with the word "acceptable" does little to ameliorate the problem and may even make things worse because for some the word has the same implications as "grammatical"[10] but for others it could deflect attention away from the sentence's structure and onto some other aspect such as the plausibility of the situation described by the sentence. Attempting to paraphrase terms such as "grammatical" or "acceptable" is also difficult. Schütze (1996, p.81) discusses an experiment carried out by Mohan (1976), in which subjects were instructed to

> Consider each of the sentences and decide if it would be possible that you would say this in conversation.

Schütze criticises this for being too "usage-oriented" and it is easy to see what he means. Asking subjects to judge a sentence based on whether it would be

[10] A letter to Dublin's *Evening Herald*, dated 21/8/2003, is a case in point. A woman complains about the newspaper's practice of starting a sentence with *and*. According to her, this is "not **acceptable**, as everyone knows a sentence is continued, not started, with 'and' ".

a normal thing to say in conversation brings with it the danger that subjects will focus less on the sentence's structure and more on its perceived frequency or on its closeness to everyday language.

Snow and Meijer (1976) interpreted the lower degree of intra-subject consistency and inter-subject agreement among the linguistically naive subjects than among the linguists in their experiment as evidence that the act of judging grammaticality is a secondary linguistic ability which is separate from the primary linguistic abilities of speaking and understanding, and which requires explicit training. They carried out a further experiment (Snow and Meijer 1976) which compared the grammaticality judgements of native Dutch speakers with those of non-native Dutch speakers of varying fluency levels and with English as a native language. The results supported their belief that the ability to make a grammaticality judgement is independent of the ability to communicate in a language: the level of agreement between the native Dutch speakers and the non-native speakers who spoke Dutch almost as well as English was no higher than the level of agreement between the native Dutch speakers and those non-native speakers who spoke Dutch considerably less well than English. Further evidence that the ability to make a grammaticality judgement is a skill which must be passed on through explicit instruction comes from studies which determine a link between literacy and the ability to reason about grammaticality (see Schütze (1996, p.124-125) for an overview). James (1998, p.50-51), in a discussion about teaching English as a second language, argues that even educated native speakers are sometimes unable to identify ungrammatical constructions.

To summarize, there are two main problems associated with asking linguistically naive people to make grammaticality judgements: the first is that the instructions given to the informants have to be unambiguous enough to ensure that they are being interpreted in the same way by everyone, and the second problem is that it is still an open issue whether judging the grammaticality of a sentence is something that someone without a grounding in grammar can actually do.

2.2.3 Degrees of Grammaticality

Chomsky's definition of a linguist's task, presented earlier, assumed a binary distinction between the grammatical and the ungrammatical. Such a distinction is easy to maintain in the field of formal language theory but becomes more difficult when one moves from studying formal languages to studying natural languages. The binary distinction does not manifest itself in the judgement data: not only do informants not agree with each other over the grammaticality status of many sentences, they often encounter sentences which do not seem fully grammatical yet which do not seem fully ungrammatical either.

Such disagreement and uncertainty can to a certain extent be traced back to issues discussed in Sections 2.2.2. Inter-subject variation, for example, is bound to arise if informants have differing ideas about what aspect of a sentence they are supposed to be giving their opinion on, or if the ability to perform the judgement task varies from individual to individual. Intra-subject variation will occur if the instructions given to the subjects are vague enough to allow subjects to change their concept of what they are supposed to be judging during the course of the experiment.

The issue of sentential context discussed in Section 2.2.1 is also relevant: it could be argued that the binary grammatical/ungrammatical distinction could be maintained if the mathematical view of language as a set of sentences was rejected and replaced by one in which sentences only existed as part of a real communication. Viewing language as a set of sentences forces us to decide upon the set membership of an utterance such as *He me* when it is clear from the discussion in Section 2.2.1 that it belongs both to the set of sentences (when appearing in the discourse context (2.4)) and to the set of non-sentences (when appearing in the discourse context (2.5)). According to Newmeyer (1983), the unclear cases are those that "with a little imagination can be given a context in which they sound perfectly acceptable." If this is true, supplying a context along with a sentence takes the imagination element out of the judgement process and should result in fewer unclear cases (a similar point was made by Snow (1975)). These and experimental factors discussed by Schütze (1996, chapters 4 and 5), e.g. the order in which sentences are presented to informants and the measurement scale used, need to be properly understood and controlled for, before the uncertain and inconsistent judgements can be interpreted with confidence.

Even assuming, firstly, that all the afore-mentioned problems with the judgement elicitation process have been ironed-out, and secondly, that uncertain judgements are still being produced, the following question remains: do the fuzzy judgements made by informants reflect their actual linguistic competence, making the concept of grammaticality graded *or* do other performance factors (e.g. short-term memory limitations) interact with the dichotomous property of grammaticality to produce graded judgements? This latter view is taken by Langendoen and Bever (1977). Fuzzy Grammar (Lakoff 1973) and, more recently, Optimality Theory (Prince and Smolensky 1993) are theories of language which have tried to accommodate the former view. The view of grammaticality as an inherently graded concept has led to the use of measurement scales such as magnitude estimation to elicit judgements. The benefit of viewing grammaticality in this way is that it takes seriously the notion that language is variable and that some grammatical constraints are more firmly entrenched than others. The drawback of viewing grammaticality as graded is that this does not necessarily confront the variety of reasons why a sentence may be at an intermediate point between grammaticality and ungrammaticality (see Sentences (2.6) and (2.7) in Section 2.2.1). Schütze (1996, Chapter Five) makes the point that the exact nature of a speaker's competence grammar is not perhaps accessible to empirical investigation, and as such, it is irrelevant whether we choose to explain graded judgements in terms of performance or in terms of competence. Such a view is echoed by those linguists (see, for example, Barlow and Kemmer (2000)) who regard the competence/performance distinction as an obstacle to understanding how people produce and comprehend language. According to this view, performance is everything, and language can only be properly understood if attempts to explain it are couched in terms of the actual processes of comprehension and production.

2.2.4 Experimental Alternatives to Grammaticality Judgements

Chomsky (1961, p.130) advocated studying grammaticality in "devious and clever ways". Experimental tasks have been used by Langendoen *et al.* (1977) and Quirk and Svartvik (1966) which attempt to determine sentence grammaticality in an indirect fashion, as a complement or an alternative to introspective data. In the compliance tests used by Quirk and Svartvik, subjects are asked to turn declarative sentences (some of them ungrammatical) into questions. Tottie (1976) applauds the use of such tests in determining grammaticality because subjects' attention is directed away from the relevant issue making them less likely to bring into play perceived ideas about correct language (see Section 2.2.2), and because, if a subject corrects an ungrammatical sentence while turning it into a question, this provides insight into why the sentence is ungrammatical. In more recent years experiments have been used to indirectly assess the grammaticality of a sentence by measuring subjects' reading times, eye movements and event-related brain potentials when confronted with the sentence (see for example Kepser and Reis (2004)). Although this kind of data does not suffer from some of the reliability problems associated with introspective data, the challenge remains of explaining what exactly is being measured, and how it relates to the grammatical structure of the sentences in the language.

2.2.5 Summary

In the previous sections, I have discussed the use of linguistic knowledge to ascertain the grammaticality status of possible sentences in the language under study. This linguistic knowledge has predominantly taken the form of introspective judgements. I looked at what I consider to be the most fundamental problems associated with introspective grammaticality judgements, namely the lack of sentential context, the choice of informants and the meaning of graded grammaticality judgements. With regard to the issue of context, I made it clear in Secton 2.2.1 that I think that grammaticality judgements made on isolated sentences are flawed. This is because context interacts with the judgement process. To ensure that it interacts with the judgement process in the same way for all those whose opinions are sought, it needs to be supplied. The problem of knowing whose judgements to trust was discussed in Section 2.2.2 but the only certainties that can be drawn from this discussion are that linguists should not test their grammars on themselves alone, and when using naive informants, due attention should be paid to the instructions that they are given so that they are applied in a consistent way. In Section 2.2.3, I discussed the various interpretations of graded judgements, before concluding in Section 2.2.4 with a brief discussion of the use of methods which access linguistic intuition but in ways less direct than explicit judgement elicitations. The references used in this section stretch over six decades, demonstrating that the "linguistic knowledge/grammaticality judgement as valid evidence" debate has been, and still is, a major concern within the field of linguistics.

2.3 Corpus Evidence

In the previous section I looked at a method for testing a theory of linguistic structure which is based on the opinions of speakers of the language in question with regard to the notion of sentence grammaticality. A second method for testing a theory of linguistic structure is to collect a set of sentences which have been naturally produced[11] and to see if the structures of these sentences coincide with the claims about structure made by the theory.

Within this view of language, the decision of whether to regard a particular structure as grammatical is made by looking for instances of that structure in language use. This empirical approach takes many forms, but I will generalize by dividing them into those that maintain the competence/performance distinction and those which abandon it: the first still views the grammar as a theory of linguistic competence and is to a certain extent still reliant on the ability of speakers of the language to distinguish the grammatical from the ungrammatical; the second views the competence/performance distinc-

[11] As opposed to those sentences whose sole reason for existence is to have their grammaticality judged. I use the term *naturalistic* to refer to naturally produced sentences.

tion as irrelevant and sees all naturalistic sentences as worthy of description. This view is compatible with the practical demands of the natural language engineering community, which in recent times has seen the development of grammars which are a direct reflection of language use with no real concern for whether the structures described in the grammar are grammatical. The first approach is described in Section 2.3.1, and the second in Section 2.3.2.

2.3.1 Empirical Competence Grammars

Sampson (2001d) confronts the two main arguments against the use of corpus data as linguistic evidence, i.e. the inability of corpus data to provide evidence that a string is *not* part of the language and the competence/performance distinction. To reiterate what was said at the beginning of Section 2.2, the negative evidence argument proceeds as follows: with a corpus of naturalistic sentences as evidence it is impossible to conclude from the fact that a certain string is not in the corpus that the same string is not in the language.[12] For example, a linguist may be interested in the grammaticality status of the strings (2.1) and (2.2), repeated for convenience as (2.9) and (2.10) below.

(2.9) *I met a friend of Leslie's.*

(2.10) *Leslie's, I met a friend of.*

[12] The negative evidence issue is used in the "poverty of the stimulus" argument in the study of child language acquisition. Proponents of this argue that the grammar acquired by a child could not be inferred from the utterances the child is exposed to, and so learning must be innate rather than data-driven. They argue that one of the reasons why children cannot infer a grammar from the environment alone is that instances of ungrammaticality do not tend to be pointed out to them, i.e. they do not receive negative evidence.

(S)he might not be able to find an occurrence of the second string in any available corpus. Can the linguist conclude its ungrammaticality from its non-occurrence?

Sampson answers this question in the affirmative. According to him, if the theory is constructed so that it excludes sentences whose structure has not actually been observed, then negative evidence becomes irrelevant. In order to exclude a sentence such as (2.10) from the grammar, it is not necessary to verify its ungrammaticality, i.e. it is not necessary to obtain negative evidence. It is enough not to have observed the sentence in practice. Once the sentence is observed, then this observation has the potential to count as a refutation of a grammar which excludes the sentence, and the grammar will need to be modified accordingly. As Sampson notes, this is Popper's view of the nature of a scientific theory: it should maximize the number of statements it makes which are refutable by observable evidence.

The second argument against the use of corpus data as linguistic data rests on the competence/performance distinction. A widely accepted view within linguistics is that grammars are theories of linguistic competence rather than performance and hence should prohibit ungrammatical yet naturalistic sentences such as Sentence (2.11):

(2.11) *Want to saving money?*

However, according to Sampson's first argument above, if such a string is observed, it should serve as a refutation of any grammar which prohibits it. A refutation falsifies a theory so the grammar in question would need to be modified so that it no longer prohibits this sentence, thus turning the grammar into something that reflects performance rather than competence. Sampson argues that the grammar should not be changed to accommodate such an observation since our knowledge that people make mistakes in language (such as omitting a word, using the wrong verb form, etc.) should allow us to relate this sentence to another sentence such as (2.12) or (2.13) below, both of which *are* accommodated by the grammar, thereby discounting the observation as a genuine refutation.

(2.12) *Want to save money?*

(2.13) *Want to start saving money?*

Although Sampson argues that a person's ability to dismiss Sentence (2.11) as a genuine refutation of a theory utilizes non-linguistic knowledge (i.e. the knowledge that mistakes occur) and is not therefore an appeal to linguistic knowledge, it could also be argued that what one is actually doing here is making a grammaticality judgement.

Baldwin *et al.* (2004) describe how a broad coverage competence grammar, the English Resource Grammar (ERG)(Copestake and Flickinger 2000), was tested using a combination of corpus evidence *and* evidence from grammaticality judgements. The corpus evidence consisted of spoken language transcriptions, emails and sentences taken from the British National Corpus (Burnard 2000). The grammaticality judgements were made by the grammar developers on both corpus sentences and invented ones in order to decide what to include and what to exclude from the grammar.[13]

The similarities between the ERG approach to grammar and Sampson's approach are that both believe that grammars should reflect linguistic competence and that grammar testing should be grounded in corpus data. Because they believe that grammars should reflect competence, they do not

[13]Section 2.2.2 contained a discussion of the argument that a grammar should not be tested by appealing to the introspective judgements of those responsible for devising it. The evidence used to test this grammar is subject to same criticism but this criticism is tempered by the fact that the judgements are in many cases made on naturalistic sentences.

believe that the ungrammatical sentences in a corpus should be described by a grammar. A similar position is taken by Aarts (1991) who believes that a grammar should be tested on corpus data, but a division should be made between corpus sentences which ought to be described by a grammar, and corpus sentences which should be ruled out by a grammar.

Sampson and the ERG developers differ, however, in their attitude to grammaticality judgements on invented sentences which have been a source of positive and negative evidence within the generative grammar tradition. Sampson, as explained above, takes the view that only positive evidence in the form of naturalistic grammatical sentences is necessary. Naturalistic ungrammatical sentences are not considered to be evidence because there are non-linguistic factors which explain their existence. Invented non-naturalistic sentences, be they grammatical such as Sentence (2.9) or ungrammatical such as Sentence (2.10) are simply irrelevant, and Sampson seems to be of the opinion that inventing such strings is an unnecessary exercise. Those responsible for the English Resource Grammar take the traditional generative grammar view that the boundaries of a grammar can be tested by inventing sentences and judging their grammaticality, in order to arrive at positive *and* negative evidence.

In this section, I have discussed two approaches to grammar development which assume that naturalistic evidence from corpora is needed to test a grammar. However, both of these approaches still have at their heart a commitment to describing linguistic competence: both believe that naturalistic ungrammatical sentences such as Sentence (2.11) should *not* be described by a grammar. Thus, both approaches are dependent on the ability of speakers to recognize an ungrammatical sentence when confronted with one. In the next section, I look at approaches to grammar development where the competence/performance distinction is unimportant.

2.3.2 Empirical Grammars: No Competence/Performance Distinction

One of the features of usage-based theories of language (Barlow and Kemmer 2000) is that the distinction between competence and performance is irrelevant:

> "Instead of viewing language processing as something external to the system, which happens only to the outputs of competence, processing is rather to be seen as an intrinsic part of the linguistic knowledge system, which cannot be treated separately from it. 'Performance errors', for example, are not viewed as due exclusively to 'processing factors', and thus are not treated as a completely separate phenomenon from other utterances not licensed by competence. Instead, all linguistic productions are seen as simply in conformance, or non-conformance, with linguistic norms, to differing degrees." (Kemmer and Barlow 2000, p.xi).

Once this distinction is rejected, the argument which employs the competence/performance distinction to invalidate corpus evidence can also be dismissed:

> "Because the linguistic system is so closely tied to usage, it follows that theories of language should be grounded in an observation of data from actual uses of language. In linguistics, the standard methodology relies on constructed examples with no naturally occurring context of production (or comprehension). This practice derives from the basic assumption, referred to above in the

> discussion of the competence/performance distinction, of a very indirect relation between linguistic knowledge and acts of language use." (Kemmer and Barlow 2000, p.xv)

Within the field of computational linguistics, the move towards grammars based on actual language in use was motivated mainly for practical rather than theoretical reasons. This trend has emerged side-by-side with the increasing use of probabilistic techniques. A motivation for the use of probabilities within the field of parsing was to disambiguate between several syntactic structures licensed by a parser's grammar for a particular sentence. One of the ways in which the most likely structure could be chosen was to exploit information about the relative frequency of certain syntactic structures. The obvious way to obtain sensible frequency information was to look at a large collection of naturalistic sentences. The logical progression from this was to use these sentences not just to obtain frequency information but to obtain the grammar itself. Since the sentences were collected with no particular regard for the notion of grammaticality, the grammar derived from these sentences was also agnostic in this respect. Grammars derived in this way can no longer be said to reflect linguistic competence since they are based directly on actual language in use.

The most frequently used set of naturalistic sentences within parsing is the Penn Treebank (Marcus *et al.* 1993) which is a 4.5 million word corpus of American English sentences from a variety of texts including Wall Street Journal articles and the Brown Corpus (Francis and Kučera 1982). The structures of the sentences in this corpus have been annotated[14] making it easier to read a grammar directly from them: this grammatical annotation transforms

[14]The sentences are first parsed using the Fidditch parser (Hindle 1994) (see Section 4.2.3) and then hand-corrected by a group of people.

it from a corpus to a treebank. Other well-known treebanks are the Susanne corpus (Sampson 1995) and the Redwoods treebank (Oepen *et al.* 2002) for English, and the TIGER treebank for German (Brants *et al.* 2002). A current source of active research is the development of treebanks for minority languages.

2.3.3 Summary

In this section I have discussed linguistic evidence which is based on observed instances of language in use. These corpus based approaches to grammar share a commitment to testing the claims of a grammar by comparing the structures predicted by the grammar to the structures which are produced by speakers of the language. I divided the corpus-based approaches into those that still depend on the grammatical/ungrammatical distinction, and those for which this distinction is irrelevant.

2.4 Linguistic Evidence and Parsing Ungrammatical Language

An important debate within the linguistics community has concerned what it is a grammar is supposed to model: should a grammar model competence or performance? Should a grammar reflect a psychological reality or a social reality? Lamb (2000), for example, distinguishes between a "theory of the linguistic extension" which is a theory of the utterances produced by a speaker or community, and a "theory of the linguistic system" which is a theory of the human cognitive system which is capable of producing and understanding such utterances. In the practical domain of natural language processing, there is no such debate. The grammar of a computer parser which is to form part of a practical application must be a theory of the linguistic extension and must describe the productions of a speech community. In proposing the competence/performance distinction, Chomsky remarked that the language produced by a speech community is rife with slips and imperfections. Therefore, if a computer parser has to accurately parse actual language, it will have to accurately parse imperfect language. It will be able to do this if it is equipped with some knowledge of deviant sentence structures. In Section 2.4.1, I explain why this knowledge will not be contained in a competence grammar, whether it is tested using corpus evidence along the lines suggested by Sampson (2001d) (see Section 2.3.1) or in the traditional generative grammar sense using grammaticality judgements (see Section 2.2). Nor will it be contained in the grammars used by probabilistic parsers which are gleaned directly from a corpus (see Section 2.3.2). An alternative definition of grammar is provided in Section 2.4.2 which takes into account deviant sentential structures. The type of linguistic evidence which is needed to test this kind of grammar is discussed.

2.4.1 Incompleteness of Previous Approaches

A competence grammar distinguishes between the grammatical and the ungrammatical and purposely describes only grammatical sentences. Baldwin *et al.* (2004) make the reasonable point that, if a grammar is to form the basis of a natural language processing system which performs not just sentence parsing but also sentence generation (e.g. the Linguistic Knowledge Base (LKB)(Copestake 2002)), it should not be able to generate ungrammatical sentences. A parser using such a competence grammar will reject ungrammatical sentences outright. However, a parser which gives the response "no" or "ungrammatical" to a sentence such as (2.11), repeated for convenience as Sentence (2.14), may be capable of distinguishing between the grammatical and ungrammatical but of what practical use is this ability if it cannot hint at the meaning of an utterance whose ill-formedness is quite commonplace?

(2.14) *Want to* ***saving*** *money?*

A parser whose grammar is derived directly from a corpus of naturally occurring sentences is not concerned with whether or not a sentence is grammatical. Typically, grammaticality is assumed, and this assumption will be quite accurate if the corpus sentences come from a high-quality newspaper such as *The Wall Street Journal*. The fact that such grammars do not purposely set out to exclude ungrammatical sentences together with the fact that such grammars are generally based upon a large body of data means that parsers equipped with such grammars are quite likely to return a parse for an ungrammatical sentence. However, since such a parser does not have a concept of ungrammaticality, it will not be aware that there is something deviant about the sentence, with the result that the parse it produces for the sentence will not necessarily be the correct one. For example, Charniak's most recent parser[15] (Charniak 2000) will provide the reasonable parse in Fig. 2.3 for Sentence (2.14) but it is less successful, for example, on the ill-formed (2.15), returning the parse in Fig. 2.4.

(2.15) *The closure* ***in*** *computed breadth-first.*

I return to the issue of probabilistic parsers and ungrammatical sentences in Chapters 4 and 6.

[15] Downloaded as *nlparser.tar.gz* from *ftp://ftp.cs.brown.edu/pub/nlparser/* in March 2003.

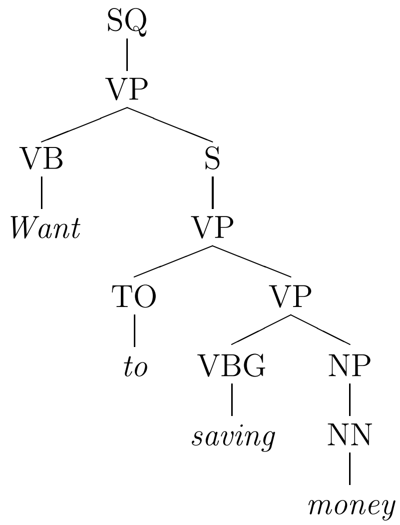


Figure 2.3: Parse Returned by Probabilistic Parser for Sentence (2.14)

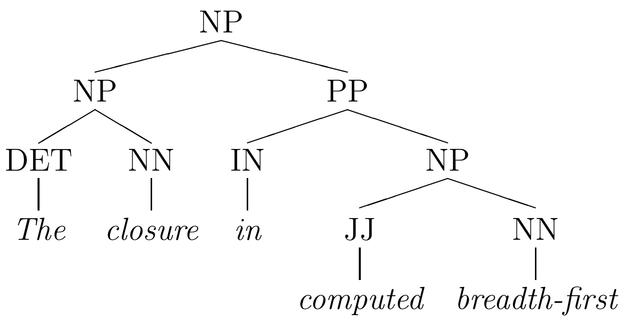


Figure 2.4: Parse Returned by Probabilistic Parser for Sentence (2.15)

2.4.2 A Parser's Grammar

The following are the requirements for the type of grammar which I believe should be used by a parsing system:

Definition 1 ***Requirements for a Parser's Grammar***

1. *The grammar should have a concept of the kind of grammatical or well-formed sentences that occur in language production.*

2. *The grammar should have a concept of the kind of ungrammatical or ill-formed sentences that occur in language production.*

It is difficult to define this conception of grammar in terms of the competence/performance distinction. It is clearly related to linguistic performance since it aims to describe the way language is used. However, unlike the kind of grammars used by probabilistic parsers, this grammar does distinguish between the grammatical and the ungrammatical. This distinction is binary, but this does not mean that the rules in each component of the grammar cannot be probabilistic. A linguistic structure described in the well-formed component of the grammar could be assigned a probability based on how frequently this structure appears in well-formed naturalistic data. Similarly, a linguistic structure described in the ill-formed component of the grammar could be assigned a probability based on how frequently this structure shows up in ill-formed naturalistic data.

The first requirement above is fulfilled by an empirical competence grammar which has been tested using a combination of grammaticality judgements and corpus evidence (see Section 2.3.1). The second requirement is not fulfilled, because an empirical competence grammar does not describe ungrammatical sentences. What kind of evidence is needed in order to test the additional component of the grammar, the part of the grammar which describes ungrammatical sentences? Since this grammar is to form the basis of a parsing system, its description of ungrammaticality must reflect the kind of ungrammaticality that actually occurs in language. This means that naturalistic ungrammatical sentences (see Sentence (2.11) in Section 2.3.1) will be needed as evidence rather than imagined ones (see Sentence (2.10) in Section 2.3.1). Baldwin *et al.* (2004) argue that naturalistic ungrammatical sentences such as Sentence (2.11) constitute "haphazard noise" and are useful only to test that a grammar does not overgenerate. I am arguing, that a grammar that is capable of generating the kind of ungrammatical sentences that people actually produce, is not guilty of overgeneration, provided that the grammar knows that these kinds of sentences are ungrammatical.

Therefore, to test the second "ungrammatical" component of the grammar, which is essentially a theory of real ungrammaticality, it is necessary to collect a corpus of naturalistic sentences which are considered by speakers of the language to be ungrammatical. The methodological issues involved in the construction of such a corpus are described in the next chapter.

2.5 Conclusion

In this chapter, I have explored two types of evidence used to test a theory of a language's sentencial structure, namely, grammaticality judgements and corpus evidence. The issue of linguistic evidence was related to the problem of parsing, in particular, the problem of parsing ungrammatical language. I argued that a parser which is capable of providing *accurate* analyses for ill-formed sentences should be equipped with a theory of ill-formedness or ungrammaticality. The two types of linguistic evidence already discussed should be brought to bear in the formation of such a theory: authentic examples of ungrammatical language should form the basis of a theory of ungrammaticality but in order to identify these examples, grammaticality judgements will also need to be carried out. A point that I have made in this chapter and one which is often ignored, is that grammaticality judgements should only be made on sentences in context. In the following chapter I describe a method for collecting evidence for a theory of ungrammaticality which involves judging naturalistic sentences in their natural context.

Chapter 3

Compiling a Corpus of Ungrammatical Language

3.1 Introduction

In this chapter, I describe the collection of a corpus of ungrammatical written English sentences. In Section 2.4.2, I defined a grammar with two components (Def. 1): the first component models the well-formed structures that are produced for a given language and the second models the ill-formed structures that are produced for the same language. The corpus of ungrammatical language described in this chapter serves as evidence for the second component of this kind of grammar. It counts as corpus evidence since the sentences it contains are naturalistic ones reflecting language in use, and it counts as evidence from linguistic knowledge since these sentences have all been judged to be ungrammatical on the basis of my linguistic knowledge.

This chapter is mainly concerned with the methodological issues involved in the compilation of this corpus. In Section 3.2, I describe the type of texts from which the ungrammatical corpus sentences were taken. In Section 3.3, I present the operational definition of *ungrammatical* which I used to decide whether or not a sentence should be included in the corpus. In Section 3.4, I explain how the ungrammatical sentences were classified based on the kind of corrections that would need to be applied in order to make them grammatical. All the ungrammatical sentences in the corpus were encountered while reading which means that the context in which they occurred was available. This issue of sentential context, first discussed in Section 2.2.1, is discussed again in Section 3.5 with particular reference to the corpus compilation process. In Section 3.6, the reliability of the data in my corpus is discussed: in Section 3.6.1, I present examples of sentences which could have been included in the corpus but which were not because, for a variety of reasons, they constituted unclear cases; in Section 3.6.2, I present a small-scale questionnaire study which was designed to investigate how other people reacted to some of the sentences I included in the corpus. In Section 3.7, I compare the corpus of ungrammatical English presented in this chapter to similar corpora. I suggest, and subsequently dismiss, an alternative methodology to the one I have used. This methodology involves the use of an existing parser and grammar in order to automatically find instances of ungrammaticality. An analysis of the type of ungrammatical language found in the corpus is delayed until Chapter 5, where it is presented in the context of a description of a parser which uses some of the results of the analysis.

The corpus of ungrammatical language itself is included in Appendix A. Sentences from the corpus which are used as examples in this and subsequent chapters have their corpus number listed in brackets beside them.

3.2 Corpus Sources

The corpus of ungrammatical sentences was manually constructed between July 2002 and January 2004. Any ungrammatical sentence which was noticed while reading was added to the corpus, resulting in 923 sentences, 993 errors and approximately 20,000 words. The table in Fig. 3.1 shows the sources of the corpus sentences along with their distribution. The first column lists the type of text where the ungrammatical sentence was encountered, and the second column gives the absolute frequency of this text type within the corpus. Emails, newspapers and academic papers are the most frequent sources of ungrammatical sentences in the corpus. The high frequencies of these text types do not necessarily indicate that these text types contain more ungrammatical sentences than others — they merely reflect my primary reading material during the period of compiling the corpus. Because I did not monitor the number of sentences I read from each text type, the ratio

Sentence Source	**Frequency in Corpus**
Emails (personal and newsletters)	279
Newspapers, Magazines	255
Academic papers, theses, etc	169
Drafts of own writing	74
Websites (including discussion forums)	72
Student Assignments	42
Technical manuals, software documentation	35
Books	28
Lecture/Presentation Handouts	17
Letters	8
Text Messages	7
Other (Signs, Teletext, Album Sleevenotes)	7

Figure 3.1: Where do the ungrammatical sentences come from?

of ungrammatical sentence to grammatical sentence for each text type is not available. Common sense dictates that this ratio will be lower for some text types than for others: the sentences in a book, for example, will be less prone to ungrammatical language than the sentences in a typical informal email since they have undergone a thorough editing process. However, the propensity of a text type to contain ungrammatical sentences cannot be concluded from the data in Fig. 3.1.

3.3 Recognizing an Ungrammatical Sentence

In order to observe instances of ungrammatical language in the written language that one encounters as a part of everyday life, it is necessary to be able to tell whether something is actually ungrammatical. In doing this what one is essentially doing is making a grammaticality judgement, albeit a grammaticality judgement on a naturalistic sentence occurring in its normal discourse context. It is clear that a definition of the term *ungrammatical* is needed in order to be able to make consistent grammaticality judgements. A definition will act as an anchor or a reference point during the corpus compilation process. The difficulty involved in the explanation of the term *ungrammatical* was discussed in Section 2.2.2. The concept is a slippery one: it can be used in an everyday and in a theoretical sense, and coming up with a definition of the term will depend on what kind of linguistic information one believes should fall under the remit of grammar. Thus, a definition such as Def. 2 is unhelpful because it merely shifts the problem onto the grammar.

Definition 2 ***Possible Definition of Ungrammatical***
A sentence is ungrammatical if it cannot be generated by the grammar rules of the language.

To overcome this problem I adopted an operational definition which avoids the term *grammar* completely. The definition has two conditions which must be satisfied for a sentence to be considered ungrammatical. It is given as follows:

Definition 3 ***Definition of Ungrammatical Used To Compile Corpus***
A sentence is ungrammatical if

1. *It contains an error.*
2. *All the individual words in the sentence are well-formed words of English.*

The first condition states that an ungrammatical sentence must contain an error. In the field of error analysis, which is concerned with the analysis of errors made by learners of a second language, a distinction is made between errors which occur as a result of lack of knowledge of the language and those which occur as an oversight (James 1998). The former are known as *errors* and the latter *mistakes*. Throughout this thesis, the word "error" is used to mean either a mistake or a language-learning error. Of course the decision over whether a sentence contains an error *is* a subjective one because it depends on a person's opinion about what constitutes an error. This subjectivity is difficulty to avoid, but a subjective definition which is clear and easy to apply, and which is applied in a consistent manner by one person will produce useful data.

The notion of error is quite broad since there are a number of levels at which language can become ill-formed. I make an assumption that a parser operates at the inter-word level. Thus, the second condition in Definition (2) focuses attention on this level rather than the intra-word level. This means that sentences such as (3.1) will satisfy this condition for ungrammaticality whereas sentences such as (3.2) will not.

(3.1) *Attempts for this seem to have not gotten* ***to*** *far.* (A.371)

(3.2) *It is probably true that all of us use some of it some of the time, whether we want to or not, whether we hear* ***outselves*** *doing it or not.* (Lakoff 1976, p.77)

The kind of spelling error contained in Sentence (3.1) is known as a *context-sensitive* spelling error (see for example Bigert (2004)) and a *real-word* spelling error (see for example Ingels (1996)). One could argue that sentences containing context-sensitive spelling errors are not ungrammatical because the error involves the orthography of the word rather than some syntactic feature such as case or agreement. However, they should be considered ungrammatical because they are a real problem for parsers. Empirical evidence of this is provided in Chapter 6.

A definition of the term *ungrammatical* was used in compiling the corpus in order to ensure that the corpus compiler's notion of the term remained fixed. In Definition 2, the notion of *grammar* is supplanted by the more practical notion of *error*. This definition of *ungrammatical* is intended to be independent of any particular linguistic theory and should be reliably applicable by both linguists and non-linguists. The interpretation of *error* is still, of course, subject to a certain amount of variation. However, in compiling this corpus, the definition was applied by only one person, thus removing inter-subject consistency as a source of concern. I return to the issue of reliability in Section 3.6. For the rest of this thesis, the word *ungrammatical* is used in the sense set out in Definition 2.

3.4 Classifying Ungrammatical Sentences

As well as merely collecting a set of ungrammatical sentences, the sentences were classified according to the type of error(s) they contained. The most convenient way of identifying error types was to classify an error in terms of what could be done to correct it. A similar classification scheme is also used in the field of error analysis (James 1998, chapter 4). Consequently, every time an ungrammatical sentence was added to the corpus, a note was made of what exactly could be done to correct the error, e.g. delete, insert, substitute or move a word. It is this information which is analysed (see Section 5.2).

If an ungrammatical sentence is classified according to how it can be corrected, an obvious extension to the classification process is to *actually* correct the sentence. Accordingly, a second corpus of grammatical sentences was compiled along with the original corpus of ungrammatical sentences. The result is a parallel corpus which is an empirical model of the grammatical/ungrammatical distinction in written English. The first corpus of ungrammatical sentences will be henceforth referred to as *corpus1*, and the second corpus of grammatical sentences as *corpus2*.

In 21% of cases, the ungrammatical sentence could be corrected in more than one way while keeping its meaning relatively constant. An example is the ungrammatical sentence (3.3) which contains an agreement error between the subject noun *journey* and the verb *begin*.

(3.3) *The longest* ***journey begin*** *with one step.* (A.569)

This sentence can be corrected by changing *journey* to *journeys* or by changing *begin* to *begins*.[1]

Another example is the ungrammatical sentence (3.4):

(3.4) *Such databases can* ***be*** *to some extent* ***be*** *improved by reference to non-corpus sources such as the native speaker's knowledge of the language.* (A.407)

This sentence can be corrected by deleting the first occurrence of *be* yielding the grammatical (3.5):

(3.5) *Such databases can to some extent* ***be*** *improved by reference to non-corpus sources such as the native speaker's knowledge of the language.*

or by deleting the second occurrence yielding the equally grammatical and synonymous (3.6):

[1]Obviously, there is some change in meaning in settling on either a singular or plural subject, but it is negligible given the context, which is the opening sentence of an email asking recipients to support a petition which aims to promote the Irish language.

(3.6) *Such databases can **be** to some extent improved by reference to non-corpus sources such as the native speaker's knowledge of the language.*

When there is a second grammatical correction for a given ungrammatical sentence in a given context, both corrected versions are added to *corpus2*.[2]

On its own, a set of ungrammatical sentences provides evidence of the kind of structures exhibited by ungrammatical sentences. Correcting the ungrammatical sentences in *corpus1* to yield *corpus2* explicitly relates the ungrammatical to the grammatical and, as such, provides evidence of the systematic ways in which grammatical sentences can become ungrammatical. *Corpus2* serves a particularly useful function in the parser evaluation method described in Chapter 6, in which parses of *corpus1* sentences are measured against parses of their *corpus2* counterparts.

3.5 The Importance of Context

There are essentially two tasks involved in compiling the parallel corpus of ungrammatical English. The first task is the identification of sentences to be included in *corpus1*. This task involves a grammaticality judgement. The second task is the correction of the ungrammatical sentence, yielding one or two grammatical sentences which are added to *corpus2*. Both of these tasks are made easier by the fact that the sentences are encountered in context, during the course of reading.

[2]For a very small number of ungrammatical sentences, I was able to propose a third correction which had the same meaning as the first two corrections. However since these cases were rare, I imposed a practical limit of two corrections per sentence.

3.5.1 Importance of Context in Ungrammatical Sentence Identification

In Section 2.2.1, I discussed the interaction between the context in which a sentence occurs and the reliability of grammaticality judgements made upon that sentence, concluding that a judgement made on a sentence in context is far more valuable than a judgement made on the same sentence out of context. This is because the context can provide information which disambiguates, making the structure of the sentence to be judged more transparent. If the context is not there, it is imagined, introducing an unnecessary variable into the judgement process. With this in mind, the ungrammatical sentences contained in the corpus have been observed (and judged) in their natural context as part of the basic linguistic process of understanding written language, i.e reading.

Another advantage of collecting instances of ungrammaticality in context is that it allows some sentences to be included in the corpus which might be excluded if they were encountered out of context. This kind of sentence needs a context to make it *ungrammatical* and is, therefore, the opposite of the elliptical *He me* case discussed in Section 2.2.1 which needed a context to be considered *grammatical*. An example is Sentence (3.7):

(3.7) *The error position is indicated by the edges and nodes build upon the words.* (A.496)

Out of context, this sentence is grammatical. In the context in which it occurs, which is a discussion of an annotated error corpus (Bredenkamp *et al.* 2000), it is ungrammatical and should be corrected as follows:

(3.8) *The error position is indicated by the edges and nodes **built** upon the words.*

Another example is (3.9) which occurs in an email:

(3.9) *We can order then directly from the web.* (A.243)

Taken out of context, it is not difficult to imagine a context in which the verb *order* is occurring without an object and is being modified by the adverb *then*. Little room is left for imagination when the sentence is viewed in its discourse context which is an email discussion about purchasing gifts. In this context it is clear that the word *then* is really a mistyped form of the object pronoun *them*. I return to this particular sentence and the issue of context in Section 3.6.2.

In the field of error analysis, the kind of error which occurs in a sentence whose structure is superficially well-formed is known as a *covert* error (James 1998).

3.5.2 Importance of Context in Ungrammatical Sentence Correction

In order to correct an ungrammatical sentence with confidence it is necessary to know what the person responsible for the sentence was trying to say, i.e. it is necessary to know what the ungrammatical sentence means. The fact that the context was always available meant that most sentences could be corrected without significant risk of ambiguity. Consider Sentence (3.10) which occurred as a headline on a Irish newspaper website and which contains a number agreement error between the plural subject noun *apartments* and the singular verbs *goes*.

(3.10) ***Apartments*** *seized by CAB* **goes** *under the hammer.* (A.843)

This sentence could be sensibly corrected by making the subject noun singular (remember, it occurs as a headline, and newspaper headlines often omit a determiner before a singular noun), or by making the verb plural. The two corrections carry distinct meanings, but which one is the true correction? This question can be answered by reading the article headed by this sentence. From this, it is clear that more than one apartment was seized, and hence, that the error resides in the verb.

I return to the issue of context and sentence correction, in Section 3.6.2.

3.6 Reliability of Corpus Data

In this section I present examples of unclear cases which were excluded from *corpus1* and *corpus2* in order to ensure that the data in the corpora is reliable. The reliability of the correction phase is tested in a questionnaire study which presents some of the *corpus1* sentences with and without a small amount of context and asks people to supply corrections if they think that the sentences contain an error.

3.6.1 The Unclear Cases

Sometimes, the tasks of identifying a sentence as ungrammatical and correcting the sentence cannot be carried out with confidence. This creates the risk of a lack of reliability and in such cases, I err on the side of caution and do not include such sentences in the corpus (although they are recorded elsewhere).

As regards the indentification task, it is not always clear, even given the context in which a sentence occurs and the operational definition of the term *ungrammatical* (see Definition 2 in Section 3.3), whether or not a sentence should be included in *corpus1*. An example is the (in my opinion) stylistically awkward (3.11), which appeared in a newspaper article about the proposed decentralisation of the Irish civil service:

(3.11) *This whole nonsense is another example* ***of where*** *local politics completely override the national interest.*

While I might prefer (3.11) to be rewritten as (3.12), it seems a step too far to describe what is amiss in terms of a structural error.

(3.12) *This whole nonsense is another example of local politics completely overriding the national interest.*

Stylistic clumsiness is not the only type of unclear case encountered. Another unclear case is the following which appeared in an email discussion between friends debating restaurant choices:

(3.13) *I hear* ***that that*** *Antonio's is a nice place but I'm sure Peter wants to go there to remind him of Erasmus!!*

In this example it is quite probable that the second *that* is a duplication error and that the sentence is ungrammatical. However, this does not represent a clear case because there is a second contextually plausible reading in which the second *that* is a demonstrative determiner and *that Antonio's* is a noun phrase.

A third type of unclear case is sentence (3.14) which occurred at the beginning of an email written by a non-native English speaker, who is introducing herself to a group of people:

(3.14) *I am Coco* ***who is doing*** *a PhD in computer science.*

The problem with (3.14) is that it seems like an unnatural way of expressing what a native speaker would express using a sentence such as (3.15):

(3.15) *I am Coco and I am doing a PhD in computer science.*

The reason why Sentence (3.14) is not considered to be ungrammatical and is not included in *corpus1* is similar to the reason why Sentence (3.11) is not: although both cases seem awkward to me, (3.11) seems unwieldy and (3.14) seems infelicitious, I did not think my intuition about them was strong enough for them to be considered erroneous. Thus, *corpus1* contains only those sentences which are considered by me to be undoubtedly ungrammatical.

The task of correcting an ungrammatical sentence also yields unclear cases. If an ungrammatical sentence is encountered whose meaning is not clear, a note is made of the sentence but it is *not* added to *corpus1*. An example is sentence (3.16) which occurred in a newspaper article discussing cuts in funding for the arts in Ireland.

(3.16) *There is a huge curtailment in the performing arts which has **affected** companies **from** doing productions all over Ireland because they are told to keep costs down.*

Even given the context in which the above sentence occurs there still remains some ambiguity: is the sentence suggesting that companies currently doing productions in Ireland have been adversely affected by the cuts as in (3.17), or does it have the stronger reading that the companies have actually been prevented from doing productions, as in (3.18)? It is not clear from reading the whole article what the ungrammatical sentence means, and so it cannot be reliably corrected.

(3.17) *There is a huge curtailment in the performing arts which has **affected** companies doing productions all over Ireland because they are told to keep costs down.*

(3.18) *There is a huge curtailment in the performing arts which has **pre-***

vented *companies* ***from*** *doing productions all over Ireland because they are told to keep costs down.*

The error in Sentence (3.16) created an ambiguity which made it difficult to correct. Some ungrammatical sentences were excluded from the corpus, not because the ungrammaticality in the sentence resulted in too many interpretions, but because the ungrammaticality resulted in no interpretation. Sentence (3.19) is an example: on encountering this sentence I had no difficulty in classifying it as ungrammatical but I did find it difficult to interpret. The sentence occurs in a message which pops up on a laptop computer screen when the laptop is either connected or disconnected from the main power supply.

(3.19) *LaptopsRUs Hotkey Utility show the indicators on your display and save brightness adjustments each power supplying conditions.*

Ungrammatical and unintelligible sentences such as (3.19) should be excluded from the corpus because they do not need to be described by parser's grammar. If a native speaker struggles to understand a sentence, it is reasonable for a parser to be unable to parse such a sentence.

To summarize, the unclear cases can be summarized as follows:

1. Unclear cases arising from identification task (grammaticality judgement)
   (a) Clumsy writing style (3.11)
   (b) Ambiguous between an interpretation which contains an error and one which does not (3.13)
   (c) Infelicitous choice of words or construction (3.14)
2. Unclear cases arising from correction task

(a) Ungrammaticality leads to two or more interpretations (3.16)

(b) Ungrammaticality leads to unintelligibility (3.19)

Since both types of unclear cases, those resulting from identifying a sentence as ungrammatical and those resulting from correcting an ungrammatical sentence, are excluded from *corpus1* (and hence also from *corpus2*), the data which I have collected constitute only those cases where my intuition is completely clear. This means that the data in the corpus is more likely to reflect truly ungrammatical cases and is therefore less vulnerable to criticisms of subjectivity. This claim is explicitly tested in a small pilot experiment described in Section 3.6.2.

3.6.2 Questionaire Study

In Section 3.6.1, I argued that the data in *corpus1* and *corpus2* should be reliable because it is data which clearly fell within the confines of my definition of *ungrammatical* (see Def 2 in Section 3.3) and it can be corrected with confidence. In this section, I describe a small experiment which tests how reliable this data is, by comparing the corrections I have made with corrections made on the same sentences by other English speakers. It is a small-scale experiment because I do not use a large pool of subjects, but if the subjects that I do use agree for the most part with my corrections, I think this suggests that the data in *corpus1* and *corpus2* is reliable.

It is impossible to mimic exactly in an experiment the conditions in which I encountered the sentences that I have included in *corpus1*, since I came across them in an opportunistic manner while reading. For practical reasons, I cannot include all the context for every sentence. Instead I include one or two sentences before and/or after the sentence in question. Also, not all the sentences will be appropriate: those extracted from technical and academic articles (for example, Sentence (3.7)) are not appropriate because I cannot expect anyone without a background in computer science or linguistics to understand them; many extracted from short personal emails from friends are also unsuitable because, quite apart from the ethical issue of presenting such sentences in context, a knowledge of the people involved is required in order to guarantee understanding. Other corpus sentences which cannot be considered for inclusion in the experiment, are those whose context is no longer available, e.g. those taken from internet pages which no longer exist.[3]

A second hypothesis which I would like to test in this experiment, is whether context really does play a role in gathering the data for *corpus1* and *corpus2*. My discussions of Sections 2.2.1 and Section 3.5 make it clear that I think that it is an important factor, but I will test this explicitly by presenting subjects with some sentences in isolation and some in context — albeit not the full context in which the same sentences were encountered by me — and seeing if there is a difference in the extent to which the corrections proposed agree with my *corpus2* sentences.

In the following sections, I describe the experimental subjects, materials and task, present the results of the experiment, and discuss these results.

Procedure

The experimental subjects were twenty fluent English speakers (nineteen native speaking and one with French as a native language) aged between twenty three and sixty three years of age. The subjects were arbitrarily

[3]Unfortunately, when I was adding sentences to *corpus1*, I only made a note of where and when I found the sentence. I did not have the foresight to copy some of the sentences surrounding the sentence of interest.

Example from Part One (sentence in context): Beckham and company want to show their support for Ferdinand, but realise they would compromise their England careers if they carry through with their threat. And for now the FA are playing *down the players action.* Director of communications, Paul Barber said: "They obviously wanted to express their opinions and views on what has happened. We were very happy to listen. They also listened to our explanation of what has happened."
Example from Part Two (sentence out of context): HP America *is ending staff over* to work on the programmes rather than basing them with a US university and already has six US researchers working full time in Ireland.

Figure 3.2: Example Items from Parts One & Two of Questionnaire

divided into two groups of ten, *SubjA* and *SubjB*.

The materials consisted of twelve sentences randomly extracted from the subset of *corpus1* containing sentences deemed to be appropriate for the experiment. The twelve sentences were randomly divided into two groups of six, *SenA* and *SenB*. *SubjA* were presented with the sentences from *SenA* in context (a few sentences before and after the sentence of interest) and with the sentences from *SenB* out of context. *SubjB* saw the sentences from *SenB* in context and those from *SenA* out of context. The sentences in context were presented first as Part One of the experiment. The sentences out of context were presented as Part Two of the experiment. The order in which the sentences were presented was randomized for each subject.

Subjects were instructed to read either the text (Part One) or the sentence (Part Two) in full and then to correct the error, which was indicated in each case by italicizing four words in the region of the error. They were told to provide more than one correction if more than one occurred to them. A space was provided after each test item to allow the subjects to write down their corrections. If a subject could not correct an error, (s)he was instructed to write down the reason why not. (S)he was told that one of the reasons why (s)he might not be able to correct the error is if (s)he did not think an error had occurred, and another reason might be that s(he) could not make sense of the sentence or the text in which the error occurred.

Fig. 3.2 shows one test item from each part of the experiment for a subject from the *SubjA* group. The example from Part One is Sentence (A.323) from *corpus1*, and the example from Part Two is Sentence (A.236). The two versions of the full questionnaire are given in Appendix B.

Results

The results of the questionnaire study are presented in Fig. 3.3. An aim of the study was to ascertain to what extent others correct sentences from *corpus1* in the same way that I have. Thus, for each sentence/respondent pair, a precision and recall score was calculated. The precision score represented the proportion of responses produced by this respondent for this particular sentence which were the same as the response(s) produced by me for the same sentence. The recall score represents the proportion of corrections produced by me for some sentence which were reproduced by the respondent for the same sentence. Fig 3.3 shows the precision and recall results for the entire set of 240 respondent/sentence pairs. If a respondent does not supply a correction for a sentence in the questionnaire, and instead supplies something like “Sentence is ok” or “Sentence makes no sense”, this counts as a response – albeit one that differs from mine since all the sentences in the questionnaire were ones which I considered to be ungrammatical *and* correctable.

Another aim of the experiment was to investigate the role of context in

	Precision	**Recall**
Overall	78.6%	72.5%
In Context	83.8%	74.3%
Out of Context	73.9%	70.6%

Figure 3.3: Questionnaire Results

the correction process. Figure 3.3 study shows that there is a difference in the precision and recall figures for the sentences which were presented in context and those which were presented out of context. The difference in context is statistically significant ($p = 0.043$) for precision but not for recall. Statistical significance was determined using the Mann-Whitney U test (see Bryman and Cramer (2001, p.133-134)).

Discussion

I will begin the discussion by analysing the cases where the questionnaire respondents did not agree with my corrections. These cases accounted for the 21.4% precision error rate. They can be categorized as follows:

1. **No correction supplied because a sentence did not seem ungrammatical**: Most of the cases where a respondent replied that there was nothing wrong with the sentence in question seemed to be cases where the respondent simply did not notice the error. This is because these cases only occurred with uncontroversially ill-formed sentences, containing an error so mundane as to be easily overlooked and which was corrected in the same way by all other informants, e.g. Sentence (3.20):

(3.20)*Please tell* ***you*** *friends.* (A.505)[4]

One respondent judged Sentence (3.21) (presented in context) to be grammatical but it is not clear whether to interpret this as a case where s(he) did not notice the missing apostrophe, or as a case where s(he) was not aware of the rule. The second case is quite possible, since one could argue that the rules for apostrophes come under the remit of "school grammar" (see Section 2.2.2):

(3.21)And for now the FA are playing *down the* ***players*** *action.* (A.323)

One of the twelve sentences in the questionnaire contained a covert error and was judged to be grammatical by some subjects when presented out of context, but by none when presented in context. This sentence was first discussed in Section 3.5.1 (Sentence (3.9)) but is repeated here for convenience as Sentence (3.22):

(3.22)We can *order* ***then*** *directly from* the web. (A.243)

2. **No correction supplied because sentence did not make sense**: There was one sentence (example (3.23) below) in the questionnaire which three respondents did not correct because it did not make sense to them. Two of these saw the sentence out of context, and one saw the sentence in context.

    (3.23)We ***might found*** *out what* was going on every six months, during which time researchers could have followed a path we didn't want them to take, wasting both our time and our money. (A.399)

[4]In the questionnaire sentences used as examples in this section, I will retain the italics which were used in the questionnaire.

Some respondents supplied a correction (sixteen out of the seventeen corrections supplied agreed with mine) for this sentence but remarked that it was difficult to interpret. This sentence is undoubtedly awkward, but it is possible that this interpretation difficulty, which appears to be related to sentence tense, would be reduced if more of the article in which the sentence appeared was presented to the subjects. Unfortunately, this was not possible for practical reasons. It would be interesting to see whether the "no correction due to no sense" response would have been produced if it had not been given as a suggested response in the questionnaire instructions.

3. **Correction supplied by respondent but does not appear in *corpus2***: Some corrections were suggested by the questionnaire respondents which did not reflect the corrections I made when compiling *corpus2*. Of these, some were clear cases of a context effect. Sentence (3.21) is an example. Many of the subjects who saw this sentence in isolation suggested two corrections: one in which the apostrophe comes before the plural "s" marker and one in which it comes after. This is because, without a context, it is not clear whether *players* is singular or plural. It is plural in the actual context which means all corrections involving a singular reading did not match the correction I produced for the sentence.

   Another phenomena which accounted for differences between my corrections and those of the respondents were instances of what I would consider to be overcorrection. This happened when a respondent corrected an error in a sentence in the same way that I had corrected it but also made another slight change to the italicized area of the sentence which contained the error. Consider, for example, Sentence (3.24), which I corrected to produce Sentence (3.25):

(3.24)We just find that it is easier to find common ground with *Irish academics* ***that*** *American.* (A.802)

(3.25)We just find that it is easier to find common ground with *Irish academics* ***than*** *American.*

(3.26)We just find that it is easier to find common ground with *Irish academics* ***than*** *American ones.*

(3.27)We just find that it is easier to find common ground with *Irish* ***than*** *American academics.*

Some respondents produced Sentences (3.26) and/or (3.27) instead of or along with Sentence (3.25). I call this phenomenon overcorrection because I think it involves a tendency on the part of respondents to paraphrase something so that it is in keeping with their particular notions of writing style, as opposed to minimally correcting it so that it no longer contains an error. A possible reason why overcorrection occurred was because the exact location of the error could not be indicated without giving away a correction. The italicizing of four words in the region of the error could have led people to assume that all four words were involved in the error.[5]

[5]The three respondents who were over fifty-five seemed particularly fond of overcorrection, not just in the italicized area which they are asked to correct, but also elsewhere in the sentence or in the preceding or following sentences for part one of the questionnaire. All three attempted to paraphrase sentences ending in a preposition. This reinforces the point made in Section 2.2.2 that our ideas about language correctness are linked to when we were educated and that older people are likely to have more rigid views about writing style.

There were only two instances of respondents producing plausible corrections which I had overlooked. The first example is the correction of Sentence (3.28) to produce Sentence (3.29). This correction was suggested by two respondents, with the other eighteen respondents producing the same correction as I did, shown in Sentence (3.30):

(3.28) HP America *is* ***ending*** *staff over* to work on the programmes rather than basing them with a US university and already has six US researchers working full time in Ireland. (A.471)

(3.29) HP America *is* ***handing*** *staff over* to work on the programmes rather than basing them with a US university and already has six US researchers working full time in Ireland.

(3.30) HP America *is* ***sending*** *staff over* to work on the programmes rather than basing them with a US university and already has six US researchers working full time in Ireland.

Although the person who produced this sentence is probably more likely to have meant *sending* than *handing*, Sentence (3.29) still represents a contextually valid correction, and it could be argued that it should be included in *corpus2*. The second example is Sentence (3.32) which is a correction produced by one respondent for Sentence (3.31). Sentence (3.33) and (3.34) are the corrections produced by me and the other respondents.

(3.31) And the average income *for* ***a players*** *in* the top half of the Premiership now runs to over EUR50,000 weekly. (A.436)

(3.32) And the average income *for* ***all players*** *in* the top half of the Premiership now runs to over EUR50,000 weekly.

(3.33) And the average income *for* ***a player*** *in* the top half of the Premiership now runs to over EUR50,000 weekly.

(3.34) And the average income *for* ***players*** *in* the top half of the Premiership now runs to over EUR50,000 weekly.

Sentences (3.32) and (3.29) are the only cases which highlight a weakness in my corpus compilation method, i.e. that there will be valid corrections of *corpus1* sentences which do not find their way into *corpus2*. However, the relative scarcity of such corrections in the questionnaire responses is encouraging.

The 27.5% error in recall represents the cases where a respondent failed to reproduce a correction for a sentence that I had produced for the same sentence when compiling the parallel *corpus1/corpus2*. How can this figure be explained? Unlike the precision results, the recall results do not display a significant context effect. That is, the failure on the part of a respondent to reproduce a *corpus2* correction cannot be explained or partly explained by the differences between the two parts of the questionnaire. I think the recall result is, to some extent, due to the attitude of the respondent. Of the twelve sentences in the questionnaire, eight of them have one correction in *corpus2* and four of them have two corrections. However, eight of the twenty respondents produced only one correction for each of these four sentences. I talked to one respondent who had supplied only one correction per questionnaire sentence after he had filled in the questionnaire: for Sentence (3.31), he had supplied Sentence (3.33) as a correction. When asked whether he thought Sentence (3.34) represented another valid correction, he replied that it did. When I asked him why he had not put it down on the questionnaire sheet, he replied that he thought one correction was enough. The instructions ask the respondents to put down more than one correction if more than one correction occurs to them, but this was obviously not emphasized enough, and is a flaw in the experimental design.

Of the four sentences with two corrections in *corpus2*, there was one sntence where none of its corrections were reproduced, i.e. one of the corrections I produced for this sentence was not reproduced by *any* of the questionnaire respondents. The correction is Sentence (3.36) which represented, in my opinion, a contextually suitable correction of Sentence (3.35):

(3.35) When I *heard was the crashing* of glass, I knew there was something coming in again. (A.462)

(3.36) When I *heard* ***what*** *was the crashing* of glass, I knew there was something coming in again.

I stand by the correction represented by Sentence (3.36), and the only explanation that I can give for its non-appearance in the questionnaire responses, is that the context that was supplied (for those who saw it in context) was not long enough to prime this correction. This is a matter for further empirical investigation.

This questionnaire study was carried out for two reasons: the first reason was to investigate to what extent the corrections I made to *corpus1* sentences would agree with the corrections made by others on a random subset of these sentences, and the second was to test the hypothesis that context plays a role during the judgement[6] and correction process. This context hypothesis was validated by the questionnaire because the difference in precision results for sentences presented in context and those presented out of context is signifi-

[6]Although this questionnaire study primarily involves error correction, the fact that the respondents had the option to conclude that a sentence did not need correcting meant that implicit grammaticality judgements (see Section 2.2.4) were also being carried out.

cant. This counts as empirical evidence of the views expressed in Section 2.2.1 and Section 3.5.

There was not 100% agreement between my corrections and those produced by the questionnaire respondents, nor did I expect there to be. I have discussed the disagreements which occurred in some detail. A certain type of disagreement in precision represents a flaw in my corpus compilation methodology, that is, some valid corrections will be overlooked by me, with the result that *corpus2* might be considered incomplete. This could perhaps have been avoided if the corrections of candidate *corpus1* sentences were discussed by a group of people. However, this would be an impractical methodology given the fact that the ungrammatical sentences are encountered in one person's personal reading material. This leads to another potential explanation of some the disagreements between my corrections and those of the respondents: it is impossible to mimic exactly what I did between January 2002 and July 2004, in an experimental situation. A plausible explanation for the fact that some of the respondents only produced one correction where I produced two, is that they were just not as motivated to find a second one as I was. However, I think that the (in context) precision/recall figures of 83.8% and 74.3% are respectable enough to conclude that *corpus1* and *corpus2* is a valuable linguistic resource reflecting the grammatical/ungrammatical distinction in everyday written English.

3.7 Comparison with Other Corpora and Methodologies

In Section 3.7.1, I compare *corpus1* and *corpus2* to other error corpora, and in Section 3.7.2, I discuss an automatic procedure for finding errors, and explain why this procedure was not applied in compiling my error corpus.

3.7.1 Other Corpora

Many attempts to compile a corpus of ungrammatical language have taken place in the context of studies into second language learning. The International Corpus of Learner English (Granger 1993), for example, is a two million word corpus of written English produced by learners of English from fourteen different mother tongue backgrounds. Emi *et al.* (2004) describe the SST corpus: a one million word corpus of transcribed spoken English produced by native Japanese speakers during oral interviews to test spoken fluency in English. A subset of this corpus has been manually annotated with information on the type of error and the correction. The analysis of the errors in learner corpora such as these allow linguists to gain insight into the process of second language learning, as well as providing useful data for computer-aided language learning applications. Although some of the sentences in *corpus1* do seem to have been produced by non-native speakers of English, it was not my aim to focus on this kind of error alone. Any ungrammatical English sentence is added to *corpus1*, regardless of whether it has been produced by someone who acquired English as a mother tongue or by someone who learnt it later in life. Often not enough information was available to make this distinction.

A long tradition within linguistics has been the analysis of the errors made in speech (see, for example, Fromkin (1980)). The aim of the analysis has been to gain insight into the processes and structures underlying the production of spoken utterances. Although *corpus1* was created in order to gain insight into the kind of ungrammatical sentences a parser of written English might be expected to encounter, it is available as a resource for linguists interested in studying the processes involved in the production of *written* utterances.

Attempts to compile error corpora in the context of robust language processing projects are described by Atwell and Elliott (1987), Ingels (1996) and Becker *et al.* (1999). Atwell and Elliott (1987) collected context-sensitive spelling errors in English from four sources: a collection of word-processed technical documents, a copy of the Gill Corpus from the Oxford Text Archive, the *Manual of Information* for the Lancaster-Oslo-Bergen Corpus which lists all the errors found in the original texts (Johansson *et al.* 1978) and business-style documents which were typed up by work experience trainees from the Information Technology Education Centre in Lancaster. The position of the erroneous word within each sentence is flagged. At the time of writing, the corpus contained 500 errors and 13,500 words. Ingels (1996) describes two corpora which represent the interactions between experimental subjects and a human operator who is simulating a computer interface to a database system: the database system for the first corpus is concerned with car modelling, and the database system for the second is concerned with travel bookings. The corpora are analysed and approximately 200 errors are found. Becker *et al.* (1999) describe the creation of a German error corpus to be used in the evaluation of grammar checking software. The errors in the corpus are annotated according to an error typology which was created using traditional grammar books. The errors were taken from emails sent to an internet discussion group. A similarity between these three approaches to error collection and my approach is that none of the three are concerned with the reason why a particular error has occurred. They are simply concerned with collecting naturalistic errors in written language, with a view to using this data in the context of robust parsing.

A difference between my error corpus and the corpora mentioned above is the manner in which the errors have been collected. In most other error corpora, the texts are chosen in advance and then scanned for ungrammatical language. The error corpus described in this chapter has been collected in a much more opportunistic fashion as a by-product of reading which would have been carried out anyway. Sampson (2001c) adopts a similar methodology when he collects examples of central embedding in his reading material over the course of several months. An obvious advantage of this approach is that it is less tedious to collect errors in texts one has an interest in reading than to collect errors in arbitrary text collections. However this is also a disadvantage: if one is absorbed in the content of a piece of writing, one is perhaps less likely to notice instances of ungrammaticality, particularly if the instance is quite mundance and trivial to correct. While it is obvious that I could not have noticed all the errors in my reading material during the corpus compilation period, this does not invalidate the errors I did find. Also, if one explicitly scans for errors, one is tempted to actually ignore content which means, firstly, that attention is more likely to wander, and, secondly, that covert errors such as those in (3.7) and (3.9) are likely to be missed. A second advantage of the opportunistic reading-based approach to error collection concerns the issue of context: broader context, in the form of background knowledge, is more likely to be available for texts which one has some interest in reading, as opposed to texts which are only being read in order to find errors. A third advantage of this approach is that, if the corpus compiler's source of reading material is broad, then the corpus will consist of errors from a wide variety of written text (see Fig. 3.1).

3.7.2 An Alternative Methodology

The task of looking for errors is a labour intensive one, whether it involves looking out for errors when reading, as I did, or explicitly scanning certain texts for them as most other approaches to the task have involved. An alternative approach to this manual one is to use an existing parser to find instances of ungrammaticality. This automatic error collection method would proceed as follows:

1. Compile a very large collection of machine-readable texts from a wide variety of sources.
2. Find a parser with a wide-coverage grammar and lexicon *and* which is capable of distinguishing between grammatical and ungrammatical.
3. Input the text to the parser. All sentences for which no parse is found are deemed ungrammatical.

An attempt was made to carry out this automatic procedure. This attempt was abandoned, mainly because step two of the procedure proved infeasible in practice. No freely available parser/grammar could be found which could be used "off-the-shelf" to parse a large body of raw text *and* make a grammatical/ungrammatical distinction. A traditional parser such as the LKB used in conjunction with a wide-coverage grammar such as the ERG was not suitable because, although it distinguishes between grammatical and ungrammatical sentences, words have to be added to its lexicon in order for it to recognize grammatical sentences (see also Section 7.6.1). A probabilistic parser such as the Penn-Treebank trained parser described in Charniak (2000) was not suitable because although it produces parses for grammatical sentences, it also produces parses for the ungrammatical ones.[7] The same can be said for a partial parser such as Fidditch (Hindle 1994). The effort required to get any parser and grammar to recognize the well-formed sentences and reject the ill-formed sentences was deemed too much for this approach to be continued.

A second problem with this automatic approach to finding examples of ungrammatical language is that a computer parser cannot make use of contextual information in the way that a human parser can, with the result that a computer parser will find analyses for sentences which seem grammatical out of context but which *are* ungrammatical in context, i.e. sentences containing covert errors. Two examples are Sentences (3.7) and (3.9) from Section 3.5.1 and a third is Sentence (3.37) taken from an email in which the writer is complaining about the behaviour of another:

(3.37) *Anyhow I was really angry, she crossed one* ***to*** *many lines there.* (A.354)

In (3.37) *to* has been typed instead of *too*. It is conceivable that such a sentence, which is obviously erroneous to the human parser, would not be rejected as ungrammatical by a computer parser. Such sentences are a valid source of ungrammatical evidence and *corpus1* is richer for their inclusion.

3.8 Conclusion

In this chapter, the creation of a corpus of ungrammatical sentences was discussed. Sentences were added to the corpus on the basis of a grammaticality

[7] It would be worthwhile to investigate to what extent a broad-coverage probabilistic parser *can* distinguish between the grammatical and the ungrammatical even though it does not explicitly encode such a distinction. This could be tested by seeing if there is a significant difference between the frequency values obtained by such a parser for ungrammatical sentences and the frequency values obtained for grammatical ones.

judgement rooted in a theory-independent definition of *ungrammatical* which makes use of the idea of an error which is detectable at an inter-word level. The sentences come from a variety of written sources and were encountered in context. An interesting aspect of the error corpus is its parallel nature. A parallel corpus usually refers to a collection of aligned corpora that are translations of one other: each aligned unit in one corpus expresses the same meaning as the corresponding units in the other corpora but in a different language. This kind of parallel corpus has applications in machine translation, word sense disambiguation and the construction of bi-lingual lexicons. *Corpus1* and *corpus2* represent a different kind of parallel corpus: a sentence in *corpus1* expresses the same meaning as a sentence in *corpus2*, but the difference between the sentences is not one of language but of grammaticality. A topic which is outside the scope of this thesis, but would nevertheless be worth exploring, is to what extent the techniques which are used to exploit traditional parallel corpora could be applied to *corpus1* and *corpus2*, or larger versions of them (see Chapter 8 for a further discussion of this point).

On its own, *corpus1* serves as evidence for the second component of the kind of grammar defined in Section 2.4.2, and as a set of test sentences for any parser which aims to parse ungrammatical language. In Section 5.2, the correction operations which were applied to the ungrammatical sentences in *corpus1* in order to make them grammatical will be analysed. The results of this analysis give some indication of the kind of ungrammaticality one might expect to find in modern written English. The combination of *corpus1* and *corpus2* allow the relationship between the grammatical and the ungrammatical in English to be explicitly analysed; the results of such analysis are useful in the development (see Chapter 5) and evaluation (see Chapter 6) of robust parsing or grammar checking techniques.

Chapter 4

Approaches to Parsing Ungrammatical Language

4.1 Introduction

The previous two chapters were concerned with the problem of identifying ungrammatical sentences. The focus in this chapter moves to the parsing of these sentences. The way ungrammatical sentences are handled by a computer parser will depend on what the parses produced by the parser are to be used for. If the parser forms part of a language tutor to aid second language learners, it will need to be able to detect errors and to provide feedback on these errors. A parser used as a grammar checker for a word processing program will need to handle ungrammatical language in a similar way. A parser which is used in a machine translation or natural language understanding system will need to produce parses for a sentence from which the sentence's meaning can be derived. When an ungrammatical sentence is given to such a parser, it should produce a parse or parses for the ungrammatical sentence which reflect(s) possible meanings of the sentence but it does not have to go a step further and provide feedback on the nature of the ungrammaticality. Such a parser should perform its task for ungrammatical language as well as it does for grammatical language.

In this chapter, I review the various approaches to the parsing of ungrammatical language, whether this parsing takes place in the context of explicit error detection or grammar checking, or in the context of natural language understanding. Although this thesis is concerned specifically with ungrammatical language, many of the parsers I describe have the general aim of being robust not only to errors but also to grammatical language which the parser's grammar for some reason does not describe. The term *extragrammatical* is generally used to describe ill-formed *or* well-formed language which falls outside the scope of the grammar used by the parsing system (see for example Carbonell and Hayes (1983)). The term *robust parsing* refers to the parsing of extragrammatical input, and the parsing of ungrammatical input is just a subset of the robust parsing problem. This chapter will include examples from the general robust parsing literature, but the emphasis throughout is on ungrammatical input (see Menzel (1995) and Ballim *et al.* (1999) for reviews of robust parsing). It should also be noted that some of the robust parsing techniques I describe in this chapter are specifically tailored towards *spoken* language. While I am interested in ungrammaticality in *written* language and although the scope of this dissertation is limited to text, I nevertheless consider it useful not to exclude robustness in spoken NLP from the general discussion.

I begin this review chapter by discussing in Section 4.2 three basic approaches to parsing, the traditional approach (Section 4.2.1), and two more recent and more robust approaches, probabilistic parsing (Section 4.2.2) and shallow or partial parsing (Section 4.2.3). The aim of this section is to provide the reader with a broad overview of general approaches to parsing and the attitude of these parsing types to the problem of ungrammatical language. In Section 4.3, I discuss the various approaches to handling ungrammatical language. Many of these approaches are based on traditional parsers, but some use probabilistic techniques and some have partial parsing as their foundation. These approaches are sub-divided into *constraint relaxation* approaches (Section 4.3.1), *error anticipation* approaches (Section 4.3.2), *minimum edit distance* approaches (Section 4.3.3) and *parse fitting* approaches (Section 4.3.4). The chapter ends with a summary in Section 4.4.

4.2 Three Parsing Paradigms

In this section, I provide a general discussion of three parsing paradigms, traditional parsing, probabilistic parsing and partial parsing. For each paradigm, its parsing philosophy and its attitude to ungrammatical language is described. The partial parsing section is a little longer than the other two because within this paradigm there are distinct groups of parsers. Ballim *et al.* (1999, p.11) remark that the most successful parser will probably be a hybrid one which combines techniques from all paradigms.

4.2.1 Traditional Parsing

To re-iterate a point made in Section 2.1, a grammar of a language is a theory of the structural conventions of the language. This theory is usually expressed in the form of rules which specify how words in the language can be pieced together in order to form sentences. A traditional parser operates by taking a sequence of words from the language and seeing if the rules of the grammar can be applied to generate this string. If they can, then the string is deemed to be grammatical and the application of the rules by the parser represents the parse or the structure of the sentence. This structure can then be returned. If the parser cannot apply the rules of the grammar in such a way that the string is generated, then the string is deemed ungrammatical and no more information is returned.

Traditional parsing is influenced by the generative grammar tradition within linguistics of distinguishing between the grammatical sentences of a language and ungrammatical word sequences which are not part of the language (see Section 2.2). For some computational linguists, the primary aim of a parser is not to analyse the structure of a sequence of words but to make a decision regarding the grammaticality of the word sequence and whether or not it is part of the language: the analysis is dependent on the recognition. Erbach (1997, p.7), for example, defines parsing as "a process of deduction, i.e. proving theorems about signs of a natural language." Within this view of parsing, a parser which returns an analysis for an ungrammatical sentence is not doing its job properly: its grammar is guilty of *overgeneration*. Not only will an overgenerating grammar accept ungrammatical sentences, it will also tend to produce spurious analyses for the grammatical ones. Probabilistic techniques (see Sections 2.3.2 and 4.2.2) or underspecified representational methods are then needed to disambiguate between the parses. An example of a current, widely used parser which adopts traditional parsing assumptions is the parser/generator LKB (Copestake 2002). In Chapter 7, I describe a change I made to the LKB parser which allows sentences containing agreement errors to be parsed.

4.2.2 Probabilistic Parsing

A parser is probabilistic if its grammar is equipped with information which says something about the relative frequency of occurrence of the linguistic structures it describes. The reason why probabilistic grammars have been adopted in parsing is that they allow the various possible analyses for a sentence to be ranked according to how likely they are to occur, and this means that the parser can return one parse rather than several. Charniak (1997) emphasizes that researchers working within the probabilistic parsing paradigm assume that it is the job of the parser to perform disambiguation between possible parses, rather than some other module within an NLP system. The simplest type of probabilistic parser has a probabilistic context-free grammar (PCFG), a context-free grammar in which each context-free rule

```
s -> np vp (1.0)
vp -> verb np (0.8)
vp -> verb np np (0.2)
np -> det noun (0.5)
np -> noun (0.3)
np -> det noun noun (0.15)
np -> np np (0.05)
```

Figure 4.1: A Toy Probabilistic Context-Free Grammar: Charniak, 1997

has been augmented with a probability. Fig. 4.1 shows a toy PCFG (from Charniak (1997)). The probability of a particular parse of a sentence is the product of the probabilities of each of the rules in the PCFG used to obtain the parse: the probability of the parse (4.2) for Sentence (4.1) (also from Charniak (1997)) given the PCFG in Fig. 4.1 is 0.009 (1.0 * 0.3 * 0.2 * 0.5 * 0.3), whereas the probability of the parse (4.3) for the same sentence is 0.036 (1.0 * 0.3 * 0.8 * 0.15):

(4.1) *Salespeople sold the dog biscuits.*

(4.2) (s (np (noun *salespeople*)) (vp (verb *sold*) (np (det *the*) (noun *dog*)) (np (noun *biscuits*))))

(4.3) (s (np (noun *salespeople*)) (vp (verb *sold*) (np (det *the*) (noun *dog*) (noun *biscuits*))))

PCFGs assume that the rule expansions within a parse tree are independent of one another but corpus studies have shown that the probability of one particular structure at one point in a parse can be dependent on the structure in which it is embedded. According to an analysis of the Switchboard corpus (see Jurafsky and Martin (2000, p.456) for a discussion), a noun phrase is more likely to be expanded as a pronoun if it is occupying the subject rather than the object position within a declarative sentence. A second limitation of PCFGs is that they do not make use of the relative co-occurrence frequencies of particular words in the input sentence when determining the most likely analysis for that sentence. This information can be useful when resolving attachment ambiguities (Hindle and Rooth 1991). It is not surprising, therefore, that probabilistic parsers which use PCFGs have been supplanted by probabilistic parsers which use more sophisticated probabilistic models which take into account, for example, the relative co-occurrence of actual words rather than just part-of-speech categories, to improve performance. Well-known probabilistic parsers are Collins (1997), (see also Collins (1999)), Bod (1998) and Charniak (2000). Toutanova and Manning (2002) use probabilistic techniques to select parses produced by the non-probabilistic LKB.

The topic of probabilistic parsing first arose in Section 2.3.2, in the context of a discussion on corpus evidence since a syntactically annotated corpus, or a treebank, is often used as the source of the probabilistic information. Treebanks can also be used as a source of test sentences to evaluate the accuracy of any parser, although they are generally used to evaluate probabilistic parsers: the parse returned by a probabilistic parser for a sentence from the treebank is compared to the treebank analysis of the sentence, which has been produced or approved by a person. The closer the parses produced by the parser are to the treebank parses, the better the parser is considered to be. The most successful probabilistic parsers to date (Collins 1999, Charniak 2000) have been tested in this manner on sentences from the Wall Street Journal section of the Penn Treebank (Marcus *et al.* 1993).

Probabilistic parsers adopt an agnostic attitude to ungrammaticality. They have the potential to be more robust to ungrammatical language than traditional parsers because their linguistic knowledge is based on a large set of naturalistic, potentially ungrammatical sentences. In reality, the corpora which are used to train probabilistic parsers tend to contain sentences which are grammatical rather than ungrammatical, e.g. the Wall Street Journal section of the Penn Treebank, but ungrammatical sentences can still be described because probabilistic parsers have the luxury of being able to produce unlikely parses. Overgeneration is not as serious a problem for probabilistic parsers as it is for traditional parsers because a probabilistic parser has an inbuilt mechanism, i.e. its probabilistic model of grammar, which allows it to select the highest-probability parse from several. This overgeneration means that a probabilistic parser is likely to return a parse for an ungrammatical sentence. This does not entail, however, that the parse is actually correct, i.e. that it actually reflects the meaning of the ungrammatical sentence. The quality of the parses returned by two probabilistic parsers (those described in Collins (1999) and Charniak (2000)) is tested explicitly in Chapter 6.

4.2.3 Partial Parsing

In traditional parsing, the parse or analysis returned by a parser groups the words in the input sentence together in such a way that the meaning of the *entire* sentence is represented. In *partial* or *shallow* parsing, sequences or *chunks* of words are grouped together to reflect aspects of the sentence's meaning but there is no commitment to combining these meaningful chunks together so that one coherent chunk representing the whole sentence is returned. Partial parsers are typically less tied to a theoretical linguistic framework than traditional parsers and the linguistic information returned by a partial parser is often not as fine-grained as that returned by a traditional parser – hence the name *shallow parsing*. The results of partial parsing are particularly useful in applications such as information extraction, where sophisticated grammatical information is not so important.

A well-known partial parser is the Fidditch parser described in Hindle (1994). It is based on the principle of *deterministic* parsing (Marcus *et al.* 1983): once a decision has been made by the parser it is never retracted. Fidditch always returns one phrase structure tree or a set of partial trees for a sentence. If, for example, Fidditch is unsure of where in a tree to attach a constituent, it simply leaves the constituent unattached. Hindle (1994, p.109) gives Sentence (4.4) as an example:

(4.4) *Who did the opposition groups approach about the issues?*

Fidditch returns two partial trees as a result: one for the phrase *Who did the opposition groups approach* and another for the phrase *about the issues*.

Another classic approach to partial parsing is to use *cascades* of finite-state processors to analyse the input sentences in an efficient and incremental fashion (Abney 1990, Abney 1996, Hobbs *et al.* 1997). Abney (1990) describes an early example, the parser CASS (Cascaded Analysis of Syntactic Structure). CASS works by applying a series of filters to tagged text. The first filter, identifies the phrasal chunks of a sentence using a regular expression recognizer; the second filter is applied to the output of the first and identifies clause boundaries and the subject and predicates within a clause; the final filter is applied to the output of the second and attaches the chunks into one parse, if possible. There is no backtracking, but it is *semi-deterministic* because decisions made at one stage can be rectified if necessary at a later stage. Apart from robustness, two major advantages of such incremental parsing is that it is fast and it is easy to maintain. Aït-Mokhtar *et al.* (2002) argue that incremental techniques can be used in conjunction with the detailed linguistic knowledge that is typically associated with traditional parsing to produce more fine-grained linguistic analyses.

Reductionistic parsing is another parsing technique which can be included under the term partial parsing: the general principle behind reductionistic parsing is to assign an input sentence all its possible analyses from the outset and to reduce these analyses by applying constraints. Thus parsing becomes a process of disambiguation. The most well-known type of reductionistic parsing is Constraint Grammar parsing (Karlsson 1990) which assigns dependency-style analyses to an input sentence. Fig. 4.2 shows an example output (from Voutilainen and Tapanainen (1993)) of a Constraint Grammar parser. The basic Constraint Grammar parser is a modular system which incrementally applies constraints at different linguistic levels (part-of-speech, morpho-syntactic, syntactic-functional) to eliminate analyses. Voutilainen and Tapanainen (1993) present an extension in which constraints from the various linguistic levels are applied in parallel rather than sequentially. These constraints are expressed as finite-state automata. A reductionistic finite-state parser for French is described in Chanod and Tapanainen (1996). Another example of reductionistic parsing is the *super-tagging* described by Joshi and Srinivas (1994) which takes place within the linguistic framework of Lexicalized Tree-Adjoining Grammar. Each word in the input sentence is assigned a set of elementary trees or supertags: unsuitable tags are eliminated from the set during a process which they describe as *almost parsing*. The only job left for the actual parser is to combine the elementary trees into a parse for the whole sentence, or sentential fragments in the event that a parse for the whole sentence cannot be found.

Lavie and Tomita (1993) describe a parser called GLR* which is an extension of the GLR parser[1] described in Tomita (1990) but which allows

[1]An LR parser is a shift-reduce parser which does not employ backtracking. It makes

```
I     PRON    @SUBJECT
see   V PRES  @MAINVERB
a     ART     @DET>N
bird  N       @OBJECT
.     FULLSTOP
```

Figure 4.2: Example Analysis Returned by a Constraint Grammar Parser

unparsable sections of the input sentence to be skipped. Given a context-free grammar G and an input string S, GLR* will find a parse of S', the maximal subset of words of S, such that S' is in the language $L(G)$. The possibilities pursued by the parser are limited heuristically and it is evaluated on spoken language within a restricted domain. Lavie and Tomita (1993) conclude that it is capable of parsing more sentences than GLR but of the sentences parsed, the ratio of good:bad parses is greater for GLR than for GLR*. Another parser which can skip words in the input sentence and which is designed to parse transcriptions of spoken language is the Left-Head Corner Island Parser (LCIP) described by Ballim and Russell (1994). The LCIP is a flexible extension of the Definite Clause Grammar (DCG) grammar formalism (Pereira and Shieber 1987): categories on the right-hand side of a rule can be marked as optional, immediate adjacency constraints between the categories on the right-hand side of a rule are not assumed but have to be explicitly marked, and thresholds can be imposed at a grammar or rule-level to limit the amount of input that may be skipped. Special 'ignore' rules can also be written to describe input which is expected but still should

use of a stack and a parse table, which lists an action for each grammar symbol in each state. A GLR parser is an extension of an LR parser that handles ambiguity in the input by allowing multiple actions for each symbol/state pair in the parse table. These actions are explored concurrently.

be ignored, e.g. the false starts common in spoken language. Consider, for example, Sentence (4.5):

(4.5) *John saw Mary and Mark saw them.*

If the thresholds are set so that nothing can be skipped and given a suitable grammar, the LCIP will produce analyses for the substrings *John saw Mary*, *Mark saw them*, *Mary and Mark saw them* and *John saw Mary and Mark*. If the threshold is lowered to the minimum, analyses will also be found for *John saw Mark*, *John and Mark saw them*, *Mary saw them* and *John saw them*. Skipping parsers can be viewed as partial parsers because they do not commit themselves to returning an analysis of the whole input sentence. They are particularly suited to handling ungrammatical sentences in which extra words have been inserted (see Section 5.2.2).

Partial parsers are by nature more robust to ungrammatical input than traditional parsers because their ultimate goal is not to determine whether or not an input sentence can be generated by a grammar. Instead they are trying to extract as much grammatical information from a sentence as possible. Since ungrammatical sentences do tend to retain a lot of grammatical structure, the response of a partial parser to ungrammatical input will tend to be quite similar to its response to grammatical input. According to Abney (1996, p.9):

> ..a common problem with traditional parsers is that correct lowlevel phrases are often rejected because they do not fit into a global parse, due to the unavoidable completeness of the grammar. This type of fragility is avoided when lowlevel phrases are judged on their own merits.

Unless an ungrammatical sentence really is just an unintelligible jumble of words, a partial parser will be able to pick out the coherent elements that remain in most ungrammatical sentences. Menzel (1995) argues that the reductionist approach to partial parsing is inherently robust (to ungrammatical and other unexpected input) because, unlike traditional parsing, it starts from a point of plentiful information and tries to reduce this information as much as possible. This means information will always be returned.

4.3 Handling Ungrammaticality

In this section, various approaches to the problem of parsing ungrammatical language will be described. What these approaches have in common is that it is their explicit aim to make a parser robust to either ungrammatical language or more generally to extragrammatical language. The format of each section is the same: the section begins with a general description of the approach along with references to work which has used this approach. Particular instances are then described in more detail, and finally, the merits and drawbacks of the approach are discussed.

4.3.1 Constraint Relaxation

A popular and intuitive approach to the problem of parsing ungrammatical (and extragrammatical) language is one in which the rules of a grammar are relaxed in some way so that an ungrammatical sentence can be parsed. This approach is known as *constraint relaxation.* Some examples are Weishedel and Black (1980), Kwasny and Sondheimer (1981), Goeser (1990), Douglas and Dale (1992), Douglas (1995), Vogel and Cooper (1995), Reuer (2000), Fouvry (2000) and Fouvry (2003b) (see also Fouvry (2003a)).

Douglas and Dale (1992) describe a constraint relaxation approach implemented in the PATR-II grammar formalism (Shieber 1986) to deal with a category of errors which they call *constraint violation* errors. Constraint violation errors are errors which can be corrected by changing the grammatical form of a word, and hence they represent a subset of the errors which are corrected by substituting one word in a sentence for another (see Section 5.2.1). For each grammar rule, the constraints associated with it are specified as either necessary or optional. Optional constraints may be relaxed. Fig. 4.3 shows an example PATR-II noun phrase rule with numbered constraints. Constraint 4 ensures that the correct form of indefinite article *a* or *an* is used, and Constraint 5 enforces number agreement between a determiner and a noun in a noun phrase. In order for the ungrammatical noun phrase *an dog* to be parsed, constraint 4 would have to be specified as relaxable, and in order for *this dogs* or *these dog* to be parsed, constraint 5 would have to be relaxable. An interesting aspect of this approach is the idea of *relaxation levels*: constraints may be relaxed at different levels, resulting in degrees of permissibility. Fig 4.4 shows the basic constraint relaxation algorithm. The constraints in the set R after this algorithm has been applied indicate the nature of the ungrammaticality. This approach takes place within the context of developing a parser for The Editor's Assistant, an automatic proofreading tool. Douglas (1995) describes a more declarative extension of the robust PATR formalism described in Douglas and Dale (1992), in which the feature paths where errors are likely to occur are specified in advance, and these are compiled into the relaxable constraints of the original rule-based approach.

In the constraint relaxation system described by Douglas and Dale (1992) and Douglas (1995), the constraints which can be relaxed are specified in advance. These relaxable constraints correspond to potential errors in the

```
X0 -> X1 X2
1. <X0 cat> = NP
2. <X1 cat> = Det
3. <X2 cat> = N
4. <X1 agr precedes> = <X2 agr begins>
5. <X1 agr num> = <X2 agr num>
6. <X0 agr num> = <X2 agr num>
```

Figure 4.3: A PATR-II rule with Numbered Constraints: Douglas & Dale, 1992

When applying rule r at relaxation level L
$N \leftarrow$ necessary constraints on r at L
$O \leftarrow$ optional constraints on r at L
$R \leftarrow \{\}$
if all $n \in N$ can be solved **then**
    incorporate any instantiations required
    for each $o \in O$ **do**
        if o can be solved **then**
            incorporate instantiations
        else
            $R \leftarrow R \cup o$
else
    return failure
return R

Figure 4.4: Basic Constraint Relaxation Algorithm: Douglas & Dale, 1992

input. A more general approach for parsers which operate on typed feature structure grammars, e.g. the LKB (Copestake 2002) or ALE (Carpenter and Penn 2001), is the one described by Fouvry (2003b). A feature structure can be viewed as a directed acyclic graph which expresses linguistic information about a word or a phrase. In a typed feature structure system, each node in a feature structure is associated with a type and the types are organized into an inheritance hierarchy with a unique most general type which subsumes all other types in the hierarchy. The type hierarchy has the property that if two types subsume another type, then there is a unique most general type which they both subsume, which is the *join* or *unification* of the two types (see Section 7.3 for a more detailed description of typed feature structure systems). Feature structure unification is based on type unification. In a typed feature structure grammar, the grammatical symbols which describe words and phrases are typed feature structures, and parsing using a typed feature structure grammar involves combining words into phrases and phrases into sentences using feature structure unification. If two feature structures fail to unify, then parsing fails. Fouvry extends the notion of a type hierarchy so that every pair of types have a join - this has the consequence that unification will never fail and therefore, that ungrammaticality will not result in parse failure. I return to the topic of constraint relaxation and unification-based parsing in more detail in Chapter 7, when I describe my own contribution to this area and compare it to others including Fouvry's.

A drawback of the constraint relaxation approach to robust parsing is that it needs to be constrained in some way. If all constraints in the grammar can be relaxed, then the grammar is not saying very much - any string of words will be accepted, with a potentially enormous set of possible syntactic analyses. While it is clear that a parser should be able to parse ill-formed input, it is also clear that there is some input that should be considered unparsable. An example is Sentence (3.19) from Section 3.6.1, which proved too much of a challenge for one human parser. Douglas and Dale place a limit on the constraint relaxation by specifying in advance what constraints can be relaxed. This is a sensible approach, although it does require a certain amount of empirical effort to decide what constraints should be relaxed. Fouvry's approach has the advantage that it requires no error anticipation, and hence has the capacity to parse anything. However, Fouvry is also faced with the challenge of limiting the constraint relaxation, and he does this by defining a notion of *information loss* which, as the name suggests, allows feature structures to be ranked in terms of the information which was lost in arriving at them.

Another drawback of the constraint relaxation approach is that it is well suited to parsing ungrammatical sentences which contain the same amount of words as their grammatical counterparts (see Section 5.2.1), but, while not incompatible with, it is not particularly suited to parsing ungrammatical sentences which contain fewer or more words than their grammatical versions (see Sections 5.2.2 and 5.2.3). In fact, Douglas and Dale (1992) suggest other techniques for the parsing of missing or extra words.

An advantage of using constraint relaxation for parsing ungrammatical sentences is that the same grammar can be used to parse both grammatical and ungrammatical sentences, with the result that there is a close pairing between the grammatical sentences and the ungrammatical ones. Douglas and Dale (1992) remark that constraint relaxation "provides a precise and systematic way of specifying the relationship between errorful and 'correct' forms, making it easier to generate suggestions for corrections." However, this property of constraint relaxation has been interpreted as a disadvantage by Gojenola and Oronoz (2000), who argue that constraint relaxation approaches suffer because they are dependent on the quality of the grammar. If the grammar does not contain a rule for a particular type of construction, then a well-formed or an ill-formed sentence exhibiting that structure will not be properly parsed.

4.3.2 Error Anticipation

An *error anticipation* approach to the problem of parsing ungrammatical language is a straightforward solution which involves anticipating the kind of ungrammaticality that might occur and building a description of this into the grammar. Examples of approaches which use error anticipation are Weischedel and Sondheimer (1983), Schneider and McCoy (1998) (see also McCoy and Masterman (1997)), Gojenola and Oronoz (2000), Bredenkamp *et al.* (2000) and Bender *et al.* (2004).

Weischedel and Sondheimer (1983) describe an augmented transition network (ATN) parser which contains *meta-rules* corresponding to patterns of ill-formedness. Each meta-rule corresponds to a conventional rule and it is invoked during a parse when a conventional rule fails.

There is no clear separation between the parser and the grammar in an ATN, and hence Weischedel and Sondheimer's approach is quite procedural in nature. Other error anticipation approaches are more declarative, where the rules which describe errors are separate from the parsing process and represent a model of ungrammatical language. *Mal-rules* or *error productions* are used within the field of applied linguistics to describe the errors typically made by the learners of a language. Schneider and McCoy (1998) describe the ICICLE system which identifies syntactic errors in the writing of native speakers of American Sign Language who are learning English as a second language. This system consists of a chart parser and a broad-coverage context-free grammar augmented with features. The grammar contains mal-rules to model the errors which are expected to be made by this community.

Bender *et al.* (2004) describe the Arboretum system which is based on the LKB (Copestake 2002) and the ERG (Copestake and Flickinger 2000), and which makes use of the LKB's dual capabilities as a parser and a generator to provide corrections for the ungrammatical sentences produced by those who are learning English as a second language. Mal-rules describing the typical syntactic errors made by learners of English are incorporated into the ERG and the semantic representations produced by the LKB parser for ill-formed sentences (using the mal-rules) are then used as a basis for generating well-formed corrections. Initial tests of this prototype system have been carried out using sentences from the SST corpus (Emi *et al.* 2004) (see Section 3.7.1): correct parses were achieved for 178 items out of 221, and the generation phase produced corrections for 167 out of these 178 items. Some of these corrections involved spurius changes to the ill-formed sentence beyond what was required, and further work in the Arboretum project involves the refinement of the correction generation strategy so that such spurious changes are minimized.

Gojenola and Oronoz (2000) describe a parser for the Basque language which uses finite-state processing techniques to handle errors. Based on an analysis of errors found in date expressions in student essays and newspapers, they wrote regular expressions for six frequent error types.[2] These regular expressions were converted into finite-state automata. The finite state processing was applied to sentences after they had been processed by a

[2]This study was considered too narrow to be included in the discussion of error corpora in Section 3.7.1. This is because they explicitly only searched for errors in date expressions.

morphological analyser followed by a unification-based partial chart parser: of the test data, thirty-one out of thirty-seven errors were detected and there were five cases where a grammatical structure was flagged as erroneous.

The main advantage of the error anticipation approach to robust parsing is that it is an empirical approach. It focuses on one aspect of extragrammaticality, ungrammaticality, and attempts to deal with that. In order to write error rules, it is necessary to know what kind of errors might occur, and in order to acquire this knowledge it is necessary to look at real ungrammatical language. This is essentially the same argument put forward in Section 2.4.2 in Chapter 2. Thus, the error corpus described in Chapter 3 can form the basis of an error anticipation approach to robust parsing. In Chapter 5, I describe a parser which I have developed which uses error anticipation based, to a certain extent, on an analysis of the kinds of errors which occur in *corpus1*.

The empirical merit of an error anticipation approach brings with it the drawback that a lot of effort is involved in writing a grammar of error rules. There are two kinds of effort involved: the first is compiling an error corpus, an arduous process, as I remarked in Section 3.7; the second is the actual analysis of the errors and the subsequent writing of the error grammar. This has the potential to be done automatically if the error corpus is large enough. Menzel (1995) is pessimistic about the prospect of learning an error grammar automatically:

> It would require huge amounts of sufficiently deviant utterances being available as training data. This renders the approach technically infeasible and cognitively implausible. (Menzel 1995, Footnote One)

A sufficiently large error corpus could be obtained if errors were artificially introduced into an existing large corpus based on an error analysis such as the one described in Chapter 5. Bigert (2004), for example, introduces artificial context-sensitive spelling errors into the sentences in a grammatical corpus, and then uses the resulting corpus as training data for an n-gram statistical approach which detects this kind of error (see also Atwell and Elliott (1987) for earlier work in this area).

Like constraint relaxation, the error anticipation approach needs to be tightly controlled so that an explosion in ambiguity does not result. For example, when a sentence is parsed using the ICICLE system, the set of mal-rules is available along with the set of normal rules from the outset, with the consequence that grammatical sentences can be parsed with mal-rules and hence flagged as ungrammatical. Since there are no restrictions on when the mal-rules can be applied and on how many can be applied at any one time, the problem of spiralling spurious ambiguity is also an issue here.

Fouvry (2003b, p.55) remarks that there is a fine line between an error anticipation approach to parsing ill-formed input and a constraint relaxation one. For example, Douglas and Dale (1992) need to anticipate errors before they can specify what constraints can be relaxed, and since Weishedel and Sondheimer's meta-rules correspond to conventional rules, they are quite similar to rules which have been relaxed in some way. Both constraint relaxation and error anticipation need to be controlled in some way so that the set of possible parses for a sentence does not get out of hand. However, error anticipation seems to be a more flexible approach since an error grammar can be derived from a conventional grammar (so that the two are closely linked) but it can also be seen as a separate entity in its own right and as a clear model of ungrammaticality. Because of this, it provides a more natural way of handling all types of errors including those involving extra or missing words.

4.3.3 Minimum Edit Distance

In the field of spelling correction, the term *edit distance* or *Levenshtein distance* refers to the number of changes (insertions, deletions, substitutions or transpositions) in letter that need to be made in order to correct the spelling of a word (Damerau 1964, Levenshtein 1966). For example, changing the string of letters *outselves* to the valid English word *ourselves* involves one substitution operation, and hence the edit distance is one. One possible correction will be favoured over another if its edit distance is smaller. This approach can be applied to parsing also: if a sentence fails to parse, it is changed in some way so that a parse can be found. So, the minimum edit distance approach to parsing involves generating possible corrections of an ill-formed sentence and trying out these corrections by re-parsing the sentence. A correction is valid if it results in a full parse, and valid corrections are ranked so that minimal change to the input sentence is favoured. The difference between minimum edit distance in robust parsing and minimum edit distance in spelling correction is that, in parsing, the changes are applied to words rather than letters. Examples of minimum edit distance robust parsing approaches are Lee *et al.* (1995) and Weng (1993). A skipping parser such as the GLR* parser described by Lavie and Tomita (1993) or the extended DCG formalism described by Ballim and Russell (1994) (see Section 4.2.3) can also be seen as a special kind of minimum edit distance parser which allows only word deletions.

Within the domain of programming language compilation, Lyon (1974) presents a minimum edit distance chart parsing algorithm for context-free grammars called the *least-errors recognition algorithm*. This algorithm assumes that an input sentence can contain errors and it tries to parse the sentence by postulating the least number of errors possible. The errors that can be hypothesized are the insertion, deletion or mutation (substitution) of a terminal symbol. Assume, for example, the active edge (4.6) at position i in the input string:[3]

(4.6) `T -> * a [i,i], 0`

The mutation-error hypothesis will result in the creation of inactive edge (4.7), the insertion-error hypothesis will result in the creation of active edge (4.8) and the deletion-error hypothesis will result in the creation of inactive edge (4.9). Note that these edges have an error cost of 1 associated with them, because, in creating them, one error was hypothesized.

(4.7) Mutation-Error Hypothesis:

```
T -> a * [i,i+1], 1
```

(4.8) Insertion-Error Hypothesis:

```
T -> * a [i,i+1], 1
```

(4.9) Deletion-Error Hypothesis:

```
T -> a * [i,i], 1
```

Lee *et al.* (1995) present an extended version of the least-errors recognition algorithm which allows both terminal and non-terminal symbols to be inserted, deleted or substituted. Their focus is on the general problem

[3] The chart edge notation in (4.6) is explained as follows: a non-terminal category `T` is looking for the terminal category `a` at position i in the input string. The * before the terminal category symbol `a` indicates that the `a` has yet to be found. This edge is active because it has not yet satisfied its expectations. The 0 after the edge indicates that there is no error cost associated with this edge.

of extragrammatical input as opposed to the more specific problem of ungrammatical input. The algorithm runs after a normal parsing algorithm has failed to produce a parse, and it uses heuristics to constrain the search space. In order to evaluate the algorithm, they derive a grammar directly from the Wall Street Journal section of the Penn Treebank, and discard the low-frequency rules. The sentences in the Penn Treebank which require these low-frequency rules are examples of the extragrammatical input which their algorithm targets. Using this grammar, the algorithm is tested on sentences from the Wall Street Journal and ATIS corpus sentences from the Penn Treebank, achieving accuracy levels of between 68% and 77%. They run their algorithm with and without heuristics, and show that the version with heuristics is more efficient and more accurate.

An advantage of minimum edit distance edit parsing is that it is flexible and does not rely on handcrafted knowledge of particular errors in order to be implemented. A second advantage is that it explicitly produces possible corrections for an ungrammatical sentence, which is a useful thing to do in the context of grammar-checking.

A third advantage of minimum edit distance parsing is that it is guaranteed to find an analysis for an ungrammatical sentence. However, Lavie and Tomita (1993) make the important point that returning a parse for an ungrammatical sentence is not the same as returning a good parse of the sentence:

> Compared with the GLR* parser, the original GLR parser [*the non-robust version of GLR**], although fragile, returned results of relatively good quality, when it succeeded in parsing the input. The GLR* parser, on the other hand, will succeed in parsing almost any input, but the parse result may be of little or no

> value in a significant portion of cases. (Lavie and Tomita 1993, p.12)

This links back to the point made in Chapter 2: a parser has a better chance of providing an accurate parse for an ungrammatical sentence if it has some knowledge of the kind of ungrammatical language that is likely to occur. Of the four approaches to robust parsing described in this chapter, the error anticipation approach is the most compatible with this way of thinking.

A major drawback of minimum edit distance parsing is linked to its flexibility: since any sentence can be transformed into another sentence using insertion, deletion, substitution and transposition, the transformations that can be applied to the ill-formed sentence need to be controlled in some way. Lee *et al.* (1995) use various heuristics to control the editing of the unparsable sentence. For example, they assume that mistakes are more likely to occur between constituents than within a constituent and so edits that occur between two constituents are assigned a lower edit cost than those that occur within constituents.[4] Penstein Rosé and Lavie (1997) carry out an experiment which compares their two stage parse fitting approach (described in more detail in the next section) to a restricted minimum-distance-edit approach which uses only insertion and deletion transformations. When the minimum edit distance approach is constrained so that a maximum of five transformations can be applied per sentence, the results are very discouraging. Not surprisingly, they are more encouraging when this threshold is reduced from five to one, but the two-stage parse fitting approach is still

[4]One can imagine that this heuristic works well for extragrammatical sentences containing interjections, e.g. the sentence *This is, **I am sure**, what he means* used as an example by Lee *et al.* (1995). However, the error analysis on *corpus1* sentences carried out in Chapter 5 does not provide evidence for this heuristic.

faster. Interestingly, according to the analysis of the differences between *corpus1* and *corpus2* described in Chapter 5, the majority of ill-formed sentences can be corrected by applying just one transformation (see Fig.5.1). This fact is also exploited in the development of an error-handling parser, also described in Chapter 5.

4.3.4 Parse Fitting

Parse fitting is a robust parsing technique which takes the results produced during normal parsing for an ungrammatical sentence and pieces together these results in a sensible way so that the ungrammatical sentence can be analysed. Thus, parse fitting approaches involve a two-stage parsing process. Examples are Jensen *et al.* (1983), Strzalkowski (1992), Grinberg *et al.* (1995), Nederhof *et al.* (1997) and Penstein Rosé and Lavie (1997).

Jensen *et al.* (1983) describe the robust parse fitting technique used in the EPISTLE system, which performs grammar checking on business English (see also Miller (1982) and Richardson and Braden-Harder (1988)). The parse fitting is applied after the parser has failed to parse a sentence using the core grammar rules. The recovery parse fitting stage has itself two stages: during the first stage a head constituent is identified, and during the second, attempts are made to attach the other constituents on either side of the head constituent. Linguistic and processing heuristics are used to guide the fitting procedure, for example, tensed verbs are viewed as the best candidates for a head constituent, and a constituent is considered more likely to be a head constituent if it spans a large portion of the input sentence. Most of the examples Jensen *et al.* (1983) give of parse-fitting in action are actually grammatical constructions which are not covered by their core grammar rules rather than ungrammatical constructions, e.g. sentence fragments and elliptical utterances. One example is Sentence 4.10 which is not parsed by the core grammar but which is fitted together by attaching the noun *Example* and the two punctuation symbols ":" and "." to the longest tensed verb phrase found by the core grammar, i.e. *Your percentage of $250.00 is $187.50*:

(4.10)*Example: Your percentage of $250.00 is $187.50.*

The Epistle system also uses constraint relaxation to detect common grammatical errors such as agreement violations.

Penstein Rosé and Lavie (1997) describe a two-stage process for parsing extragrammatical input. Their approach works as follows: a version of the GLR* skipping parser (Lavie and Tomita 1993) (see also Section 4.2.3) produces partial parses for utterances for which no full parse can be found. A genetic programming algorithm is then used to find the best combination of these partial parses so that a meaning representation for the full utterance can be returned. The two-stage parser is part of a spoken language machine translation system called Janus in the restricted domain of meeting scheduling. They evaluate this approach by comparing it to a minimum edit distance parser which allows only insertions and deletions, and conclude that it is considerably faster (see Section 4.3.3).

Grinberg *et al.* (1995) present a robust parsing algorithm for the link grammar formalism (Sleator and Temperley 1993). If a full parse cannot be found for an input string during the first parsing phase, the second phase allows null links to be established between words in the input, and calculates the minimum number of null links that are needed to parse each substring. They compare this algorithm to one in which cost is calculated not in terms of null links but in terms of an edit-distance metric, and conclude that the former is more efficient.

A significant advantage of the two-stage parse fitting parsing technique is that there is no interference with the normal parsing process. This means that unlike other approaches, for example the mal-rule approach to parsing second language errors (Schneider and McCoy 1998) described in Section 4.3.2, there is no deterioration in the time it takes to parse a grammatical sentence.[5] Penstein Rosé and Lavie (1997) demonstrate that the parse fitting approach is a good deal more efficient than the minimum edit distance approach. However, it shares with the minimum edit distance approach the advantage of not needing any handcrafted error rules.

A disadvantage of the parse fitting approach is that, in fitting together the pieces of a parse, it is quite reliant on heuristics, and to some extent, its performance will depend on how sophisticated these heuristics are. Although Rosé and Lavie's approach is decidedly more efficient than a minimum edit distance approach it is only slightly more accurate.

Mellish (1989) describes a two-stage chart parsing process which can be used to parse ungrammatical sentences containing unknown words, missing words and extra words using a context-free grammar. During the first stage, a sentence is parsed using a bottom-up chart parser. A bottom-up strategy ensures that all potential partial parses are found. If no parse spanning the entire input sentence is found, then the second recovery stage is embarked upon. During this stage, a generalized top-down chart parsing algorithm is applied to the same input in order to locate the errors. This top-down parsing process makes use of the edges found during the initial phase so that no work is repeated. A top-down chart parsing algorithm (Earley 1970) is modified so that the fundamental rule can combine active edges with inactive

[5]Note that the minimum edit distance approach(Lee *et al.* 1995) is also a two-stage approach, indicating that a precise categorization of each approach is not possible.

edges at any position in the chart. Recovery chart parsing rules relating to unknown words, missing words and extra words are also applied during the second phase in order to allow the sentence to be parsed. The search space is constrained by using heuristics to order the edges on the chart agenda. An example of one such heuristic is the preference of edges which will result in a parse which spans a larger portion of the input string. Another heuristic prefers edges which arise from the fundamental rule over those arising from the top-down rule, presumably because the top-down rule introduces more active edges whereas the fundamental rule brings active edges closer to completion. The approach was evaluated on a very small grammar using sentences containing less than thirteen words generated from the grammar and into which errors were randomly introduced, and it was found that the time it takes to enumerate all possible parses of an ungrammatical sentence using this method was no more than ten times longer than the time it takes to parse the corresponding error-free sentence. A similar strategy is adopted by Kato (1994). It is difficult to know how to classify these two approaches. They could be classified as minimum edit distance approaches because the parsing rules which correspond to the identification of an extra word, substituted word or missing word are strongly reminiscent of minimum edit distance parsing in which the parsing algorithm assumes that a terminal or non-terminal symbol has been inserted, substituted or omitted. They can be classified also as parse fitting approaches because the second top-down phase proposes ways in which the partial parses found during the initial bottom-up phase can be fitted together.

4.4 Conclusion

In this chapter, I have given a brief overview of three parsing paradigms, the traditional parsing paradigm in which ungrammatical sentences are rejected; the probabilistic parsing paradigm which aims to parse realistic language and which does not sharply separate the grammatical from the ungrammatical; partial parsing which aims to extract as much grammatical information as possible from a sentence, be it grammatical or ungrammatical. The identification of the three parsing paradigms leads into a discussion of the many and various attempts that have been made in the last thirty years to allow parsers to cope with the problem of ungrammaticality. I have crudely categorized these attempts by dividing them into four groups: the constraint relaxation solutions in which the parser's grammar is made more lenient to allow ungrammaticality to be analysed, the error anticipation solutions in which ungrammatical language is explicitly described, the minimum edit distance solutions in which corrections are proposed for the ungrammatical sentence so that it can be parsed and the parse fitting solutions in which the partial parses for an ungrammatical sentence are fitted together into one parse. Of these four approaches, error anticipation seems to concentrate the most on the specific problem of *ungrammatical* input, whereas parse fitting seems the most suited to the more general problem of *extragrammatical* input. In Chapters 5 and 7, I present my own attempts to parse ungrammatical language, both attempts within the traditional parsing paradigm. The approach described in Chapter 5 is an error anticipation approach with aspects of constraint relaxation, and to lesser extent, minimum edit distance and parse-fitting, whereas the approach in Chapter 7 is a constraint relaxation approach with aspects of error anticipation. The corpus described in Chapter 3 has a role to play in both.

Chapter 5

Error Analysis and A Parsing Experiment Using an Error Grammar

5.1 Introduction

This chapter has two aims: to present an analysis of the errors in the ungrammatical *corpus1* sentences, and to describe and present the results of a parsing experiment which makes use of some of the analysis results. In Section 5.2, the differences between *corpus1* and *corpus2* are analysed. This analysis is carried out by looking at how the ungrammatical sentences in *corpus1* were transformed into the grammatical sentences of *corpus2*. Section 5.3, the second part of this chapter, presents a technique for parsing ungrammatical sentences. The parsing approach is an error anticipation approach because it uses a grammar of error rules to parse ungrammatical sentences. However it also displays features of constraint relaxation, parse fitting and minimum edit distance parsing (see Section 4.3).

5.2 Analysing the Ungrammatical Sentences

Every time an ungrammatical sentence is added to *corpus1*, a note is made of the operation that is needed to correct the sentence. It is this information which is analysed. Fig. 5.1 indicates the correction operators which were applied to the sentences in the corpus. The figures in brackets beside the correction operator indicate the number of times a sentence was corrected by applying this operator (as a percentage and an absolute frequency). For each type of correction, two examples are given from *corpus1* and *corpus2*. According to James (1998, p.114), the same ordering of the substitute, delete and insert correction operators was found in a study of native speaker spoken language slips carried out by Stemberger (1982).

A possible correction operator not listed in Fig.5.1 is the *move* operator which moves a word from the sentence into another position within the

Correction Operator	**Examples**
Substitute a word (48.2%, 590)	*the theory* ***in*** *empirical*→ *the theory* ***is*** *empirical* *staff* ***was*** *allowed to return*→ *staff* ***were*** *allowed to return*
Insert a word (24.5%, 299)	*Will be declaring their undying love for each other?*→ *Will* ***they*** *be declaring their undying love for each other?* *we must assume the validity this induction principle*→ *we must assume the validity* ***of*** *this induction principle*
Delete a word (16.7%, 204)	*We* ***we*** *have them all*→ *We have them all* *a joint development which will* ***the*** *provide 10 new apartments*→ *a joint development which will provide 10 new apartments*
More than one of above (10.6%, 129)	*This means* ***to allow*** *structure-sharing*→ *This means structure-sharing* ***is allowed*** *What does a single line* ***yellow*** *mean?*→ *What does a single* ***yellow*** *line mean?*

Figure 5.1: Some Error Analysis Results

sentence. However, since such an operation occurs relatively infrequently (a sentence was corrected by applying this operator alone 1.5% of the time), the decision was taken to define it in terms of the insert and delete operators, and to include errors which could be corrected in this way in the class of *composite errors* or errors which can be corrected by applying more than one correction operator. Of course, it is also possible to define the substitute operator in terms of the insert and delete operators, but since this is the most frequent correction operator, it is seen as a valid operator in its own right. I will return to this issue in Section 5.3.2.

In the following sections, the substitute, insert and delete correction operators are described in more detail. The instances of each type of correction are classified in the way that seems most natural for the correction operator in question. For each classification an example is given. Please note that when a correction is supplied for an example ungrammatical sentence, this is not necessarily the only way that the sentence has actually been corrected: some *corpus1* sentences have not one but two valid corrections (see discussion of this point in Section 3.4).

5.2.1 Substitution Errors

590 errors could be corrected by substituting one word within a sentence for another. These substitutions were examined and classified in terms of the similarity between the word that was substituted and the substituting word. The classifications are shown in descending order of frequency.

1. **Context-Sensitive Spelling (188)**: A word can be substituted for a word similar to it in spelling. I classified two words as similar in spelling if the Levenshtein distance between them is one (e.g. *the* and *they*), or if the Levenshtein distance is two and the first letter in each word is

the same (e.g. *to* and *the*).[1] Sentence (5.1) is an example containing a context-sensitive spelling error:

(5.1) *It was one part Del Boy, one part Minder and the rest Fair City, with Roy* ***form*** *Corrie parachuted in as the groundsman.* (A.107)

106 of these context-sensitive spelling errors involved the substitution of 1–5 letter function words from the following list: *a*, *as*, *an*, *at*, *from*, *in*, *is*, *it*, *its*, *my*, *no*, *our*, *or*, *of*, *on*, *the*, *they*, *to*, *too*, *that*, *there*, *he*, *then*, *you* and *your*.

2. **Number/Person (114)**: If the word is a noun, pronoun, verb or determiner, it can be substituted by the same word with a different value for a number or person feature, e.g. third person *awaits* with first person *await* in Sentence (5.2):

   (5.2) *I* ***awaits*** *your response.* (A.745)

   Sentence (5.2) is an example of an agreement error since there is a person clash between the first person pronoun *I* and the third person verb *awaits*. Sentence (5.3) is an example which could be corrected by changing the number of a word (in this case, the noun *hideaway*) but there is no agreement clash between this word and another word in the sentence:

   (5.3) *So the only way we could feel at ease was by going abroad together to secret* ***hideaway***. (A.104)

[1] There were two cases falling outside this definition which were nevertheless classified as context-sensitive spelling errors. Both of these were caused by the predictive text messaging facility on mobile phones – both were two letter words which were corrected by replacing the two letters with another two letters alphabetically adjacent to the original two, e.g. *of* with *me* or *on* with *no*.

I return to the subject of agreement errors in more detail in Chapter 7.

3. **Verb Form (65)**: If the word is a verb, it can be substituted by the same verb with a different form, e.g. finite *became* with infinitival *become* in Sentence (5.4):

   (5.4) *In his presence, Brent would often* ***became*** *stunned by resentment...* (A.565)

4. **Homophone confusion (53)**: A word can be substituted by a word with the same or very similar pronunciation but with a different meaning, e.g. *there* with *their* in Sentence (5.5):

   (5.5) *Everybody who's anybody will be at* ***there*** *gig in Whelans on the 13th August.* (A.296)

5. **Same Root/Different Syntactic category (41)**: A word can be substituted by a word with the same lexical root but with a different syntactic category, e.g. noun *plausibility* with adverb *plausibly* in Sentence (5.6):

   (5.6) *Yet it captures quite* ***plausibility*** *a certain predominant form of life today.* (A.581)

6. **Case (37)**: All case marking errors found in the corpus involved the incorrect use of the genitive case on pronouns and nouns, e.g. the noun *rivers* instead of the possessive *river's* in Sentence (5.7):

   (5.7) *As it was Sunday the river was lined with Thai families having picnics in the bamboo picnic huts at the* ***rivers*** *edge* (A.196)

   In Sentence (5.8) the possessive noun *client's* has been used instead of the plural noun *clients*:

(5.8) *They are responsible for purchasing ad break slots for* ***client's*** *who want their brands associated with a given programme.* (A.527)

7. **Prepositions (33)**: If the word is a preposition, it can be substituted by another preposition, e.g. *of* with *from* in Sentence (5.9):

(5.9) *Chapter 6 describes the implementation that resulted* ***of*** *the experiment.* (A.380)

8. **Indefinite articles (12)**: If the word is the indefinite article *a*, it is substituted by the indefinite article *an*, or vice versa:

(5.10) *Head-driven Phrase Structure Grammar is* ***an*** *theoretical framework for expressing generalizations about linguistic phenomena in a way that integrates information supplied from syntax, semantics and pragmatics.* (A.874)

9. **Auxiliary Verbs (11)**: Sometimes the wrong auxiliary verb was used, e.g. the verb *has* was used instead of the copula *is* in Sentence (5.11):

(5.11) *A new object* ***has*** *created.* (A.281)

10. **Synonymous Verbs (6)**: If the word is a verb it can be substituted by a synonymous verb which has a different grammatical distribution, e.g. the verb *comprises* can be substituted by *consists* before the preposition *of* in Sentence (5.12):

(5.12) *The accommodation* ***comprises*** *of 1 bedroom, 1 reception, fully equipped kitchen.* (A.712)

11. **Infinitival *to* and prepositions (3)**: The infinitival marker *to* can be substituted by the preposition *of* in Sentence (5.13):

(5.13) *I actually think I've got a better chance of winning Pop Idol than I had* ***to***[2] *winning You're A Star in Ireland.* (A.222)

Similarly, the preposition *for* can be replaced with the infinitival marker *to* in Sentence (5.14):

(5.14) *The jury of 10 men and two women agreed with the prosecution that Milroy-Sloan was a "cunning fantasist" who was out* ***for*** *find fame and fortune.* (A.424)

12. **Same root/same part of speech/different meaning(2)**: This occurred when two words with the same root and part of speech category but with a similar yet different meaning were confused, e.g. the covert error in Sentence (5.15) where the noun *compilation* has been used instead of the noun *compiler*:

    (5.15) *The decision to classify a sentence as ungrammatical will depend on whether or not the corpus compiler thinks that an error has occured, and as such is dependent on the linguistic intuition of the corpus* ***compilation***. (A.724)

13. **Conjunctions (2)**: If the word is a conjunction it can be substituted by another conjunction, e.g. *and* can be substituted by *then* in Sentence (5.16):

    (5.16) *Once we allow the scientist to invent the facts he uses for hypothesis testing,* ***and*** *the door is wide open to self-fulfilling predictions.* (A.255)

[2] *To* is considered an infinitival marker rather than a preposition if it is followed by any form of a verb.

14. **Full Stop with Complementizer (2)**: A full stop can be replaced by the complementizer *that*. Sentence (5.17) is an example:

    (5.17) *Don't let anyone tell you**.** You are over-ambitious.* (A.862)

15. **Pronoun with Noun (1)**: If the word is a pronoun it can be replaced by a noun: an example is Sentence (5.18) where the pronoun *he* is replaced by a noun:

    (5.18) *Everyone was looking at me to see what I would do, but the **he** just reached in...* (A.571)

16. **Adjectival Form (1)**: If the word is an adjective, it can be substituted by the same adjective with a different form, e.g. the comparative form *clearer* can be changed to *clear* in Sentence (5.19):

    (5.19) *It is convenient to use the setter method from the syntaxClass, making it more **clearer** which subtrees are set to what nodes.* (A.175)

17. **Other (19)**: There is no obvious similarity between the word to be substituted and the word substituting it, e.g. *when* with *with* in Sentence (5.20):

    (5.20) *But **when** one as doggedly uninteresting as this it seems like an aeon.* (A.287)

5.2.2 Insertion Errors

299 errors could be corrected by inserting a word into the sentence. These insertions were examined and classified in terms of the syntactic category of the word that was inserted:

1. **Determiners (81)**: In all but one of these cases, a determiner needed to be added before a *singular* noun. The determiners which were added were either *the* (40) or *a/an* (39), *this* (1) or a possessive pronoun (1). An example of an ungrammatical sentence containing a missing determiner is (5.21):

   (5.21) *Write a program that generates sentences* ***in German*** *language.* (A.883)

2. **Verbs (67)**: In 25 cases it was an auxiliary verb that needed to be inserted as in Sentence (5.22):

   (5.22) ***I planning*** *on organising some gigs on friday nights....* (A.831)

   In 25 of the 42 cases involving a missing main verb, the main verb in question was some form of the copula *be*, e.g.

   (5.23) *The if-statement condition needs to check whether the index* ***selected greater*** *than or equal to six.* (A.907)

3. **Prepositions (60)**: A frequently omitted preposition was the possessive *of* (20), e.g.

   (5.24) *These basic phrases are produced by a* ***collection finite-state*** *rules compiled into a transducer.* (A.798)

   In another 21 cases, the omitted preposition was required by another lexical item in the sentence, e.g. the preposition *with* after the verb *deal* in Sentence (5.25):

   (5.25) *Of course,* ***dealing this*** *kind of rule requires some changes to our parsing scheme.* (A.318)

The remaining cases (19) comprised omissions of the following prepositions: *to*, *by*, *with*, *from*, *in*, *for*, *about*, *up* and *between*.

4. **Pronouns (29)**: Pronouns needing to be added to the object position accounted for 12 cases, subject pronouns 7, relative pronouns 9, reflexive pronouns 1. An example of an ungrammatical sentence which can be corrected by adding a subject pronoun is (5.26):

   (5.26) *Being so* ***gifted can*** *find fulfillment in many employments.* (A.759)

5. **Nouns (20)**: An example of an ungrammatical sentence in the corpus which was corrected by adding a noun (in this case, the noun *sign*) is (5.27):[3]

   (5.27) *And, as he made clear in last week's speech to the American Enterprise Institute, it would be an* ***earnest of*** *America's commitment to the democratic transformation of the Middle East, which would help to solve the Israel-Palestine problem.* (A.146)

6. **Infinitival *to* (19)**: An example of an ungrammatical sentence in the corpus which was corrected by inserting the infinitival marker *to* is (5.28):

   (5.28) *It is also obvious that many further rules* ***remain be*** *written.* (A.791)

7. **Punctuation (11)**: In 9 of the 11 cases involving a missing punctuation symbol, the symbol that needed to be added was a full stop, e.g.

[3] In cases where a noun has been omitted from *corpus1*, the corresponding *corpus2* sentence contains what I believe to be the most contextually plausible noun. This does not rule out corrections involving other nouns, since I consider all corrections involving sensible noun insertions to be equivalent.

Sentence (5.29)[4]:

(5.29) *Event organisers effectively laid siege to an enduring* ***tradition Hollywood*** *frivolity, it seems, has been a casualty of the war in Iraq.* (A.816)

The other two missing punctuation symbols were a hyphen and a bracket.

8. **Conjunctions (5)**: The corpus included examples of omissions of the following conjunctions: *and*(2), *as*(1), *if*(1) and *but*(1). An example of a sentence from *corpus1* which is missing a conjunction (in this case, *if*) is (5.30):

   (5.30) *The attraction is that it is a passive system, it sits on the side of the network* ***and it*** *fails, which none have to date, there is no negative effect on the network.* (A.819)

9. **Adjectives (2)**: An example of a sentence in *corpus1* which can be corrected by adding an adjective is Sentence (5.31), in which the adjective *Australian* can be inserted between the words *an* and *footie*:

[4] I took the decision when compiling the corpus not to include punctuation errors involving a comma. This is because there does not seem to be a widely agreed upon set of rules surrounding their use in modern written English. Truss (2003, p.81) remarks, "I have seen an essay on the Internet seriously accusing John Updike, that wicked man, of bending the rules of the comma to his own ends "with fragments, comma splices, coordinate clauses without commas, ellipted co-ordinate clauses with commas, *and more*" – charges to which, of course, those of us with no idea what an ellipted-coordinate-clause-with-a-comma might look like can only comment, "Tsk"". Truss goes on to list the rules for the use of the comma, but concludes that, more than any other punctuation symbol, the comma should be used with "intelligent discretion".

(5.31) *Tom also is taking us to* ***an footie*** *match which shall be interesting.* (A.540)

10. **Other (5)**: There were one or two omissions of parts of set phrases such as *how* in *how much* and *such* in *such as*

5.2.3 Deletion errors

204 errors were corrected by deleting one word from a sentence. These deletions were examined and classified in terms of the sentential environment in which they occurred:

1. **Repeated words (62)**: The word which was deleted appeared elsewhere in the sentence, for example

   (5.32) *The bronzed Bermondsey belle has openly revealed she is a bottle blonde, and* ***roots*** *as the roots started to burst through, her barnet was in need of some urgent TLC.* (A.425)

   In almost half of these cases, the two occurrences were immediately adjacent, for example

   (5.33) *A large body of literature exists which criticizes the reliability of such* ***such*** *data.* (A.891)

2. **Repeated similar words (58)**: These kinds of deletions are similar to the repeated word deletions, the difference being that the word to be deleted occurs in the vicinity of another word either of the same syntactic category (54, e.g.(5.34)) or of similar spelling (4, e.g.(5.35)). Douglas and Dale (1992) call these *double syntactic function* errors and remark that they seem to be a "surprisingly frequent" occurrence for proofreaders.

(5.34) *Why **is do** they appear in this one section of the beach and not another?* (A.207)

(5.35) *I don't know but **it's its** crawling up your leg.* (A.827)

3. **Unnecessary words (84)**: These kind of deletions involve no repetition and are simply cases where an extraneous word appears in the sentence. Determiners (25), prepositions (20), verbs (19), full stops (8), pronouns (4) and conjunctions (4) make up the majority of these. An example involving a preposition is Sentence (5.36):

   (5.36) *But as soon as **of** the strangely creepy Jack started chatting up Kate the audience forgot they could see the strings being pulled.* (A.26)

5.2.4 Composite Errors

129 corrections involved a composite error, or an error which is corrected by applying some combination of the insert, delete and substitute operators. These errors have been classified according to the particular combination of correction operators they involve.

1. **Insertion and Deletion (30)**: Of these insertion and deletion composite errors, 18 consisted of errors which could be corrected by deleting a word from one position in a sentence and inserting it in another position, i.e. moving the word. An example is Sentence (5.37) where the word *pay* needs to be deleted from its position after the word *cheques*, and added to the position before *cheques*:

   (5.37) *As I went outside to join the others, it struck me that Sirius Recruitment must have known about the layoffs in order to deliver the monthly cheques **pay** several days early.* (A.460)

A further 7 of the 30 insertion and deletion composite errors involve the moving of a word from one position in the sentence to another, as well as another insertion or deletion. Sentence (5.38) is an example: this can be corrected by moving the determiner *the* from before the word *any* to after it, and inserting the preposition *of* immediately after *any* and before the moved *the*:

(5.38) *The proposed traffic lay-out has more in common with the Tallaght bypass than with* ***the*** *any current thinking on sustainable urban design.* (A.398)

The remaining 5 insertion and deletion composite errors do not involve the insertion and deletion of the same word. Sentence (5.39) is an example which can be corrected by deleting the two instances of the verb *call* and adding the copula verb *are* after the noun *judgements*.

(5.39) *According to Gleitman et al 1972 call call grammaticality judgements 'very derived and rather artificial psycholinguistic phenomena which develop late in language acquisition'.* (A.701)

2. **Insertion only (25)**: 22 out of 25 of the insertion only composite errors are cases where a sequence of words has been omitted from the sentence. An example is Sentence (5.40), which can be corrected by adding the word sequence *the fact that* at the beginning of the sentence:

   (5.40) *Big Brother had earlier reaffirmed its message to the housemates that reasons for nomination must be valid may have been a factor in the long duration of the Diary Room sessions.* (A.62)

An example of an insertion only composite error which can be corrected by inserting words at different points within the sentence is (5.41), which can be corrected by adding the word *the* before the word *describing*, and adding the word *of* after *describing*:

(5.41) *Phrase-structure syntax* ***allows describing the*** *syntactic structure of utterances in terms of the phrases which constitute them, using a hierarchical and recursive set of concepts.* (A.595)

3. **Deletion only (17)**: The deletion only composite errors are corrected by deleting a sequence of words from the ill-formed sentence. Of these, 9 involve deleting a sequence which appears elsewhere in the sentence, e.g:

(5.42) *The body of this for loop checks if value i.e. the number of a particular pronoun, is equal to* ***the number of the number of*** *a verb in the verb array at position i.* (A.90)

Sentence (5.43) is an example which can be corrected by deleting a word sequence which is *not* repeated elsewhere in the sentence. The word sequence to be deleted is *it was*:

(5.43) *They reassured them that* ***it was*** *they came back because they truly wanted to.* (A.742)

4. **Insertion and Substitution (17)**: There are no obvious patterns within the group of composite errors which are corrected by carrying out a combination of insertions and substitutions. In 12 of the 17 errors of this type, a verb is changed in some way. An example is Sentence (5.44) which can be corrected by adding the word *be* af-

ter *possible to* and changing the form of the verb *analyse* to the past participle *analysed*:

(5.44) *On the basis of (i) these annotated examples and (ii) written documentation, new utterances should be possible **to analyse** with a high degree of consistency.* (A.406)

5. **Insertion and Deletion and Substitution (15)**: A composite error of this kind is corrected by applying some combination of the three correction operators. 11 of these involve an error which is corrected by changing the form of a verb to past participle, by moving this verb, by moving or adding an infinitival marker and by adding a passive auxiliary verb. Sentence (5.45) is an example:

   (5.45) *Validating parsers allow **to control** the input and the output of XML conversion routines.* (A.469)

   In (5.45), the sequence *to control* can be deleted from its position before its object noun phrase and added to the position after the noun phrase. The form of the verb *control* can be changed from infinitival to past participle, and the infinitival auxiliary *be* can be added between *to* and *controlled*.

6. **Substitution only (14)**: 6 of the 14 substitution only composite errors involve an agreement error which is corrected by changing an agreement value on more than one word in the sentence. An example is Sentence (5.46) which can be corrected by changing the singular pronoun *it* to its plural counterpart *they*, and by changing the singular verb *is* to its plural counterpart *are*:

(5.46) *One of the most important arguments for declarative approaches to NLP is that **it is** monotonic.* (A.257)

A further 4 of these substitution only composite errors can by correcting by replacing an infinitival marker with the preposition *of*, and by replacing an infinitival verb form with a present participle. Sentence (5.47) is an example:

(5.47) *But no one of them is capable **to deal** with robustness as a whole.* (A.18)

7. **Deletion and Substitution (11)**: In 7 out of the 11 composite errors which are corrected by a combination of deletion and substitution, a verb is substituted, either for a verb or another part of speech. An example is Sentence (5.48) which is corrected by deleting an unnecessary infinitival marker and changing the form of the verb *remark* from infinitival to present participle:

   (5.48) *It is also worth **to remark** that robustness is a necessary prerequisite in any system designed for real-life applications.* (A.850)

Fig. 5.2 summarizes the information presented within this section.

5.3 A Parsing Experiment

In this section, an approach to parsing ungrammatical language is described and evaluated. A general overview of the approach is given in Section 5.3.1 and it is compared to some of the error-handling solutions described in Chapter 4. Using the error analysis results presented in Section 5.2 as a basis, an

Correction	Freq	Correction	Freq
Substitution	**590**	**Insertion**	**299**
Context-Sensitive Spelling	188	Determiners	81
Number/Person	114	Verbs	67
Verb Form	65	Prepositions	60
Homophone Confusion	53	Pronouns	29
Same Root/Different POS	41	Nouns	20
Case (All Possessive Errors)	37	Infinitival *to*	19
Prepositions	33	Punctuation	11
Indefinite Articles	12	Conjunctions	5
Auxiliary Verbs	11	Adjectives	2
Synonymous Verbs	6	Other	5
Infinitival *to* and Prepositions	3	**Composite Corrections**	**129**
Same Root,Same POS,Different Meaning	2	Insertion and Deletion	30
Conjunctions	2	Insertion Only	25
Full Stop/Complementizer	2	Deletion Only	17
Pronoun/Noun	1	Insertion and Substitution	17
Adjectival Form	1	Insertion and Deletion and Substitution	15
Other	19	Substitution Only	14
Deletion	**204**	Deletion and Substitution	11
Unnecessary Words	84		
Repeated Words	62		
Repeated Similar Words	58		

Figure 5.2: Summary of Error Analysis Results

error grammar is derived from a conventional grammar. The process of deriving the error grammar is described in Section 5.3.2. The robust chart parsing algorithm which parses ill-formed sentences using the error grammar is described in Section 5.3.3. Finally, an experiment is carried out using this algorithm and the error grammar, and the results of this are presented in Section 5.3.4. Suggestions for how these results could be improved upon are also provided.

5.3.1 General Procedure

My approach is to use an empirically grounded model of ungrammaticality or an *error grammar* as the basis of a chart parsing algorithm which is designed to run after an initial bottom-up chart parsing process has failed to deliver a parse for a sentence. The error grammar contains a sequence of *error rules*, where each error rule is derived from a conventional rule in such a way that the relationship between them reflects an error that might be expected to occur and the correction that would be applied to make it well-formed. The recovery chart parsing process uses the complete and incomplete edges found during the normal parsing phase together with edges arising from an error rule in its attempt to arrive at a complete parse. The approach incorporates aspects of error anticipation, constraint relaxation, minimum edit distance parsing and parse fitting. I will describe each of these aspects in turn, making reference, where appropriate, to particular parsers discussed in Chapter 4:

1. Error Anticipation: The approach is essentially one of error anticipation because it uses error rules which explicitly model ungrammatical language. In Section 4.2.2, I made the point that probabilistic parsers, such as those described in Charniak (2000), are by their very nature robust since they are exposed in their training phase to a large body of empirical data where no explicit distinction is made between the grammatical and the ungrammatical. According to Sampson (2001b), this is a good thing since a clear line cannot always be drawn between the two (see also Section 2.2.3). However, to re-iterate the point made in Section 2.4.1, the fact that language errors do undoubtedly occur means that the concept of ungrammaticality cannot be dismissed. The use of an error grammar has the advantage of providing a linguistic model of ungrammaticality, informed by knowledge of constructions and not just likelihood-analysis based on frequencies alone – this means that ill-formed sentences can be diagnosed as such, instead of being viewed merely as sentences occurring with a low frequency. The ability to detect that an error has occurred can sometimes mean the difference between finding a parse for a sentence and finding the *right* parse.

The use of an error grammar is, however, compatible with a probabilistic view of language processing and I do not in any way mean to suggest that the work presented in this chapter should be viewed as an alternative to a broad-coverage probabilistic parsing system but rather could be used to provide such a system with an explicit model of ungrammaticality.[5] The error rules described in this chapter are derived from non-probabilistic grammar rules but there is no reason why they could not be derived from probabilistic grammar rules. The probability of an error rule could then be viewed as a function of the probability of the error it represents (derived from a corpus of ungrammatical sentences such as a larger version of *corpus1*), and the probability of the normal rule from which it is derived, thus allowing one robust parse to be preferred over another. I take up this discussion again in

[5]Indeed, the work presented in Chapter 6 will make it clear that broad-coverage probabilistic parsers are doing quite a good and robust job even without an explicit model of ungrammaticality.

Section 5.3.4.

Another advantage of using an error grammar is that it is allows all error types, substitution, insertion, deletion and composite errors, to be handled within a uniform framework. Some other approaches to the problem of parsing erroneous text are more suited to one error type than another. The constraint relaxation approach described by Douglas and Dale (1992), for example, is designed to handle substitution errors where the words involved in the substitution clash with respect to some grammatical property, e.g. case or verb form. Jensen *et al.*'s (1983) parse-fitting approach is designed to run after a constraint relaxation error-handling strategy has already been applied to substitution errors, and is more suited to errors involving missing or extra words.

Schneider and McCoy (1998) and Bender *et al.* (2004) also use error rules or mal-rules to parse ungrammatical sentences. However, their attention is restricted to second language errors and, in Schneider and McCoys' case, they explicitly attempt to model the second language learning process. The class of errors handled by the approach outlined in this chapter is more general – this class includes any type of syntactic error, be it a language learning error or a performance error.

2. Constraint Relaxation: The error rules used in this parser are derived from conventional grammar rules, and in this respect, the approach outlined here is similar to constraint relaxation approaches where there is a close relationship between the analyses produced for well-formed input and those produced for ill-formed input. The advantage of this is that corrections of ungrammatical sentences can be easily proposed, but the disadvantage is that the parser's coverage with respect to ungrammatical sentences will be dependent on its coverage with respect to grammatical sentences.

The error rules described in this chapter are quite similar to the meta-rules described by Weischedel and Sondheimer (1983) because they are both derived from the conventional grammar rules. However, there are differences between the two approaches. Weischedel and Sondheimer's (1983) meta-rules are integrated into the main parsing process, whereas in this approach, the error rules are only applied after a normal parse has failed. The integration of the meta-rules into the main parsing phase is a consequence of the fact that the parser is a top-down Augmented Transition Network (ATN) rather than a chart parser. An ATN is simultaneously both a parser and a grammar. This means that there is less control over when the meta-rules are invoked, since they can be applied even before sections of the input have been encountered. Another consequence of the ATN-based approach is that the meta-rules are more procedural than declarative — it is difficult to view them in isolation from the actual parsing algorithm and so they cannot be used as a model of ungrammaticality in the same way that the error rules described in this chapter can.

3. Minimum Edit Distance: One of the principles of minimum edit distance parsing is that corrections should be preferred for an ungrammatical sentence which involve minimal change to the ungrammatical sentence. This principle is emphasised in my parsing approach because I focus on ungrammatical sentences which have an edit distance of one from their grammatical counterpart, the majority of sentences according to the data in *corpus1* and *corpus2*. By imposing this restriction, the search space in the parse chart is constrained because interaction between error rules can be eliminated. This will mean that nonsensical parses of ungrammatical sentences should also be kept to a minimum.

The error rules which are used to describe deletion and insertion error rules are very general. In fact, they do not make use of the subclassifications determined in the analysis of these kinds of errors. Thus, the handling of deletions and insertions is similar to the way they are treated in the minimum edit distance approach described by Lee *et al.* (1995) in which insertion and deletions of any non-terminal symbols can be postulated at any point in the input string. I return to this issue in Section 5.3.2, when I describe how the error grammar was written.

4. Parse Fitting: Another difference between the error rule approach described in this chapter and the mal-rule approach employed by Schneider and McCoy (1998) is that the approach described here is carried out in two stages. The advantages of using a two-stage process are that grammatical sentences will never be flagged as ungrammatical, and there will be no deterioration in efficiency during the normal parsing phase. The use of a two stage parsing process makes this approach similar to the parse fitting approaches discussed in Section 4.3.4.

In this section, I have given a brief overview of the parser described in this chapter: it employs a two-stage parsing process, and during the second stage uses error rules introduced in a controlled way into the parse chart to find parses for ungrammatical sentences. The error rules are derived from conventional grammar rules based on an analysis of the differences between *corpus1* sentences and *corpus2* sentences (see Section 5.2). This approach was compared to some of the parsing approaches discussed in Chapter 4.

5.3.2 Generating the Error Grammar

In this section, I describe how an error grammar was derived from a conventional grammar.

A grammar of well-formed language is used in order to generate a grammar of ill-formed. To test this error grammar approach, a context-free phrase structure grammar containing 1,705 rules was used. The following linguistic phenomena are covered by the grammar: several verb subcategorization types, auxiliary and modal verbs, pre- and post-nominal modification by adjectives, prepositional phrases and relative clauses, verb modification by adverbs and prepositional phrases, declarative and interrogative sentences, and co-ordination. The non-terminal symbols of the grammar are Prolog terms with argument slots for appropriate features. The parser which operates on this grammar is implemented in Prolog, and the rules (the original well-formed rules and the error rules) are compiled into Prolog predicates. For each error correction type, the process of generating error rules for this type is described.

Substitution Error Rules

For each rule in the grammar which expands a pre-terminal symbol (i.e. a part-of-speech category), an error rule is generated for all the words in the lexicon which are similar to the word on the right-hand side of this rule, according to the similarity criteria described in 5.2.1. For the rule (5.49) the error rules in (5.50) are generated:

(5.49) `verb(fin(sing,third),cop) --> [is].`

(5.50)
```
verb(fin(sing,third),cop) spellop [in].
verb(fin(sing,third),cop) spellop [it].
verb(fin(sing,third),cop) spellop [its].
verb(fin(sing,third),cop) spellop [if].
verb(fin(sing,third),cop) spellop [as].
verb(fin(sing,third),cop) agreeop [are].
```

```
verb(fin(sing,third),cop) agreeop [am].
verb(fin(sing,third),cop) vformop [be].
verb(fin(sing,third),cop) vformop [being].
verb(fin(sing,third),cop) vformop [been].
```

To distinguish error rules from conventional grammar rules, the $-->$ connective is substituted by a connective which describes the error, e.g. the connective `spellop` refers to errors which result when the correct and incorrect word are similar in spelling, the connective `agreeop` refers to errors where the correct and incorrect word have conflicting values for an agreement feature, and the connective `vformop` refers to errors where the two words (or verbs) have conflicting verb form values. In all other respects these connectives mean the same thing as the conventional rule connective $-->$. The connective in the well-formed rules is converted into a binary Prolog predicate where the first argument is the left-hand side of the rule and the second argument is the right-hand side represented as a list. The error rule connectives are converted into predicates with similar argument structure. (5.51) is the predicate for rule (5.49) and (5.52) is the predicate for the first error rule in (5.50):

(5.51)`rule(verb(fin(sing,third),cop),[[is]]).`

(5.52)`spell_error_rule(verb(fin(sing,third),cop),[[in]]).`

1,371 rules in the original grammar expand pre-terminal symbols. 2,869 substitution error rules were generated from this set. They were generated manually in this case, but rules of this sort could be generated using a process which would automatically detect similar pairs of words in the lexicon, according to the similarity classifications presented in Section 5.2.1.

Insertion Error Rules

For each rule in the grammar (excluding rules expanding pre-terminal symbols), an error rule is generated which has the same right-hand side as the original rule except that a symbol on the right-hand side is removed.[6] So, for example, for the rule (5.53) the error rules in (5.54) are generated:

(5.53)`vp(Form) --> verb(Form,trans), np(Num,Per).`

(5.54)`vp(Form) missingop np(Num,Per).`
`vp(Form) missingop verb(Form,trans).`

An insertion error rule is not generated if its right-hand side is the same as the right-hand side of a conventional rule, e.g. given the rule (5.55) the error rule (5.56) is not generated since this already exists in the grammar as a conventional rule.

(5.55)`np(Num,Per) --> det(Num,Per), adj, n(Num,Per).`

(5.56)`np(Num,Per) missingop  det(Num,Per), n(Num,Per).`

There is also no need to generate insertion error rules with an empty right-hand side.
334 rules in the original grammar do not expand pre-terminal categories. From this set, 512 insertion error rules were generated automatically.

Deletion Error Rules

For each rule in the grammar (excluding rules expanding pre-terminal symbols), an error rule is generated which has the same right-hand side as the

[6]These insertion error rules could also provide a treatment for the grammatical phenomenon of ellipsis. This would make this parser more like some of the parsers discussed in Chapter 4, where the problem of parsing ungrammatical input is conflated with the problem of parsing constructions which are not described by the core grammar system.

original rule except that a symbol is added after a pre-terminal symbol. This symbol must be capable of matching with any non-terminal symbol in the grammar. Given, for example, the rule (5.57) the error rule (5.58) is generated, where `nonterm` is compatible with every nonterminal category in the grammar.

(5.57)`vp(Form) --> verb(Form,trans), np(Num,Per).`

(5.58)`vp(Form) extraop verb(Form,trans), nonterm, np(Num,Per).`

The rule (5.59) results in the two deletion error rules in (5.60):

(5.59)`np(Num,Per) --> det(Num,Per), n(Num,Per).`

(5.60)`np(Num,Per) extraop det(Num,Per), nonterm, n(Num,Per).`
`np(Num,Per) extraop det(Num,Per), n(Num,Per), nonterm.`

The error rules will describe any ill-formed sentence where an extraneous word or a sequence of extraneous words representing a particular linguistic catgory appear *after* any word in the well-formed sentence. To allow for extraneous words at the beginning of a sentence, error rules can be generated from rules which expand sentential categories. So, for example, given the classic sentential rule (5.61), the error rule in (5.62) will be generated:

(5.61)`s --> np(Num,Per), vp(fin(Num,Per)).`

(5.62)`s extraop nonterm, np(Num,Per),vp(fin(Num,Per)).`

563 deletion error rules were generated automatically from a set of 334 well-formed rules.

Error Grammar Coverage

The deletion and insertion error rules are general enough to describe any type of error whose correction involves the deletion of a word from a sentence or the insertion of a word into a sentence. As such, these error rules do not reflect the detailed analysis carried out upon these kinds of correction operators (described in Sections 5.2.2 and 5.2.3) but are a generalization of them. In its very general treatment of deletion and insertion errors, this approach is similar to the minimum edit distances approaches discussed in Section 4.3.3, e.g. (Lee *et al.* 1995).

The insertion and deletion error rules will also handle a subset of the insertion only and deletion only composite errors (see Section 5.2.4), namely, those composite errors which involve an extraneous or missing sequence of words which have been recognized as a category during the normal parsing phase (see, for example, Sentences (5.40) and (5.43)). No other composite errors are handled by the error grammar, as it stands. This can be justified by noting that almost 90% of the corrections made on sentences in *corpus1* involved the application of just *one* correction operator.[7] In fact, 92.5% of the sentences in *corpus1* have at least one correction which involves just one operator. Although the error grammar does not at present cover most composite errors, they are not incompatible with this approach, and handling them is a goal for further research (see Section 5.3.4).

The substitution error rules are a more direct reflection of the analysis carried out on this operator (see Section 5.2.1) since they are defined on

[7]This mirrors findings in the field of spelling correction: Ingels (1996, p.7-8) reports on studies which have found that approximately 80% of spelling errors can be corrected with an edit-distance of one. I found a similar proportion in the 188 context-sensitive spelling errors in *corpus1* (see Section 5.2.1).

the basis of the similarities, noticed during the error analysis, between the substituting and substituted word. The only exception is the last category where the confusion between two words could not be explained in terms of the similarity between them. I have made the decision to treat such corrections as composite error corrections involving an insertion and deletion operation.

To summarize, I expect that the error grammar, once it has been combined with a suitable parsing algorithm (which I prove in Section 5.3.3), will be able to parse all errors which are correctable by inserting or deleting one word or a sequences of words representing a syntactic category, and most errors which are correctable by substituting one word for another.

5.3.3 Robust Recovery Algorithm

In this section the robust recovery algorithm is described. This algorithm takes as input a sentence which has failed to parse using an active bottom-up chart parsing algorithm, along with the chart edges which were created while trying to find a complete parse. A bottom-up parsing strategy is used during the normal parsing phase because a bottom-up parser is driven by the words in the input sentence and will be guaranteed to find all partial parses in the ungrammatical sentence.[8]

The bottom-up active chart parsing algorithm used by the parser is shown in Fig. 5.3. The chart edge notation

```
CM --> CM1...CMN* [y,z]
C --> C1...*CM...CN [x,y]
```

[8]By contrast, a standard top-down parser will run out of steam leaving sections of the input sentence untouched. However, Van Noord (1996) describes how a head-corner parser which uses a combination of top-down and bottom-up processing can make use of underspecification to find partial parses throughout the input string.

is explained as follows: a category `CM` consisting of the sequence of categories `CM1...CMN` has been found from positions `y` to `z` in the input string. The `*` symbol is used to separate categories that have already been found by the parser from those which are still to be found. In the `CM` edge the `*` appears after the category sequence `CM1...CMN` which means that this edge has satisfied its expectations and is *inactive*. The second edge is *active* because it is still looking for the categories `CM...CN`. It has already found the categories `C1..CM-1` between positions `x` and `y` in the input string. When it has fulfilled its outstanding expectations, a category `C` will have been found. The chart parser packs inactive edges of the same category to prevent local ambiguites from slowing down the parsing process. For all inactive edges of the same category but with different daughter constituents which span the same section of the input string, only one 'representative' edge is added to the chart, and the equivalent edges are unpacked when the parses are enumerated.

The robust recovery parsing algorithm is given in Fig 5.4. The `errorop` connective corresponds to any connective which appears in an error rule, e.g. `missingop` or `spellop` (see Section 5.3.2). All the chart edges built during the normal parsing phase, active and inactive, are available to the recovery algorithm. The recovery algorithm works by repeatedly adding new edges corresponding to error rules to the chart and reinvoking the bottom-up chart parsing algorithm to see if a complete parse can be found. After an error rule has been tried (whether successfully or unsuccessfully), the slate is wiped clean for the next error rule. This is done by deleting from the chart all "error edges", or edges which have been created as a result of this error rule. This ensures that there is no interaction between error rules.

Substitution error rules are chosen on the basis of the actual words in the input sentence, in a process similar to the chart initialization process.

Insertion and deletion error rules are chosen in a selection process similar to the bottom-up rule. A forward scan of the chart is carried out to ensure that only error rules whose right-hand side are already contained in the chart are added to the chart agenda. They are added as inactive edges because their right-hand side has been found. This forward scan procedure can also be added to the bottom-up rule of chart parsing. The fundamental rule of chart parsing still combines active and inactive edges, but these edges can be edges found during the first phase or error edges resulting from an error rule.

Consider Sentence (5.63) which is a shortened version of one of the sentences in the test data (the data which is used to test this algorithm is described in the next section). (5.64) is the appropriate correction given the context in which it was written (a restaurant review).

(5.63)*Everything was* ***is*** *perfect amounts.* (C.1)

(5.64)*Everything was* ***in*** *perfect amounts.*

The diagram in Fig. 5.5 is an example of what the parse chart might look like after the normal parsing phase. Edges not directly relevant to the example are omitted in order to make the diagram easier to read. No inactive edge spans the whole sentence so the recovery phase is needed. I will discuss two examples, one where a substitution error rule is proposed on the basis of a word in the input sentence, and another where a deletion error rule is proposed on the basis of an inactive edge found during the first phase.

For each word in the input string, a search is made in the lexicon for error rules which have this word on their right-hand side. The word *is* will be on the right-hand side of error rule (5.65) due to its similarity in spelling to the preposition *in*. The inactive edge (5.66) will be added to the chart at position [2,3]. This will trigger the creation of the error edges in (5.67), resulting in a `bottom-up parse:`

initialize the chart

while there are more edges on agenda **do**

E = edge taken from top of agenda

if E does not span the chart **then**

use E to find more edges using fundamental rule

use E to find more edges using bottom-up rule

add E to the chart

enumerate all parses

`chart initialization:`

for each word $W[x, y]$ in the input string **do**

for each rule $C - - > W$ in the grammar **do**

add an inactive edge $C - - > W * [x, y]$ to agenda

`fundamental rule:`

if there is an inactive edge $CM - - > CM1...CMN * [y, z]$ **then**

for each active edge $C - - > C1... * CM...CN[x, y]$ **do**

add an edge $C - - > C1...CM * ...CN[x, z]$ to agenda

if there is an active edge $C - - > C1... * CM...CN[x, y]$ **then**

for each inactive edge $CM - - > CM1...CMN * [y, z]$ **do**

add an edge $C - - > C1...CM * ...CN[x, z]$ to agenda

`bottom-up rule:`

if there is an inactive edge $C1 - - > C11...C1N * [x, y]$ **then**

for each rule in the grammar $C - - > C1...CN$ **do**

add an active edge $C - - > *C1...CN[x, x]$ to agenda

Figure 5.3: Bottom-up Active Chart Parsing Algorithm

```
top-level:
```

for all appropriate error rules (see below) **do**

reinvoke the bottom-up chart parser (see Fig. 5.3 and new bottom-up rule below)

remove all edges from chart which have arisen from this error rule

```
find appropriate substitution error rules:
```

for all words $W[x, y]$ in the input string **do**

for each error rule C *errorop* W in the error grammar **do**

add inactive edge C *errorop* $W * [x, y]$ to agenda

```
find appropriate insertion and deletion error
    rules:
```

for all inactive edges $C1 - - > C11..C1N * [x, y]$ found during the normal parsing phase **do**

for each error rule C *errorop* $C1...CN$ **do**

for each sequence of inactive edges between y and some z for the categories $C2..CN$ **do**

Add an inactive edge C *errorop* $C1...CN * [x, z]$ to agenda

```
new bottom-up rule:
```

if there is an inactive edge $C1$ *errorop* $C11...C1N * [x, y]$ **then**

for each rule $C - - > C1...CN$ **do**

for each sequence of inactive edges between y and some z for the categories $C2...CN$ **do**

add an inactive edge C *errorop* $C1...CN * [x, z]$ to agenda

Figure 5.4: Recovery Algorithm

complete edge which spans the input. This situation is illustrated in Fig. 5.6. Dashed lines are used to indicate edges which are added during the recovery phase. The parse is reported and the edges (5.66) and (5.67) are removed from the chart.

(5.65)`prep spellop [is].`

(5.66)`prep spellop [is] * [2,3]`

(5.67)`pp spellop prep np * [2,5]`
```
vp spellop verb pp * [1,5]
s  spellop np vp * [0,5]
```

For each inactive edge in the chart, a search is made for an error rule which has this edge's category as the first category on its right-hand side, and where all subsequent categories on its right-hand side are present in the correct position in the chart. Given the inactive edge (5.68), the error rule in (5.69) fulfils these criteria, and the error edge (5.70) is added to the chart. This edge causes the error edge in (5.71) to be added to the chart. This situation is illustrated in Fig. 5.7. The algorithm continues to explore possible error rules until all possibilities have been exhausted.

(5.68)`verb --> [was]* [1,2]`

(5.69)`vp extraop verb verb np`

(5.70)`vp extraop  verb verb np * [1,5]`

(5.71)`s extraop np vp * [0,5]`

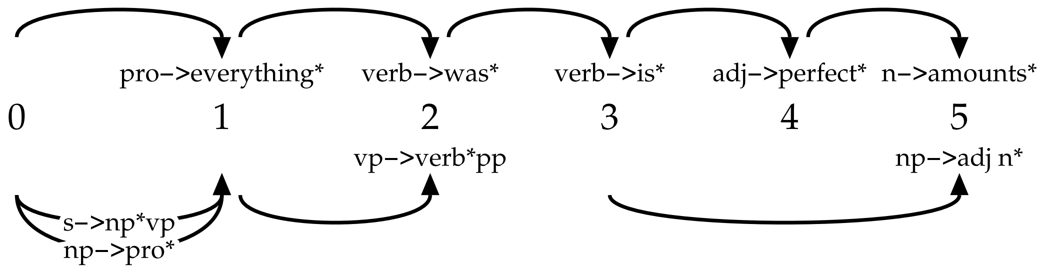


Figure 5.5: Chart after Initial Parse

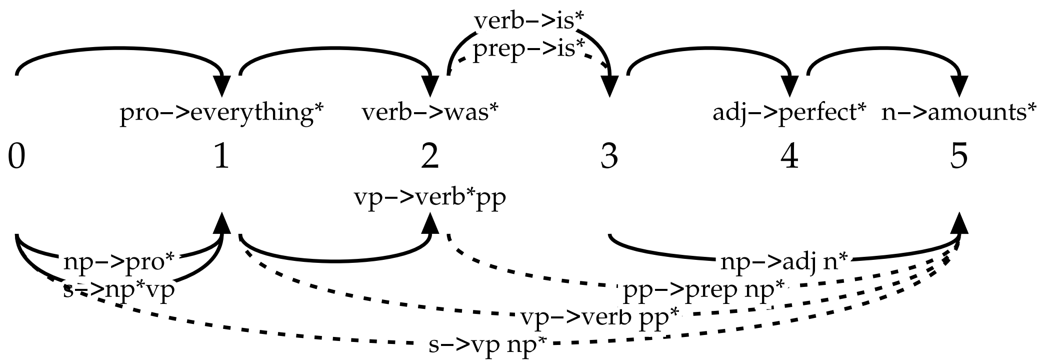


Figure 5.6: Recovery Phase: Substitution Error Rule

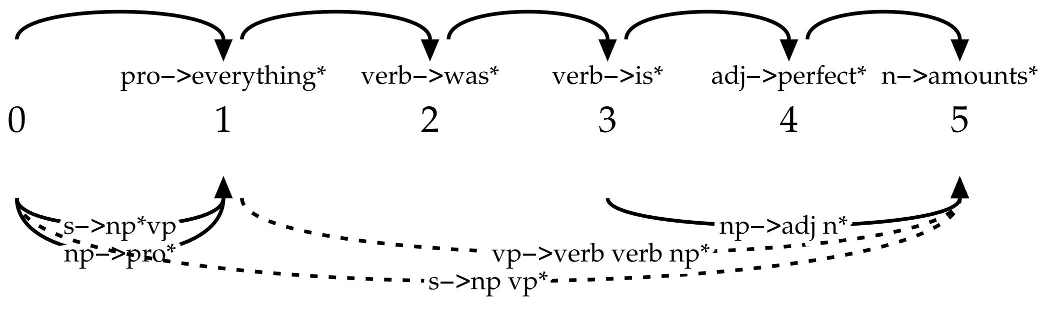


Figure 5.7: Recovery Phase: Deletion Error Rule

5.3.4 Evaluation

The robust recovery algorithm described in Fig. 5.4 was applied to forty-four ungrammatical sentences which are not contained in *corpus1*. *Corpus1* sentences could not be used in testing the parsing approach because the error grammar is partly based on the analysis of these sentences. The test sentences were collected after the corpus was completed, but they were collected using the method described in Chapter 3 which was used to collect the corpus sentences: the ungrammatical sentences were noticed while reading, and they were corrected using the context in which they were found as a basis. The longest sentence in this test set has fifty four words, the shortest has two words, and the average sentence length is eighteen words. The test sentences are contained in Appendix C. Before the ungrammatical sentences were tested, their corrected versions were tested using the conventional grammar and the bottom-up chart parsing algorithm described in Fig. 5.3. Any necessary expansions were made to the grammar to ensure that the grammatical sentences received appropriate parses. The error grammar is derived from the well-formed grammar, and its ability to parse ungrammatical sentences will depend on the ability of the well-formed grammar to parse the corresponding grammatical sentences.

Accuracy

The recovery process was deemed to have worked if the corrections proposed for an ill-formed sentence matched in structure the corrected versions of the sentence. This was the case if an error rule used during the recovery phase parsing of the ill-formed sentence was derived from a conventional rule which was used at the same point in the parse of the corrected sentence. The robust recovery procedure achieved an accuracy rate of almost 80%. The nine failed attempts can be explained as follows:

1. More than one error in the sentence: two sentences in the test data contained two separate errors, e.g. Sentence (5.72) which contains a substitution error involving *a/an* along with either a deletion error (superfluous preposition *for*) or an insertion error (a missing pronoun *they* before *insisted*).

(5.72) *Following on from* ***a*** *emotional reunion* ***for*** *the lads insisted they would not get back together as a band.* (C.36)

The recovery algorithm (see Fig. 5.4) ensures that only one error rule is ever considered during any one parse attempt, with the result that this algorithm will not handle sentences containing two or more errors. A further step in this research is to modify the algorithm so that it can consider more than one error rule under certain controlled circumstances.[9]

2. Composite errors: One sentence in the test data contained a composite error:

(5.73) *In this paper an evaluation method is presented which is similar to the above yet* ***is which*** *specifically designed to measure a parser's accuracy in the face of input which is ungrammatical.* (C.10)

Sentence (5.73) can be corrected by exchanging *is* and *which* (a deletion and insertion composite error). Sentences containing a composite error will not be dealt with under this approach for the same reason that sentences containing more than one error will not: only one error rule can be considered at any one time. To deal with composite errors, the recovery algorithm will

[9] 67 of the 924 sentences in *corpus1* contain more than one error.

need to be modified so that it has the potential to use more than one error rule at a time, or error rules could be constructed which describe certain types of regularly occurring composite errors (see Section 5.2.4) .

3. Covert errors: six sentences in the corpus were not recognized as ill-formed by the normal parser. Sentence (5.74) is an example:

(5.74) *The dependency-based method allows us to evaluate not only the overall performance of the parser, but also its performance with respect* ***different*** *grammatical relationships and phenomena.* (C.39)

The preposition *to* has been left out of Sentence (5.74), and although this may be obvious to the human parsing mechanism, it is not obvious to the computer parser described in this chapter. Given its grammar, the parser finds a full parse in which *respect, grammatical relationships and phenomena* is treated as a co-ordinated noun phrase. Sentence (5.75) is another example of an ill-formed sentence for which an implausible reading can be found. The ill-formed verb *view* should be conjoined to the present participle verb *following*, but in its erroneous form, it can be conjoined to the infinitival *be*:

(5.75) *Successful exploitation requires that a user can be tricked into following a link or* ***view*** *a malicious HTML document.* (C.44)

Examples from *corpus1* of this kind of sentence were discussed in Chapter 3. If the recovery algorithm was applied to these sentences the appropriate robust parse would be suggested, but since they are not ungrammatical according to the test grammar, the recovery process will never be applied. This kind of case will remain a problem until the grammar is extended so that it can take into account sophisticated semantic and contextual information

	Average Parse No.	**Median Sentence Parse No.**
Corrected	54	3
Ill-Formed	362	11
	Average Cycle No.	**Median Sentence Cycle No**
Corrected	1444	1131
Ill-formed	6094	4220

Figure 5.8: Performance Results

(in itself a formidable task). If the conventional grammar is a probabilistic one, it might be possible to calculate a probabilistic threshold which could be used to determine if an error rule should be fired.

Efficiency

Fig. 5.8 compares the successfully parsed ungrammatical sentences in the test data to their corrected counterparts, in terms of average number of parses and average number of parse cycles. The notion of a parse cycle was proposed by Mellish (1989) as an implementation-independent measure of the work done by a chart parser. In one parse cycle, an edge is taken from the agenda, added to the chart and used (via the fundamental and bottom-up rules of chart parsing) to propose more edges. The average parse figures are slightly skewed by a couple of very long sentences in the test data containing multiple prepositional phrase attachment and co-ordination ambiguities. For this reason, the sentence with the median parse number in the corrected test set was determined - its parse and cycle numbers are shown, along with the parse and cycle numbers for the corresponding ungrammatical sentence.

The increase in number of parses for the set of ungrammatical sentences is expected, since the robust recovery algorithm is essentially proposing ways of correcting an ungrammatical sentence, and each correction will have a certain number of parses associated with it. Consider, for example, the ungrammatical (5.76) which is in the test data and which originally appeared in the documentation for a piece of software:

(5.76) *Please note, some basic knowledge of* ***computer*** *is necessary.* (C.17)

The corrected version of Sentence (5.76) is Sentence (5.77) and this receives two parses.

(5.77) *Please note, some basic knowledge of* ***computers*** *is necessary.*

The robust parsing algorithm finds these two parses, along with another eight associated with the following grammatical sentences of varying degrees of plausibility:

(5.78) *Please note, some basic knowledge of* ***a/the/this*** *computer is necessary.* (an insertion error involving a singular determiner)

(5.79) *Please note, some basic knowledge of computer* ***science/hardware/graphics/processing, etc*** *is necessary.* (an insertion error involving a noun or a present participle verb)

(5.80) *Please note, some basic knowlege* ***per*** *computer is necessary.* (a substitution error, involving a confusion between the prepositions *per* and *of*)

The parse cycle number for an ungrammatical sentence includes the parse cycle number for the original parse of the sentence (when no parses are found) plus the parse cycle number for the recovery parse. Moving from the grammatical sentences to the ungrammatical sentences, there is an approximately four-fold increase in the number of parse cycles. Kato (1994) reports comparable results for his two-stage parser which employs bottom-up strategies during the first phase and top-down strategies during the second (his figures are an improvement on those reported by Mellish (1989)). However, the approaches employed by Mellish (1989) and Kato (1994) (see Section 4.3.4) are tested on short invented sentences of up to thirteen words. The approach described here is tested on longer naturalistic sentences. It must be noted, however, that the parse cycle figures only reflect the parsing process and do not reflect the time spent searching through the set of error rules to find appropriate ones to add to the parse chart. The length of time spent doing this will depend on the size of the error grammar[10], the size of the input string and the number of inactive edges in the chart after the first recovery phase has finished. The latter two can be measured by extending the notion of parse cycle to include the process of taking a word in the input string or an inactive edge in the chart and using it to propose appropriate error rules. With this extended notion of a parse cycle, the average parse cycle number increases from 6094 cycles to 6208 cycles.

The parse cycle figure could be reduced by limiting the number of error rules that are investigated. There are two ways in which this could be done: a threshold could be introduced after which no more error rules are explored, *or*, an attempt could be made to find out where in the ungrammatical sentence the error has occurred, so that error rules are only introduced from this portion of the chart. I will discuss both solutions in turn. The best solution will involve a combination of both strategies.

The problem with introducing an error rule threshold is that there is a

[10] In the grammar used in this experiment, the ratio of error rules to conventional rules is 2.3:1.

possibility that valuable parses will be lost. In order to improve the accuracy associated with stopping at a certain point during the parsing of an ungrammatical sentence, it will be necessary to ensure that the edges corresponding to error rules are added to the chart in a way that mirrors error frequency (Magerman and Weir 1992). At the moment the error rules are ordered such that substitution corrections are proposed before insertion and deletion corrections since these are the most frequent in the corpus (see Fig 5.1). The individual substitution error rules are also ordered in terms of the frequencies given in Section 5.2.1. A more sophisticated ordering of the error rules which takes into account some of the error analysis results discussed in Sections 5.2.2 and 5.2.3 and not included in the error grammar, is a goal for further research. An example is an ordering which would prefer deletion error rules where the extra category on the right-hand side is of the same category as the pre-terminal symbol immediately to its left, since these errors show up frequently in the corpus data.

Since edges corresponding to error rules are added to the agenda based on the words in the input sentence and the inactive edges found during the initial parsing phase, another way to reduce the number of parse cycles is to reduce the number of inactive edges and/or words in the sentence which are used to choose the appropriate error rules or to order these words and edges appropriately. This could be done by guessing where in the input sentence the error is most likely to have occurred. One possibility, following Mellish (1989), is to explore the use of top-down parsing in the second phase to pinpoint the location of an error. This will have to be done in such a way that the parsing involved in guessing the location of an error does not itself take up too much time. Otherwise the benefit of using such a technique is lost. Heuristic strategies which select the words and inactive edges based on their category could also be tested. Over half of the errors found in *corpus1* involve the insertion, deletion or substitution of a verb, preposition or a determiner. This observation could be used to prioritize certain words in the input string over others. Another possibility is to exploit probabilistic techniques to determine the location of an error within a sentence, e.g. an error is perhaps more likely to have occurred in a portion of the chart where the partial parses have relatively low frequency ratings. This is a topic for further investigation. The figures in Fig. 5.8 represent an encouraging non-optimized scenario, and I am confident that they can be improved upon without a significant deterioration in accuracy.

5.4 Conclusions

In this chapter, the differences between the sentences in *corpus1* and *corpus2* were analysed. An error typology was presented which classifies errors on the basis of the correction operator which was applied in order to eliminate the error: substitution, insertion, deletion and composite. The errors in each group were counted and examined. On the basis of this examination, further groups were identified and the errors within these groups counted.

The second part of this chapter presented a robust parsing technique, based on the concept of an error grammar, which provides a set of error rules describing ungrammatical sentences in the same way that a conventional grammar provides rules which describe grammatical sentences. The rules in an error grammar are distinguished by means of a rule connective from rules in the normal grammar but they can be parsed using the same parsing techniques. The error rules are derived from normal grammar rules on the basis of actual error data, while remaining general enough to ensure good coverage. The recovery algorithm described in this chapter is designed to use one error rule together with the results produced by a bottom-up chart parser in order to find a parse for an ungrammatical sentence. The decision to use just one error rule at a time has also been justified by the data in *corpus1* and *corpus2*. The recovery parsing algorithm has achieved promising results in its first evaluation, results which can be improved when the suggestions for further work have been carried out.

Chapter 6

Evaluating the Parsing of Ungrammatical Language

6.1 Introduction

In a discussion on the viability of corpora as linguistic evidence in Section 2.3.2, I referred to the increased use of realistic data in the field of computational linguistics. Along with the increased use of the corpus, there has been increased interest in tools and methods which make use of corpora to evaluate the accuracy of various computational linguistic applications, including parsers. Carroll (1998b) distinguishes between the *intrinsic* evaluation of a parser and the *extrinsic* evaluation: intrinsic evaluation is concerned with how well a parser performs with respect to some particular application in which it is embedded, e.g. a machine translation system. Extrinisic evaluation concerns the evaluation of a parser's output with respect to some gold standard set of parses. This chapter is concerned with extrinisic rather than intrinsic evaluation.

A standard extrinsic evaluation method has been adopted for evaluating the accuracy of a parser: the analyses produced by a parser for a set of test sentences are compared to gold standard analyses produced by human judges for the same sentences. A treebank is the source of both the test sentences and the human parses of these sentences, and some quantitative measure is usually adopted to compare analyses. Although the sentences on which a parser is tested are generally assumed to be grammatical, it is also necessary to evaluate how well a parser copes with ungrammatical sentences, and in this chapter, I present a method to do this. This method does not measure how well a parser provides feedback on the errors in an ungrammatical sentence, but rather it measures how well the parses of the ungrammatical sentence reflect its meaning. It is similar to the standard method, but it does not involve a humanly annotated treebank. It does, however, depend on the parallel corpus described in Chapter 3.

Chapter 6 is organized as follows: in Section 6.2, I provide an overview of the evaluation procedure. In Section 6.3, I go into more detail and explain, by way of examples using different kinds of ungrammaticality and different kinds of parse structures, how a parse for an ungrammatical sentence can be considered to be completely accurate. In Section 6.4, the results of applying the evaluation procedure to two popular probabilistic parsers are provided and discussed. In Section 6.5, I discuss a criticism of the evaluation procedure and present a modified version of the evaluation procedure which takes this criticism into account. The results from Secton 6.4 are also modified and re-displayed in light of this. Finally, in Section 6.6, the work presented in this chapter is summarized, and possible improvements and extensions are discussed.

6.2 The Evaluation Procedure: an Overview

The evaluation method presented in this chapter is based on three assumptions. The first assumption is that every ungrammatical sentence has a grammatical counterpart which expresses the same meaning as the ungrammatical sentence and which would have been produced had the source of error been removed. This assumption is at the heart of the parallel *corpus1* and *corpus2*, described in Chapter 3. For example, the person who was responsible for Sentence (6.1) from *corpus1* really meant to write (6.2) from *corpus2*.[1]

(6.1) *I didn't* ***wanted*** *to delete the track.* (A.649)

[1]The example sentences in this chapter are abbreviated versions of sentences in *corpus1* and *corpus2*. They are abbreviated because I will be providing sample parse trees for many of them, and small parse trees are easier to read than larger ones.

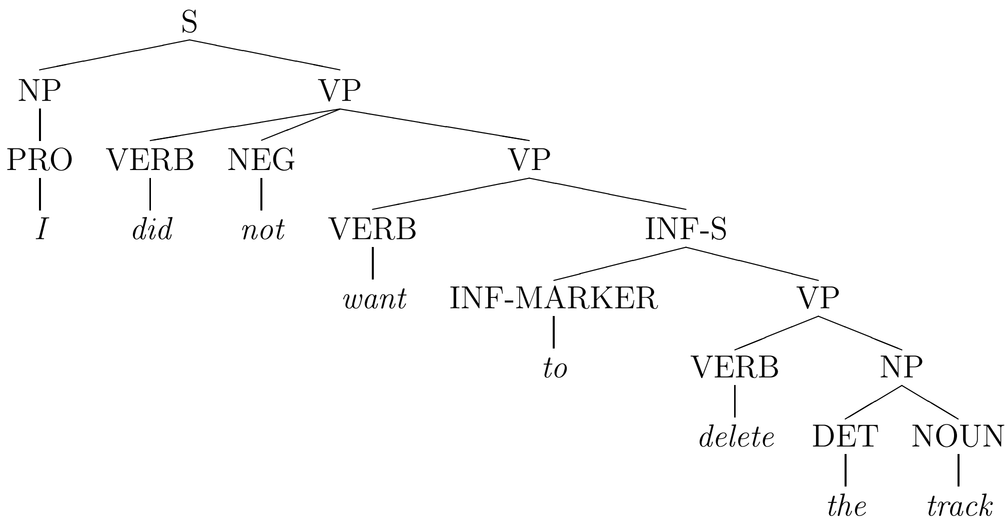


Figure 6.1: A Parse of Sentence (6.2)

(6.2) *I didn't **want** to delete the track.*

The second assumption upon which this evaluation method is based, is that a parse for a particular sentence reflects that sentence's meaning. This is an uncontroversial assumption: we parse sentences in order to grasp their structure; once their structure is clear, their overall meaning is closer at hand (see Section 2.2.1 for a related discussion). For example, the parse tree in Fig. 6.1 shows how groupings of the words in (6.2) reflect meaningful components which are combined into a coherent whole.

The third assumption follows on from the first two: if every ungrammatical sentence is really a (distorted) grammatical sentence, and the analysis or parse of a sentence reflects its meaning, then a parser should return an analysis for an ungrammatical sentence which is as close as possible to the analysis it returns for the corresponding grammatical sentence. In other words, a parser that is truly robust to errors should be able to see right through an ungrammatical sentence to the grammatical sentence at its heart, and produce a parse which reflects the meaning of the grammatical sentence. Thus, a truly robust parser will return the analysis in Fig. 6.1 for the ungrammatical (6.1).

The test data for the evaluation procedure are the ungrammatical sentences in *corpus1*. The grammatical sentences which make up *corpus2*, or rather, *the parses of these sentences produced by the parser being evaluated*, provide the gold standard or reference for the evaluation method. The evaluation procedure is presented in Fig. 6.2. The procedure takes into account the fact that an ungrammatical *corpus1* sentence might have more than one correction in *corpus2*, and hence, that there is more one than reference parse to which the parse of the ungrammatical sentence must be compared. When this happens, the parse of the ungrammatical sentence is compared to each reference parse in turn[2], and the reference which has the most number of parse constituents in common with the test parse is used. For example, consider the ungrammatical sentence (2.11), repeated for convenience as (6.3), which has two corrections in *corpus2*, (2.12 and 2.13), repeated for convenience as (6.4) and (6.5):

(6.3) *Want to **saving** money?* (A.91)

(6.4) *Want to **save** money?*

(6.5) *Want to **start** saving money?*

If some parse comparison metric deems that the parse for (6.3) has, for example, four of its six parse constituents in common with the parse for (6.4) and five out of six in common with the parse for (6.5), the higher score will be preferred. It should also be noted, that the method presented in

[2]In practice, there are only ever two reference parses because I imposed a practical limit of two corrections per sentence (see Section 3.4).

PRELIMINARIES:

Let $parserA$ be the parser being evaluated.

Let $parseset1$ be the set of parses produced by $parserA$ for the sentences in $corpus1$.

Let $parseset2$ be the set of parses produced by $parserA$ for the sentences in $corpus2$.

$parseset2$ are the *reference* parses.

Let $totalcons1$ be the number of parse constituents in $parseset1$.

Let $totalcons2$ be the number of parse constituents in $parseset2$.

Let $parse : corpus1 \cup corpus2 \rightarrow parseset1 \cup parseset2$ be a function which maps a sentence to the parse produced by $parserA$ for the sentence.

Let $correct : corpus1 \rightarrow \{x : x \in corpus2\}$ be a function which maps an ungrammatical sentence in *corpus1* to its corrected version(s) in *corpus2*

ALGORITHM:

$totalcorrect \leftarrow 0$

for each sentence *ungram* in *corpus1* **do**

  $best \leftarrow 0$

  for each grammatical counterpart *gram* in $correct(ungram)$ **do**

    $correctcons \leftarrow 0$

    for each constituent $con \in parse(ungram)$ **do**

      if $con \in parse(gram)$ **then**

        $correctcons \leftarrow correctcons + 1$

    if $correctcons > best$ **then**

      $best \leftarrow correctcons$

  $totalcorrect \leftarrow totalcorrect + best$

$precision \leftarrow \frac{totalcorrect}{totalcons1}$

$recall \leftarrow \frac{totalcorrect}{totalcons2}$

Figure 6.2: Evaluation Procedure

Fig. 6.2 assumes that only one parse is returned for a sentence but it could be adapted in order to evaluate a parser which returns a set of parses for a sentence.

A difficulty associated with reference parse sets such as the Penn Treebank (Marcus *et al.* 1993) is that they are unsuitable for evaluating parses whose structure is not directly comparable with the structure of the treebank parses. Various solutions to this problem have been suggested: one solution is to devise mappings from treebank trees to another linguistic representation format (see, for example, Lin (1998), Clark and Hockenmaier (2002) and Frank *et al.* (2003)). Gaizauskas *et al.* (1998) argue that the reference parses should be relatively flat structures encoding information upon which there is general agreement within the linguistics community, and these reference parses should be generately automatically from existing treebanks; Musillo and Sima'an (2002) argue that parallel treebanks for the same corpus of sentences should be created for the various linguistic frameworks. This problem of comparing parse representations from different frameworks does not, however, occur for the evaluation method outlined in Fig. 6.2: the parses being evaluated and the reference parses are guaranteed to be directly comparable since they have been produced by the same parser. Thus, this method is general enough to be used with any kind of parser, regardless of how it represents its output.

The quantitative measures of precision and recall are used to compute similarity. The flexibility in the procedure lies in deciding how to define a parse constituent and when to consider two constituents to be equal. The Parseval method (Black *et al.* 1991), for example, is specifically designed to handle parses which are represented using phrase structure trees. Constituents are sequences of words in the sentence organized into phrases and a constituent in a test parse is considered to be equal to another in the reference parse if they enclose the same word sequence at the same point in the sentence. Consider a possible parse for sentence (6.2) shown in Fig. 6.3. Assume that this tree is the test parse and that the parse tree shown in Fig. 6.1 is the reference parse against which it will be evaluated. The only difference between the test parse tree and the reference parse tree is that the phrase *to delete the track* is considered to be a noun phrase in the test parse and an infinitival clause in the reference parse. Thus, according to the Parseval measures, the test parse will receive a score of 100% precision and 100% recall. A stricter version of equality is one where a test constituent must enclose the same word sequence *and* be labelled by the same category as a constituent in the reference parse. According to these stricter measures, the test parse in Fig. 6.3 receives a precision score of 85.7% and a recall score of 85.7%. This is because both the test parse and the reference parse contain seven constituents (excluding pre-terminal constituents) and one of the seven constituents in the test parse, the one labelled by *NP* and enclosing the word sequence *to delete the track* is no longer considered to be equal to any constituent in the reference parse. These stricter precision/recall measures are known as *labelled* precision/recall.

The Parseval measures, although quite widely used to evaluate parser accuracy, have many critics. Sampson and Babarczy (2002) propose the Leaf-Ancestor (LA) measure as an alternative to the Parseval measure for comparing phrase structure trees. The LA measure compares the lineage (sequence of categories from root to terminal) of each word in the test parse to the lineages of the corresponding words in the reference parses, and they argue that this measure reflects our intuitive notion of what constitutes a good parse better than the Parseval measure. Others have argued for less

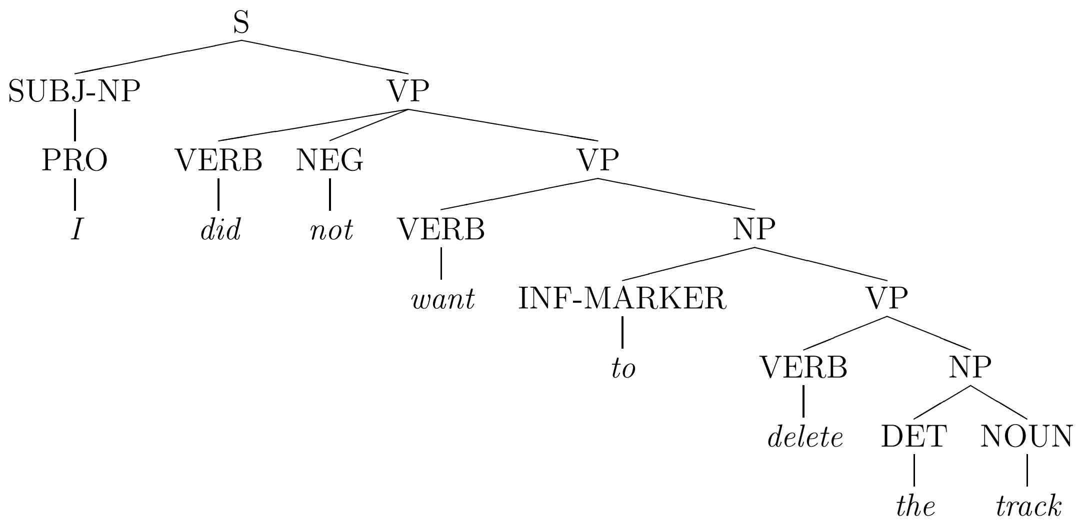


Figure 6.3: Another Parse of Sentence (6.2)

tree-oriented measures which allow a parse to be evaluated in terms of the information it returns concerning the semantic relationships between the words in a sentence (see for example Briscoe *et al.* (2002)). Lin (1998) and Kübler and Telljohann (2002) have argued that dependency-based evaluation measures are a better way to evaluate the information returned by a parser. According to a dependency-based evaluation method, a constituent is a dependency relationship between one word in the sentence and another. A dependency relationship in a test analysis is considered correct if the same relationship holds between the same two words in the reference analysis. It is outside the scope of this thesis to discuss the relative merits of each particular parse accuracy measure: the evaluation procedure described in Fig. 6.2 is general enough to be compatible with any measure which breaks a parse into components. In the following section, the notion of parse comparison will be explored in more detail, using phrase-structure trees and dependency analyses as examples.

6.3 Comparing Grammatical and Ungrammatical Parses

Of course there is an obvious difference between the general evaluation of a parser's accuracy and the evaluation of its accuracy with respect to ungrammatical sentences: in the former case, the two parses being compared reflect the same sentence whereas in the latter case the two parses reflect slightly different sentences (since the reference parse(s) describes the corrected version(s) of the sentence described by the test parse). The comparison metric needs to take this into account so that the ungrammatical sentence is not penalized for consisting of a slightly different word sequence to its grammatical counterpart(s). In fact, the comparison metric should allow a parse for an ungrammatical sentence to attain 100% on its precision and recall scores. This will be illustrated in the following sections, with each section describing a particular type of ill-formed sentence (see Fig. 5.1). For each example, a phrase-structure analysis *and* a dependency analysis will be provided. The phrase-structure analyses will be compared using the Parseval labelled precision/recall metrics, but they could just as easily be compared using the LA metric. The example analyses shown are hypothetical analyses and have not been produced by any existing parser.

Section 6.3.1 deals with ungrammatical sentences containing a substitution error (see Section 5.2.1), Section 6.3.2 is concerned with sentences which contain an insertion error (see Section 5.2.2) and Section 6.3.3 deals with sentences containing a deletion error (see Section 5.2.3). Recall that a sentence containing a substitution error is corrected by replacing one word in the sentence with another, a sentence containing a deletion error is corrected by deleting a word from the sentence and a sentence containing an insertion error is corrected by adding a word to the sentence. Sentences involving composite errors or more than one error are not discussed explicitly but can be viewed as a combination of the other three.

6.3.1 Substitution Errors

Consider, for example, the ungrammatical sentence (6.6):

(6.6) *A romance* ***in*** *coming your way.* (A.659)

(6.6) contains a common error involving the mistyping of *is* to produce *in*. The corrected version of this sentence is the grammatical (6.7):

(6.7) *A romance* ***is*** *coming your way.*

Fig. 6.4 depicts the phrase structure parse tree produced by some hypothetical parser (and its grammar) for the grammatical sentence (6.7). The parse for the ungrammatical sentence (6.6) shown in Fig. 6.5 should be considered completely accurate. This parse makes the crucial recognition that the preposition *in* is part of a verb phrase and contrasts in this way with the quite possible but less accurate parse for the same sentence, shown in Fig. 6.6. In this case, the presence of the preposition *in* has caused the parser to build the false structure of a noun phrase containing a prepositional phrase and to omit the verb phrase, with the result that this parse fails to reflect the sentence's true meaning. Four of the seven constituents[3] in the parse are correct, when measured against the parse for the grammatical sentence in Fig.6.4 (which contains five constituents).

Of course, a parse can be represented using a dependency analysis instead of a phrase structure analysis. Following Lin (1998), a dependency analysis

[3]Constituents labelled by pre-terminal symbols are not included in the evaluation.

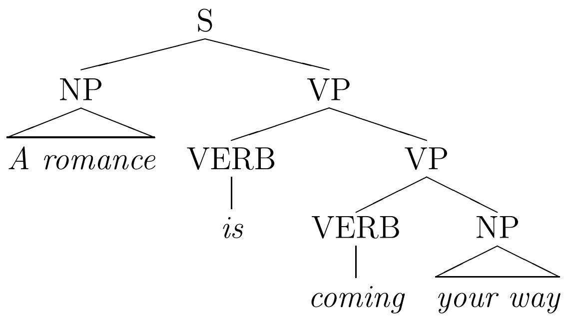


Figure 6.4: A Parse for Sentence (6.7)

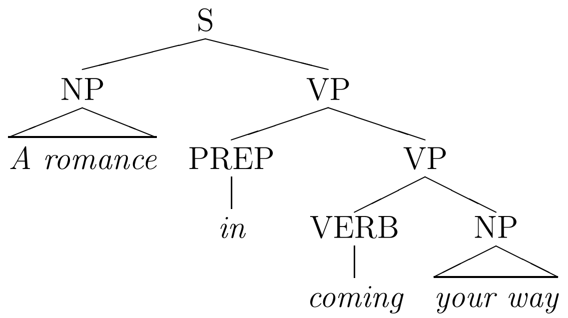


Figure 6.5: A Completely Accurate Parse for Sentence (6.6)

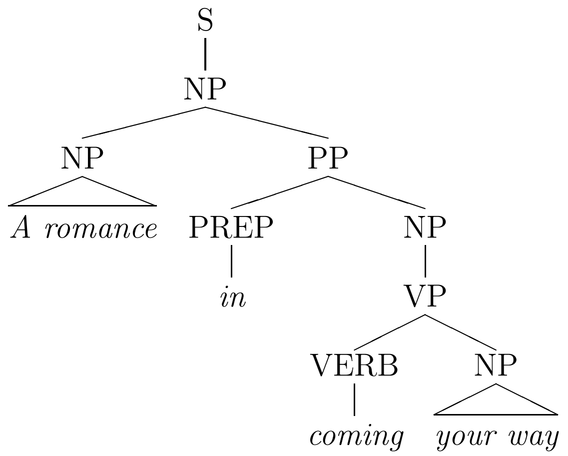


Figure 6.6: A Less Accurate Parse for Sentence (6.6)

consists of a set of tuples where each tuple represents a word in the sentence and has the following form:

(`Word`,`Category`,[`Head`], [`Relationship`])

Word is the actual word in the sentence, *Category* is its part of speech category, *Head* is another word in the sentence which *Word* is dependent on. *Relationship* specifies the nature of the dependency relationship between *Word* and *Head*. *Head* and *Relationship* are optional and can be omitted for words in the sentence which are not dependent on any other word. An example is the *head* word of the sentence which is not dependent on any other word and upon which all other words are directly or indirectly dependent.

Fig. 6.7 shows a dependency analysis produced by an imaginary parser for the grammatical example (6.7). This analysis specifies that *coming* is the head of the sentence, its subject is headed by *romance* and its modifier is headed by *way*. A completely accurate dependency analysis for the ungrammatical sentence (6.6) is depicted in Fig. 6.8. This analysis recognizes that the word *in*, although a preposition, is dependent on the verb *coming* and the nature of the dependency is an auxiliary verb relationship. The less accurate dependency analysis in Fig. 6.9 fails to make such a recognition. The important difference between this analysis and the correct one is the difference in the head word: according to this analysis, the head is the noun *romance* instead of the verb *coming*. Three of the five dependency relationships in this analysis are correct when measured against the five dependency relationships in the analysis for the grammatical sentence in Fig. 6.7.

(**A**,det,romance,det),
(**romance**,noun,coming,subj),
(**is**,verb,coming,aux),
(**coming**,verb),
(**your**,noun,way,gen),
(**way**,noun,coming,mod)

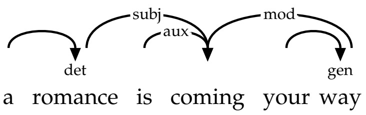


Figure 6.7: A Dependency Analysis of Sentence (6.7)

(**A**,det,romance,det),
(**romance**,noun,coming,subj),
(**in**,preposition,coming,aux)
(**coming**,verb),
(**your**,noun,way,gen),
(**way**,noun,coming,mod)

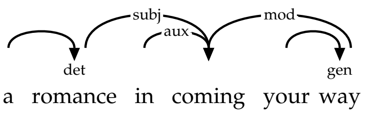


Figure 6.8: A Completely Accurate Dependency Analysis of Sentence (6.6)

(**A**,det,romance,det),
(**romance**,noun),
(**in**,preposition,romance,mod)
(**coming**,verb,in,pcomp),
(**your**,noun,way,gen),
(**way**,noun,coming,mod)

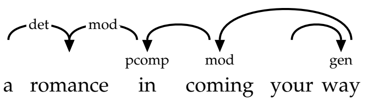


Figure 6.9: A Less Accurate Dependency Analysis of Sentence (6.6)

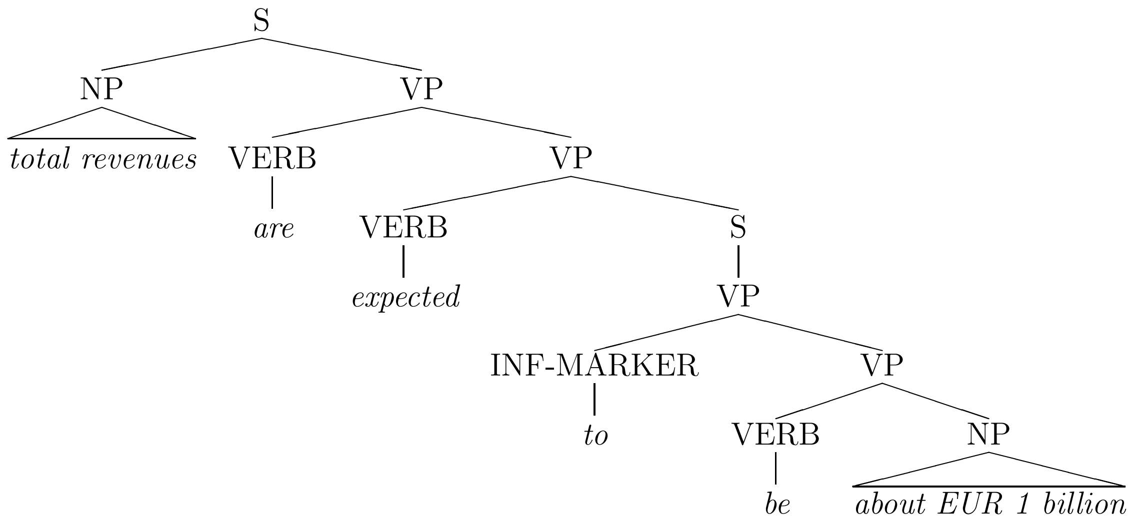


Figure 6.10: A Parse of Sentence (6.9)

6.3.2 Insertion Errors

As a second example, consider the ungrammatical sentence (6.8) and a correction of this sentence in *corpus2*, the grammatical sentence (6.9):

(6.8) *Total revenues are expected* ***to about*** *EUR 1 billion.* (A.180)

(6.9) *Total revenues are expected* ***to be about*** *EUR 1 billion.*

Assume that a hypothetical parser produces the parse tree in Fig. 6.10 for the grammatical (6.9). Correspondingly, a completely accurate parse of the ungrammatical (6.8) would be the parse tree shown in Fig. 6.11. In contrast, the plausible parse in Fig. 6.12 would receive a lower score because it misdiagnoses the infinitival marker *to* as introducing a prepositional phrase: five of the six constituents in this parse are correct, and three of the eight constituents in the correct analysis do not appear at all in this analysis.

A dependency analysis of the grammatical sentence (6.9) is shown in Fig. 6.13. A completely accurate dependency analysis for the ungrammati-

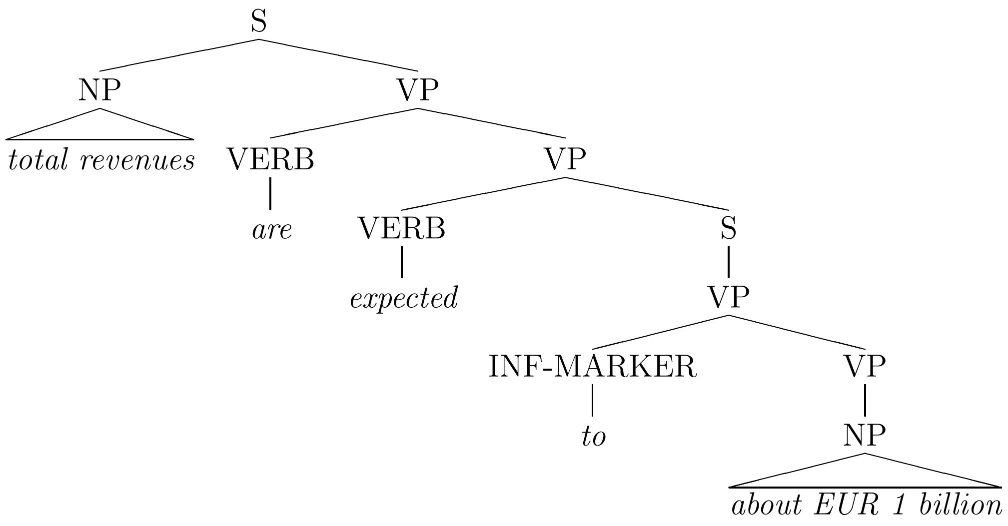


Figure 6.11: An Accurate Parse of Sentence (6.8)

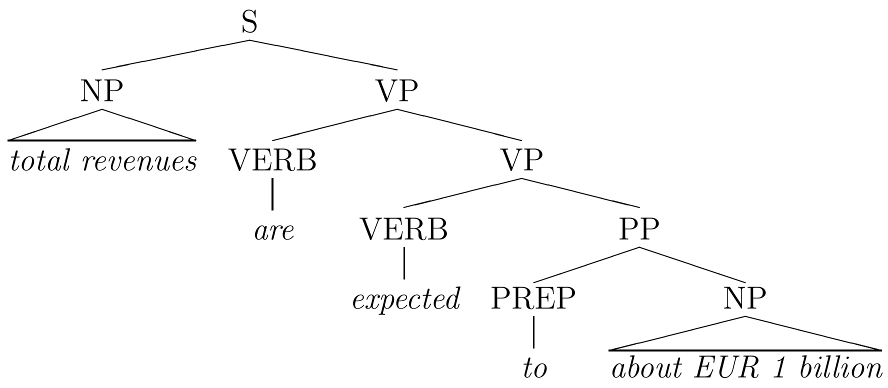


Figure 6.12: A Less Accurate Parse of Sentence (6.8)

(**Total**,adjective,revenues,mod),
(**revenues**,noun,expected,obj1),
(**are**,verb,expected,aux),
(**expected**,verb),
(**to**,infmarker,be,aux),
(**be**,verb,expected,obj2),
(**about**,prep,be,pred),
(**EUR**,noun,billion,mod),
(**1**,noun,billion,mod),
(**billion**,noun,about,pcomp)

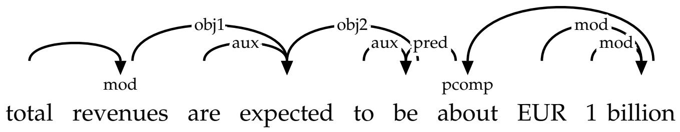


Figure 6.13: A Dependency Analysis of Sentence (6.9)

cal sentence (6.8) is shown in Fig. 6.14. In this analysis, () is used to indicate a trace or a non-overt element in the sentence. This analysis should be considered to be 100% accurate since it captures all and only the dependencies present in the reference analysis in Fig. 6.13. In contrast, the analysis in Fig. 6.15 loses some of the dependencies and proposes others which should not be there: there are five dependencies shared by this analysis (containing eight dependencies in total) and by the reference analysis (containing nine dependencies).

6.3.3 Deletion Errors

As a third and final example, consider the ungrammatical sentence (6.10) which contains a superfluous word:

(6.10)*Annotators parse* ***to*** *the sentences in a corpus.* (A.451)

The corrected version of this sentence in *corpus2* is Sentence (6.11):

(**Total**,adjective,revenues,mod),
(**revenues**,noun,expected,obj1),
(**are**,verb,expected,aux),
(**expected**,verb),
(**to**,infmarker,be,aux),
(**()**,verb,expected,obj2),
(**about**,prep,(),pred),
(**EUR**,noun,billion,mod),
(**1**,noun,billion,mod),
(**billion**,noun,about,pcomp)

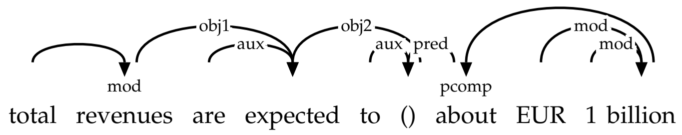


Figure 6.14: A Completely Accurate Dependency Analysis of Sentence (6.8)

(**Total**,adjective,revenues,mod),
(**revenues**,noun,expected,obj1),
(**are**,verb,expected,aux),
(**expected**,verb),
(**to**,preposition,expected,mod),
(**about**,adjective,billion,mod),
(**EUR**,noun,billion,mod),
(**1**,noun,billion,mod),
(**billion**,noun,to,pcomp)

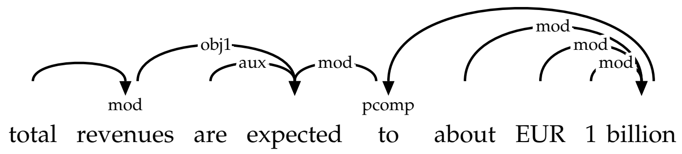


Figure 6.15: A Less Accurate Dependency Analysis of Sentence (6.8)

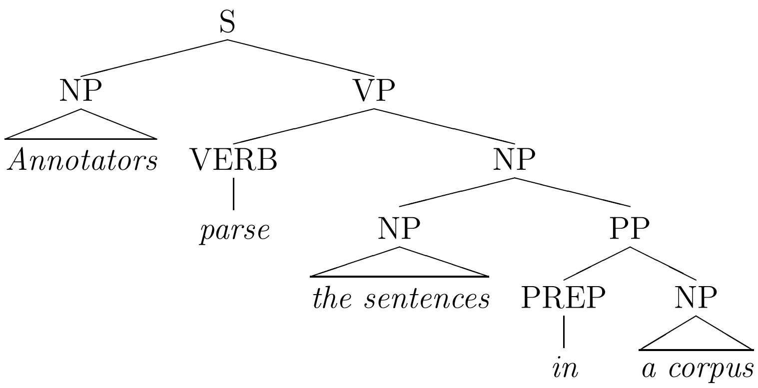


Figure 6.16: A Parse of Sentence (6.11)

(6.11)*Annotators parse the sentences in a corpus.*

Fig. 6.16 shows the phrase-structure parse tree for the corrected sentence (6.11). An accurate parse for the ungrammatical sentence (6.10) is shown in Fig. 6.17. Here, the superfluous *to* does not affect the constituent structure of the sentence. Equally accurate parses are shown in Figs. 6.18 and 6.19. The only difference between the parses in Figs. 6.17, 6.18 and 6.19 is the level where the word *to* is attached. In all three parses, *to* has not introduced any extra structure, which is a desirable result since the word does not contribute to the sentence's meaning. This can be seen more clearly if it is contrasted with another possible and less accurate parse, shown in Fig. 6.20, where *to the sentences in a corpus* is analysed as a prepositional phrase: seven of the eight constituents in this parse are considered correct.

A dependency parse for the grammatical (6.11) is shown in Fig. 6.21. An accurate dependency analysis for the ungrammatical sentence (6.10) is shown in Fig. 6.22. In this analysis, the word *to* is recognized as a preposition[4] but it is not linked to the other words in the sentence since it is not dependent on any of them and none are dependent on it. Thus, this analysis preserves all

[4]It could, of course, also be recognised as an infinitival marker.

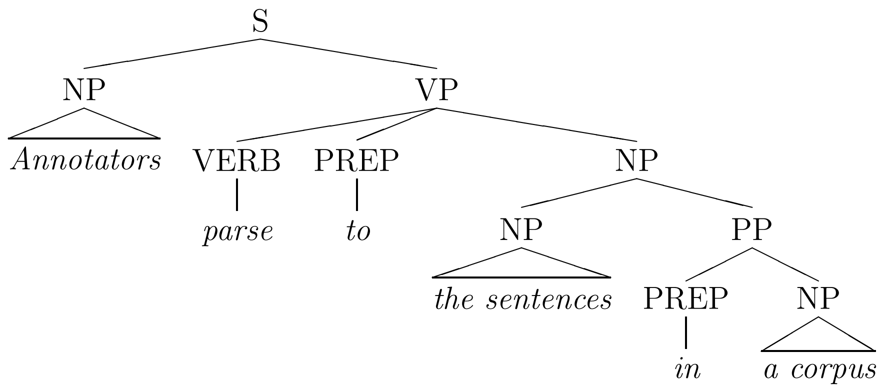


Figure 6.17: An Accurate parse of Sentence (6.10)

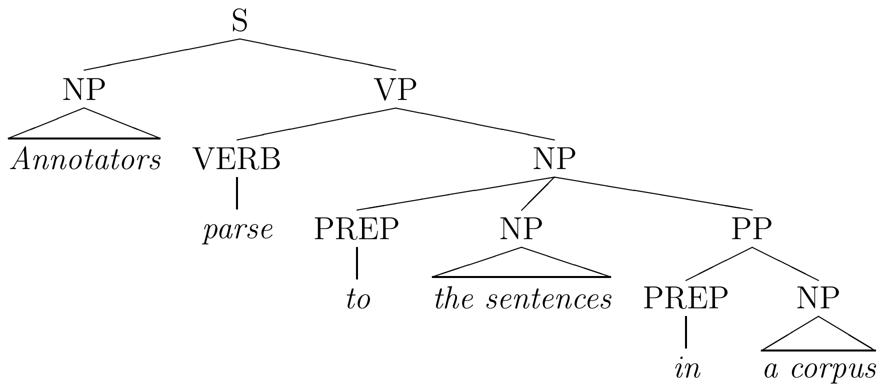


Figure 6.18: An Accurate Parse of Sentence (6.10)

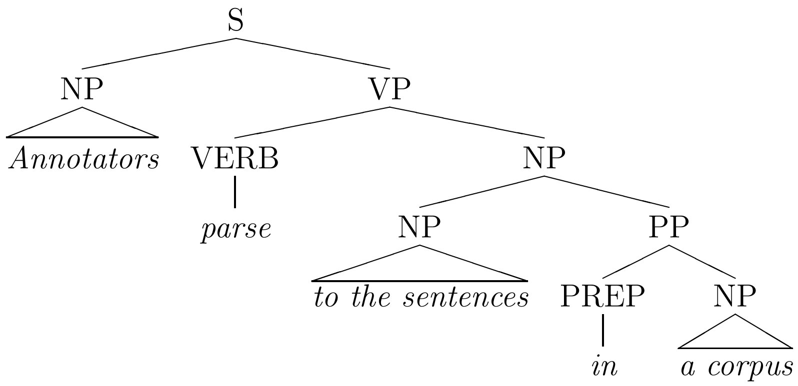


Figure 6.19: An Accurate Parse of Sentence (6.10)

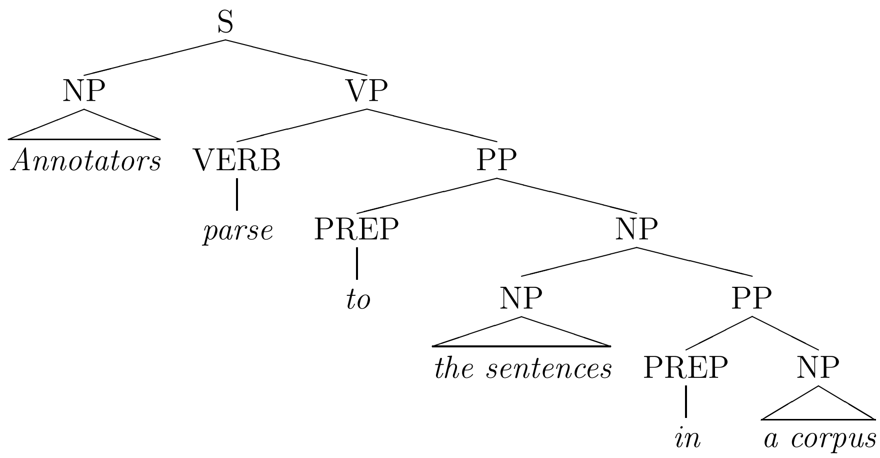


Figure 6.20: A Less Accurate parse of Sentence (6.10)

the dependencies present in the analysis for the grammatical sentence and introduces no others. A less accurate analysis links the superfluous word to the other words in some way, e.g. in the analysis in Fig. 6.23, the object dependency between the verb *parse* and the noun *sentences* has been lost, and by building the superfluous *to* into the dependency structure of the sentence, it has been imparted meaning which it does not in reality possess. Five of the seven dependencies are correct when measured against the analysis of the grammatical sentence in Fig. 6.21 (which contains six dependencies).

This section has provided examples of how the evaluation procedure presented in Fig. 6.2 can be applied, demonstrating its applicability within various grammatical frameworks by using two very different frameworks as hypothetical examples. In the following section, the procedure is applied to two real parsers.

(**Annotators**,noun,parse,subj),
(**parse**,verb),
(**the**,det,sentences,det),
(**sentences**,noun,parse,obj),
(**in**,preposition,sentences,mod),
(**a**,det,corpus,det),
(**corpus**,noun,in,pcomp)

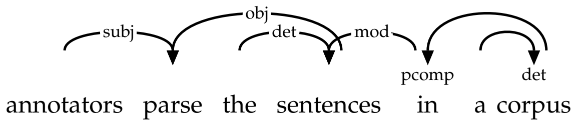


Figure 6.21: A Dependency Analysis of Sentence (6.11)

(**Annotators**,noun,parse,subj),
(**parse**,verb),
(**to**,preposition),
(**the**,det,sentences,det),
(**sentences**,noun,parse,obj),
(**in**,preposition,sentences,mod)
(**a**,det,corpus,det),
(**corpus**,noun,in,pcomp)

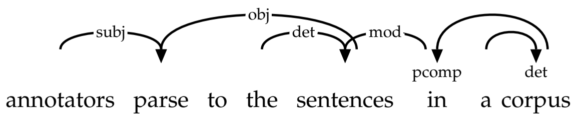


Figure 6.22: An Accurate Dependency Analysis of Sentence (6.10)

(**Annotators**,noun,parse,subj),
(**parse**,verb),
(**to**,preposition,parse,mod),
(**the**,det,sentences,det),
(**sentences**,noun,to,pcomp),
(**in**,preposition,sentences,mod),
(**a**,det,corpus,det),
(**corpus**,noun,in,pcomp)

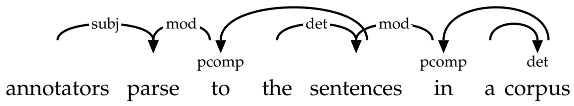


Figure 6.23: A Less Accurate Dependency Analysis of Sentence (6.10)

6.4 Applying the Evaluation Measure

In this section, I describe how the evaluation procedure presented in the previous two sections was applied to two very similar parsers. In Section 6.4.1, I describe how the evaluation procedure was applied, in Section 6.4.2, I present the results produced by the procedure for both parsers, and, in Section 6.4.3, I discuss these results.

6.4.1 Procedure

The evaluation method was applied to the wide-coverage probabilistic parser described by Charniak (2000) and the wide-coverage probabilistic parser described by Collins (1999).[5] These parsers are capable of returning parses

[5]Charniak's parser was downloaded from *ftp://ftp.cs.brown.edu/pub/nlparser/* in March 2003. Collins' parser was downloaded from *http://www.ai.mit.edu/people/mcollins* in March 2003. Collins' parser provides three different linguistic models and the third one of these was used. The third model extends the other two by providing a probabilistic treatment of wh-movement. Of the three models, the second and third slightly outperform the first when evaluated in the standard way on sentences from the Penn

for most of the ungrammatical sentences in *corpus1* which means that their ability to handle ungrammaticality can be meaningfully evaluated using the full corpus. However, twenty-seven *corpus1* sentences were not included in the test data: seventeen of these were sentences such as Sentence (6.12) and Sentence (6.13) where the error involved a missing or extra full stop. These sentences were excluded because either the *corpus1* sentence or its *corpus2* equivalent involves two sentences instead of one.

(6.12) *Event organisers effectively laid siege to an enduring **tradition Hollywood** frivolity, it seems, has been a casualty of the war in Iraq.* (A.816)

(6.13) *Don't let anyone tell you**.** You are being overambitious.* (A.862)

The other ten sentences were those for which no parse was returned by one or both of the two parsers under evaluation. Of these, eight were sentences whose length exceeds fifty words.

The procedure described in Fig. 6.2 is applied to both parsers. Labelled precision/recall is the metric used to compute parse similarity. In order to take into account the fact that the reference parses and the test parses describe slightly different word sequences and in order to facilitate easier comparison of the two parses, the reference parses are converted into parses which describe the ungrammatical sentences but which would achieve 100% precision and recall scores if compared to the reference parse. This means, for example, that the parse returned by the hypothetical parser in Section 6.3 for the ungrammatical sentence (6.8) is measured against the parse for this sentence in Fig. 6.11 instead of the parse for its grammatical equivalent in

Treebank. The input to the parser must be tagged for part-of-speech categories and this was carried out using the part-of-speech tagger recommended in the parser documentation (Ratnaparkhi 1996).

Parser	Precision	Recall	F-Score	100% Match	Problematic
Charniak	90.8	90.7	90.7	32.8	15.1
Collins	90.4	90.1	90.2	34.4	17.9

Figure 6.24: Overall Results for Charniak and Collins' Parsers

Fig. 6.10. If the error is a deletion error there is potentially more than one 100% accurate ungrammatical sentence parse for every grammatical sentence parse, and so the parse returned for the ungrammatical sentence (6.10) is measured against the parses in Figs. 6.17, 6.18 and 6.19 instead of the parse in Fig. 6.16. The code which performs this transformation to the set of grammatical sentence parses returned by both parsers is provided in Appendix D.

6.4.2 Results

Charniak's parser achieved a precision score of 90.8% and a recall score of 90.7%. The overall precision/recall figures were calculated for the *corpus1* sentence set using the equations shown in Fig. 6.2. In 32.8% of cases, there was a complete match between the parse for an ungrammatical test sentence and the parse for (one of) its reference parses. The percentage of *problematic* cases was also calculated: a parse for an ungrammatical sentence was deemed to be problematic if it achieved an individual precision or recall value of less than 75%. 15.1% of ungrammatical sentences fell into this category.

Collin's parser achieved an overall precision score of 90.4% and an overall recall score of 90.1%. In 34.4% of cases there was a complete match between the ungrammatical test parse and a reference parse, and in 17.9% of cases, an ungrammatical sentence parse achieved a relatively low-scoring precision or

Error Type	Precision	Recall	F-Score	100%	Problematic
Substitution Error					
Charniak	90.9	91.6	91.2	47.9	15.5
Collins	91.0	91.3	91.1	47.7	16.2
Insertion Error					
Charniak	92.5	88.1	90.2	20.0	16.3
Collins	91.7	87.4	89.5	24.6	20.2
Deletion Error					
Charniak	89.8	94.3	92.0	27.1	9.8
Collins	88.8	92.9	90.8	28.1	15.8
Composite Error					
Charniak	88.9	89.0	88.9	10.0	16.0
Collins	87.1	88.6	87.8	13.8	20.7
More Than One Error					
Charniak	88.3	89.1	88.7	17.6	17.6
Collins	88.9	87.8	88.3	21.4	23.2

Figure 6.25: Results according to Error Type

Error Type	Precision	Recall	F-Score	100%	Problematic
Agreement Errors					
Charniak	93.9	94.2	94.0	75.3	10.4
Collins	96.4	96.8	96.6	76.6	6.5
Wrong Preposition					
Charniak	95.3	95.1	95.2	72.7	3.0
Collins	96.1	95.8	95.9	62.5	0.0
Wrong Verb Form					
Charniak	93.1	94.2	93.6	40.0	6.7
Collins	89.3	90.1	89.7	41.5	14.6
Misspelling Category Change					
Charniak	85.4	86.5	85.9	22.8	27.5
Collins	86.1	86.8	86.4	25.0	28.0

Figure 6.26: Frequent Substitution Errors

recall value of less than 75%. The results for the two parsers are summarised in Figure 6.24. The f-score combines precision and recall into one score according to equation (6.14):

(6.14) $Fscore = \frac{2*Precision*Recall}{Precision+Recall}$

Fig. 6.25 shows the results for each particular type of error occurring in *corpus1.* Recall that a composite error involves two or more of the three main error types, e.g. the ungrammatical clause in (6.15) which is a combination of an incorrect word form (*remark* instead of *remarking*) and an extra word (*to*).

(6.15) *It is also* ***worth to remark*** *that.....* (A.850)

This is distinguished from a sentence which contains two or more individual errors, e.g. sentence (6.16) where a noun has been omitted after the verb *has* and the verb *derived* has been repeated.

(6.16) *The following roadmap for the* ***has*** *been derived* ***derived*** *for this presentation.* (A.797)

Substitution errors constitute the largest category. The results for the most common type of errors within this category are shown in Fig. 6.26.

6.4.3 Discussion

It seems that both parsers react in a very similar way to the ungrammatical sentences in *corpus1*, which is not surprising, since both have been trained on the same grammatical data. Overall, both parsers have performed well on the ungrammatical sentences in *corpus1*, achieving 100% accuracy for approximately one third of the test data. Charniak's parser slightly outperforms Collins' parser. Across error types, there is a tendency on the part of Collins' parser to produce more complete matches *and* more problematic cases than Charniak's. For both parsers, the deletion errors achieve a higher recall score in comparison to their precision score which suggests that these kinds of errors tend to introduce unwanted structure into a parse. Similarly, for both parsers, the insertion errors achieve a higher precision score in comparison to their recall score, suggesting that a lack of relevant structure is associated with these kinds of errors. This is expected. It is also not surprising that both parsers' f-scores for composite errors and multiple errors are slightly lower than the f-scores for errors involving one correction operator.

For some error types (agreement errors and the use of the wrong preposition), between 60% and 70% of cases obtained a complete match for both parsers, suggesting that these kinds of errors do not tend to affect these parsers in a negative way. Sentence (3.10), repeated for convenience as (6.17), is an example, however, of a sentence from *corpus1* where the agreement error in the sentence does cause it to be misparsed by Charniak's parser:

(6.17)*Apartments seized by CAB* ***goes*** *under the hammer.* (A.843)

(6.18)*Apartments seized by CAB* ***go*** *under the hammer.*

This ungrammatical sentence receives the inaccurate parse shown in Fig. 6.27. Its corrected version in (6.18) receives the quite different parse in Fig. 6.28. According to Charniak's parser, the verb *goes* in the ungrammatical (6.17) is more likely to agree with the noun *CAB* than with the noun *Apartments*, a situation which might not occur if an agreement error was anticipated by the parser. The same sentence achieved a precision/recall score of 100% when parsed with Collins' parser.[6]

[6]It does not fall under the remit of this thesis to explain why these parsers have produced the parses they have for the set of *corpus1* sentences. This section is simply high-

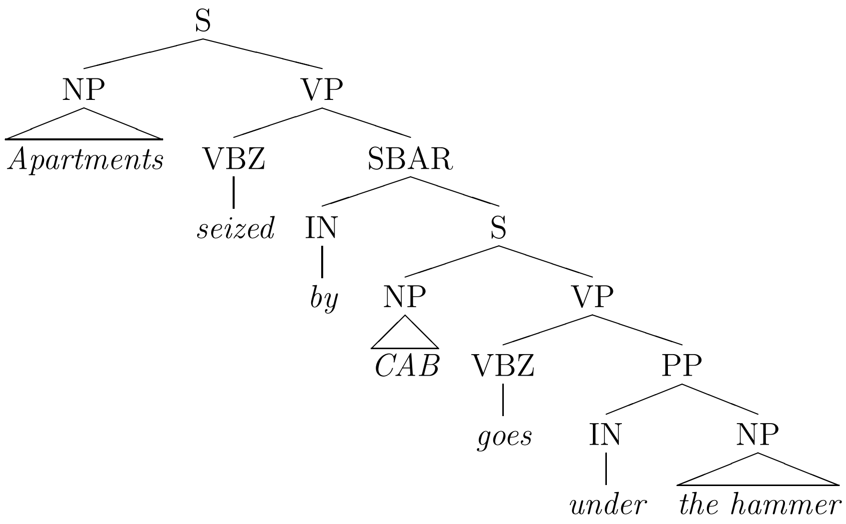


Figure 6.27: Parse Produced by Charniak's Parser for (6.17)

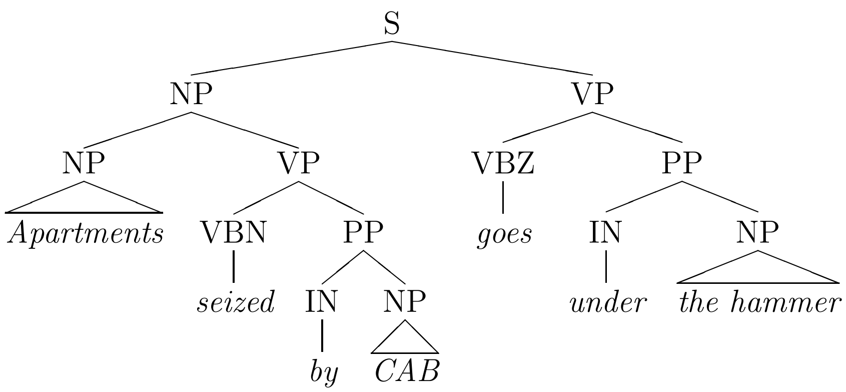


Figure 6.28: Parse Produced by Charniak's Parser for (6.18)

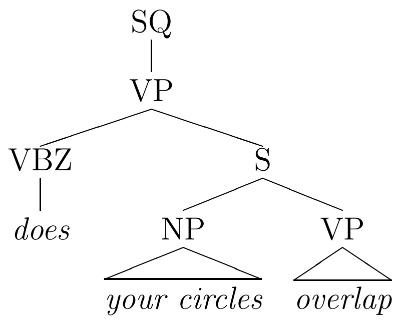


Figure 6.29: Parse Produced by Collins' Parser for (6.19)

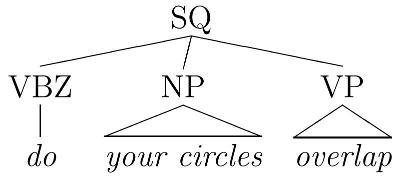


Figure 6.30: Parse Produced by Collins' Parser for (6.20)

In contrast, the ungrammatical (6.19) receives the parse in Fig. 6.29 from Collins' parser.

(6.19)***Does your circles*** *overlap?* (A.210)

(6.20)***Do your circles*** *overlap?*

Evaluated against the parse for its corrected version (6.20) shown in Fig. 6.30, the test parse receives a precision score of 3/5 and a recall score of 3/3. Charniak's parser returns a complete match for this sentence and its corrected version.

The problematic cases increased to nearly 30% when the ungrammatical sentence contained an erroneous word of a different category to the one it should have been. A typical example is the ill-formed (6.21) which contains

lighting interesting aspects of the evaluation results. These results will, I hope, be useful to those involved in the development of these and similar parsers.

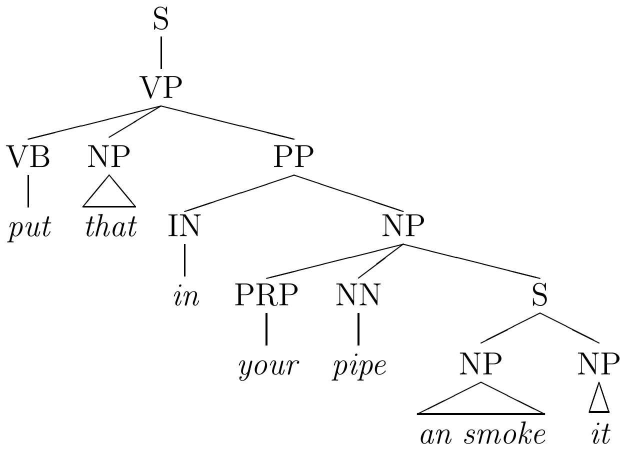


Figure 6.31: Parse Produced by Charniak's Parser for Sentence (6.21)

a context-sensitive spelling error and whose corrected version in *corpus2* is (6.22).

(6.21)*Put that in your pipe **an** smoke it.* (A.343)

(6.22)*Put that in your pipe **and** smoke it.*

Charniak's parser produces the peculiar parse in Fig. 6.31 for Sentence (6.21). Evaluated against the parse for its grammatical counterpart in Fig. 6.32, only three of its eight constituents are correct. The same sentence achieved a better but still relatively low score with Collins' parser.

In Section 2.4.1 of Chapter 2, I made the point that the probabilistic grammars used by such parsers need to be aware of ungrammaticality in language in order to be able to parse ungrammatical sentences properly. The success of Charniak's and Collins' parsers on the *corpus1* sentences does not invalidate this point. Although both parsers did well, there is certainly room for improvement: there are ungrammatical sentences whose meaning is not properly reflected by the parses produced for them by Charniak's and Collins' parsers. These occur for every error type but more frequently

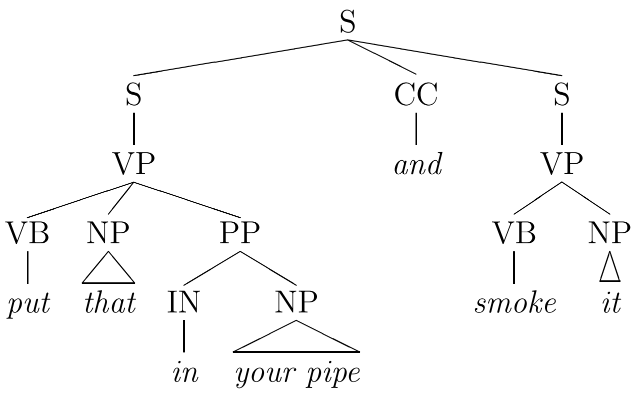


Figure 6.32: Parse produced by Charniak's Parser for Sentence (6.22)

for sentences containing context-sensitive spelling errors where there is a category difference between the substituting word and the substituted word, for sentences containing more than one error and for sentences containing a composite error.

6.5 Adapting the Evaluation Measure

There are two problems associated with using as reference parses the parses produced by the parser being evaluated for the grammatical sentences in *corpus2*. Both problems are rooted in the fact that there is no guarantee that the parser being evaluated is able to provide accurate parses for the grammatical sentences in *corpus2*. How can we be sure that these so-called gold standard parses deserve the name "gold standard" at all?

The first problem related to this is quite straightforward: if the parser being evaluated gives the inaccurate parse in Fig. 6.33 for the grammatical sentence (6.23) and the parse in Fig. 6.34 for the ungrammatical sentence (6.24), it will score 100% for its parse of the ungrammatical sentence according to the evaluation procedure which has been used up until now.

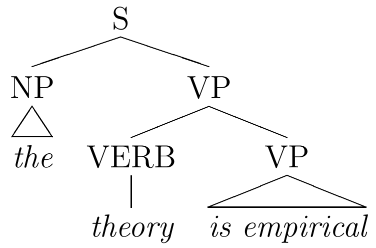


Figure 6.33: A (Bad) Parse of Sentence (6.23)

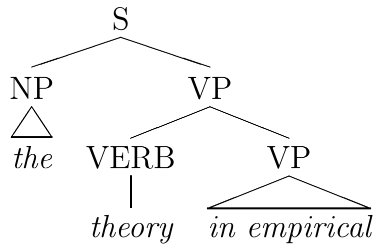


Figure 6.34: A (Bad) Parse of Sentence (6.24)

(6.23) *The theory* ***is*** *empirical.* (A.783)

(6.24) *The theory* ***in*** *empirical.*

It does not seem satisfactory that this parser attains a score of 100% because, although it copes well with the *in/is* confusion, it still does not capture the meaning of the sentence. The only way around this problem is to take the accuracy of the parser on grammatical sentences into consideration in the evaluation procedure.

According to the evalution procedure which was specified in Fig. 6.2, a constituent in an ungrammatical test parse is considered correct if it occurs in the grammatical reference parse. I will use the term *gram_and_ungram* to refer to the set of constituents which appear in both a test parse and its reference and which up until now have been considered to be correct. The set *gram_and_ungram* can be further partitioned into two disjoint sets *correct_gram_and_ungram* and *incorrect_gram_and_ungram*.

correct_gram_and_ungram is the set of constituents which are in both the reference parse set and the test parse set and which are actually correct, according to some human gold standard. Similarly, *incorrect_gram_and_ungram* is the set of constituents which are in both the reference parse set and the test parse set and which are incorrect, according to some human gold standard. Rather than actually producing gold standard parses of the grammatical sentences in *corpus2* – a worthwhile venture, but beyond the scope of this thesis – the *correct_gram_and_ungram* : *incorrect_gram_and_ungram* ratio can by estimated by looking at the score achieved by a parser when it has been evaluated on other grammatical text.[7]

Consider, for example, some hypothetical parser which has an accuracy level of 80% precision and 90% recall when evaluated in the standard way on grammatical treebank sentences. I will adopt the term *precision_gram* to refer to a parser's precision value on grammatical text, and the term *recall_gram* to refer to its recall value on grammatical text. Similarly, *precision_ungram* will refer to its precision on ungrammatical text and *recall_ungram* to its recall. When the evaluation procedure from Fig. 6.2 is applied to this parser, it has 4,000 constituents from its test parses, 4,500 constituents from its reference parses and 3,000 "correct" constituents. According to the old procedure for evaluating robustness, the *precision_ungram* value for this parser would be 3000/4000 or 75%. However, we can assume that only 80% of these 3000 are actually correct - the remaining 20% belong in the set *incorrect_gram_and_ungram* and can be discarded. Thus, the new *precision_ungram* value for this parser is 2400/4000 which is 60%. Similarly, its *recall_ungram* value is not 3000/4500 but 2700/4500 or 60%.

[7] I am grateful to Jonas Sjöbergh for pointing this out to me at LREC '04.

Charniak's parser and Collins' parser have both been evaluated against the same set of sentences from the Wall Street Journal section of the Penn Treebank. Charniak's parser receives a *precision_gram* and *recall_gram* score of 89.5% and Collins receives a *precision_gram* score of 88.04% and a *recall_gram* score of 88.25%.[8] Re-interpreting the results from Section 6.4, Charniak's parser's *precision_ungram* score is 81.3% and its *recall_ungram* value is 81.2%. Collins' parser's *precision_ungram* score is 79.6% and its *recall_ungram* score is 79.5%.

The second problem with the evaluation procedure occurs when the parser being evaluated produces an accurate parse for an ungrammatical sentence but an inaccurate parse for the corresponding grammatical sentence. Charniak's parser, for example, misparses the grammatical clause in (6.25) by treating *P* as a verb. The ill-formed clause (6.26) does not suffer the same problem and *P* is recognized as part of a name.

(6.25)*Melissa P probably* ***without*** *significant other* (A.190)

(6.26)*Melissa P probably* ***with out*** *significant other*

This means that the new evaluation procedure is possibly too pessismistic since it does not take into account the fact that an ungrammatical sentence parse which differs from its corresponding grammatical sentence parse might actually be a good parse. As it stands, the evaluation procedure still considers a constituent in a test parse to be correct only if the same constituent is also in the reference parse. The evaluation procedure could be modified so that a constituent which is not in the reference parse can also be considered correct. I will use the term *ungram_only* to refer to the set of parse

[8]These are the figures that are reported in the README documentation accompanying both parsers.

constituents which are in test parses but not in reference parses. The set *ungram_only* can be partitioned into two disjoint sets *correct_ungram_only* and *incorrect_ungram_only*.

correct_ungram_only is the set of constituents which are in the test parse set but not in the reference parse set and which are actually correct, according to some human gold standard. Similarly, *incorrect_ungram_only* is the set of constituents which are in the test parse set but not in the reference parse set and which are incorrect, according to some human gold standard. As was the case with *correct_gram_and_ungram* and *incorrect_gram_and_ungram* there is no automatic way of knowing how many parse constituents are in each of the subsets *correct_ungram_only* and *incorrect_ungram_only*. I propose an estimate of these numbers which can be computed by looking at a sample of one hundred parse constituents which are in the ungrammatical parse set ($parseset1$) but not in the grammatical parse set ($parseset2$), and counting how many of these are actually correct. I will refer to the percentage which are correct as $correct\_ungram\_only_{est}$. $correct\_ungram\_only_{est}$ can be multiplied by *ungram_only* to give an approximate value for *correct_ungram_only* and this value can be added to the number of correct constituents. Fig. 6.35 gives the revised evaluation procedure which takes the two problems discussed in this section into account.

$correct\_ungram\_only_{est}$ was calculated from the set of parses produced by both Charniak's parser and Collins' parser. Charniak's parser has an $incorrect\_ungram\_only_{est}$ value of 11% and Collins of 10%. The precision, recall and f-scores presented in Section 6.4 have been modified to reflect the new evaluation procedure in Fig. 6.35. These are shown in Figs. 6.36, 6.37 and 6.38.

PRELIMINARIES:

Let $parserA$ be the parser being evaluated.

Let $parseset1$ be the set of parses produced by $parserA$ for $corpus1$ sentences.

Let $parseset2$ be the set of parses produced by $parserA$ for the sentences in $corpus2$. $parseset$ are the $reference$ parses.

Let $totalcons1$ be the number of parse constituents in $parseset1$.

Let $totalcons2$ be the number of parse constituents in $parseset2$.

Let $parse : corpus1 \cup corpus2 \rightarrow parseset1 \cup parseset2$ be a function which maps a sentence to the parse produced by $parserA$ for the sentence.

Let $correct : corpus1 \rightarrow \{x : x \in corpus2\}$ be a function which maps an ungrammatical sentence in *corpus1* to its corrected version(s) in *corpus2*

ALGORITHM:

$totalcorrect \leftarrow 0$

for each sentence $ungram$ in *corpus1* **do**

  $best \leftarrow 0$

  for each grammatical counterpart c in $correct(ungram)$ **do**

    $correctcons \leftarrow 0$

    for each constituent $con \in parse(ungram)$ **do**

      if $con \in parse(c)$ **then**

        $correctcons \leftarrow correctcons + 1$

    if $correctcons > best$ **then**

      $best \leftarrow correctcons$

  $totalcorrect \leftarrow totalcorrect + best$

$totalincorrect \leftarrow totalcons1 - totalcorrect$

$precision \leftarrow \frac{(totalcorrect*precision\_gram)+(totalincorrect*correct\_ungram\_only_{est})}{totalcons1}$

$recall \leftarrow \frac{(totalcorrect*recall\_gram)+(totalincorrect*correct\_ungram\_only_{est})}{totalcons2}$

Figure 6.35: Revised Evaluation Procedure

Parser	Precision	Recall	F-Score
Charniak	82.3	82.2	82.2
Collins	80.6	80.4	80.5

Figure 6.36: Overall Results for Charniak and Collins' Parsers: Modified Results

Error Type	Precision	Recall	F-Score
Substitution Error			
Charniak	82.4	83.0	82.7
Collins	81.0	81.5	81.2
Insertion Error			
Charniak	83.6	79.6	81.6
Collins	81.5	77.9	79.7
Deletion Error			
Charniak	81.5	85.6	83.5
Collins	79.3	83.1	81.2
Composite Error			
Charniak	80.8	80.9	80.8
Collins	78.0	79.5	78.7
More Than One Error			
Charniak	80.3	81.0	80.7
Collins	79.4	78.6	79.0

Figure 6.37: Results according to Error Type: Modified Results

Error Type	Precision	Recall	F-Score
Agreement Errors			
Charniak	84.7	85.0	84.9
Collins	85.2	85.8	85.5
Wrong Preposition			
Charniak	85.8	85.7	85.7
Collins	85.0	84.9	85.0
Wrong Verb Form			
Charniak	84.1	85.0	84.6
Collins	79.7	80.6	80.1
Misspelling Category Change			
Charniak	78.1	79.0	78.5
Collins	77.2	78.0	77.6

Figure 6.38: Frequent Substitution Errors: Modified Results

6.6 Conclusion

This chapter has presented a procedure for evaluating a parser's ability to produce accurate parses for ungrammatical sentences. The corpus of ungrammatical language described in Chapter 3 provides the test sentences and the reference parses are provided by the parses produced by the parser under evaluation for the corrected versions of the test sentences. The basic procedure is presented and then adapted slightly to take into account the fact that no parser will ever be perfectly accurate in the face of grammatical input. The field of robustness evaluation is relatively new, and I know of only one other work which attempts to automatically evaluate the parsing of ungrammatical language as opposed to evaluating robustness in more general terms (see, for example, Basili and Zanzotto (2002)). Bigert *et al.* (2004) describe work which is similar to the work presented in this chapter and has been produced concurrently to it: their approach is based on the same assumption as mine, i.e. that a parser should produce an analysis for an ill-formed input which is as close as possible to the analysis of the corresponding well-formed input. The two main differences between their approach and this one is that they concentrate solely on spelling errors, and instead of starting with a set of ungrammatical sentences and deriving from them a set of grammatical sentences, they proceed in the opposite direction, by automatically introducing artificial spelling errors into a grammatical text.

A second contribution of this chapter to the field of robust parsing has been the application of this evaluation procedure to two wide-coverage probabilistic parsers (Collins 1999, Charniak 2000). The results show that these parsers return structures for ungrammatical sentences which are quite close to those returned for their corrected versions and that such parsers react quite robustly in the face of ungrammatical input. However, the results are not perfect, and if these parsers were equipped with knowledge of how grammatical sentences can sometimes become ungrammatical, their ability to parse ungrammatical sentences would improve.

I have emphasized that the evalution procedure presented in this chapter is general enough to be applied to any parser, regardless of the way it represents its parses. It would certainly be interesting to apply the evaluation procedure to a parser which returns linguistic structures informationally richer than the trees returned by Charniak's and Collins' parsers, e.g. the typed feature structures returned by the LKB parser (Copestake 2002). However, this procedure cannot be applied directly to the LKB parser because this parser does not produce analyses for ungrammatical sentences (see Section 4.2.1). It would need to be applied instead to a robust version of the LKB (the work in Chapter 7 is a step in that direction). Another step in this research would be to apply the procedure to a parser which returns dependency analyses instead of phrase structure trees, e.g. the Minipar parser (Lin 1998).

Chapter 7

Parsing Agreement Errors using Inconsistent Typed Feature Structures

7.1 Introduction

I ended the previous chapter by suggesting that the evaluation procedure presented in that chapter could eventually be applied to parsers which produce as output structures which are informationally richer than phrase structure trees. An example of such a structure is a typed feature structure, and examples of parsers which produce such structures are the Linguistic Knowledge Base (LKB) (Copestake 2002) and the Attribute Logic Engine (ALE) (Carpenter and Penn 2001). The work in this chapter is concerned with encoding ungrammatical sentences as typed feature structures.

The precise aim of this chapter is to provide a constraint relaxation treatment for agreement errors, a subset of the class of substitution errors (see Section 5.2.1), within the typed feature structure framework provided by Carpenter (1992). Since this is the underlying framework for Head-Driven Phrase Structure Grammar (HPSG)(Pollard and Sag 1994), the work presented in this chapter has applications in HPSG-based parsing. Indeed, it is tested using the English Resource Grammar (ERG) (Copestake and Flickinger 2000) which encodes HPSG principles. However, throughout this thesis I have tried to remain neutral with respect to any particular linguistic theory, and the feature structures presented will be clear to anyone with a basic knowledge of unification based approaches to grammar.

This chapter is laid out as follows: in Section 7.2, I discuss agreement errors and make a distinction between substitution errors which involve an agreement mismatch and those which do not. Examples from the ungrammatical sentence corpus, *corpus1*, of both kinds of error are provided. In Section 7.3, an overview of Carpenter's basic feature structure system is provided. The concepts of *type hierarchy, feature structure, subsumption* and *unification* are explained and the failure of classical unification to behave robustly when confronted with inconsistent inputs is outlined.

In Section 7.4, the operation of *robust agreement unification* is informally introduced. This is a new kind of unification specifically designed to treat agreement errors in a robust fashion. Alternative robust unification strategies are discussed and then the concept of *inconsistent agreement identity* which is central to the operation of *robust agreement unification* is introduced. In Section 7.5, the structures which arise from the operation of robust agreement unification, namely, *robust agreement feature structures* (RAFs), are defined, followed by a definition of robust agreement unification. In Section 7.6, I describe how I adapted the LKB parser (Copestake 2002) so that it employs robust agreement unification. I then describe how this implementation was tested using the ERG and the sentences from *corpus1* which contain an agreement error. In Section 7.7, I argue that robust agreement unification is not an appropriate treatment for substitution errors which do not involve agreement. The final section of this chapter reviews the main parts of this work and identifies how this work could be extended in future research.

7.2 Agreement Errors

Ferguson and Barlow (1988, p.1) define grammatical agreement as the phenomenon "by which a grammatical element X matches a grammatical element Y in property Z within some grammatical configuration." This is a deliberately vague definition since it is designed to apply to any particular manifestation of the phenomenon in any language. In this chapter, I restrict my attention to the two main types of grammatical agreement in English, namely, subject-verb agreement, in which X is the subject noun or pronoun in a clause, Y is the main verb and Z is the property of number or person, and, determiner-noun agreement, in which X is the determiner in a noun phrase, Y is the main noun and Z is the property of number or countability (mass versus count). I define an (English) agreement error as follows:

Definition 4 ***Agreement Errors in English***
A sentence contains an ***agreement error*** *if there is a mismatch in number or countability between a noun and its determiner in a noun phrase, or if there is a mismatch in number or person between a verb and its subject in a clause.*

Cases of anaphoric agreement errors such as the *corpus1* sentence (7.1), in which a singular subject and verb are co-referenced with a plural noun phrase in another clause, are not covered by Def. (4). When tested on the ERG using the LKB parser, this sentence was accepted as grammatical, because the agreement clash is an instance of a covert error in a syntactically well-formed sentence.

(7.1) *One of the most important arguments for* ***declarative approaches to NLP*** *is that* ***it is*** *monotonic.* (A.257)

The agreement errors which occurred in *corpus1* are analysed in Section 7.2.1 and in Section 7.2.2 the differences between agreement errors and other substitution errors are discussed.

7.2.1 Agreement Errors and *Corpus1*

There are sixty-eight sentences in *corpus1* containing an agreement error according to Def. 4. Some sentences in *corpus1* contained a composite error which involved agreement amongst other things, e.g. Sentence (7.2) which is

Error	Example	No.
Sing Subj, Plur Verb	***The command have*** *only optional arguments.*	18
Plur Subj, Sing Verb	***Several attempts has*** *been carried out.*	28
1stPer Subj, 3rdPer Verb	***I awaits*** *your response.*	1
Sing Det, Plur Noun	*Each subject judged* ***one lexicalizations****.*	10
Plur Det, Sing Noun	***These assumption*** *are stated explicitly.*	9
Count Det, Mass Noun	*They should possess* ***these information***	2

Figure 7.1: Agreement Errors in *corpus1*

a corrected version of Sentence (7.3):[1]

(7.2) *But also their real concerns* ***were*** *voiced for their safety in the future without a fire escape.* (A.273)

(7.3) *But they also* ***was*** *voiced their real concerns for their safety in the future without a fire escape.*

In (7.2), the plural *were* has replaced the singular *was* in (7.3) but this is only one of the things which needs to be done in order to make the first half of the sentence grammatical. Sentences such as (7.2) were not included in the set of agreement error sentences.

The table in Fig. 7.1 shows the frequency of agreement error types in the corpus. The table in Fig. 7.2 shows how these errors were corrected. Note that sixteen sentences containing an agreement error could be corrected in more than one way in the context in which the sentence appeared. An example is Sentence (7.4) in which there is a subject-verb number disagreement. This can be sensibly corrected in the given context by changing the number

[1]The more obvious correction, which is also in *corpus2*, is one where the verb *was* is simply deleted.

Correction	No.
Sing Subj, Plur Verb → Sing Subj, Sing Verb	12
Sing Subj, Plur Verb → Sing Subj, Sing Verb AND Plur Subj, Plur Verb	6
Plur Subj, Sing Verb → Plur Subj, Plur Verb	18
Plur Subj, Sing Verb → Plur Subj, Plur Verb AND Sing Subj, Sing Verb	4
Plur Subj, Sing Verb → Sing Subj, Sing Verb	6
First Subj, Third Verb → First Subj, First Verb	1
Sing Det, Plur Noun → Sing Det, Sing Noun	6
Sing Det, Plur Noun → Sing Det, Sing Noun AND Plur Det, Plur Noun	3
Sing Det, Plur Noun → Plur Det, Plur Noun	1
Plur Det, Sing Noun → Plur Det, Plur Noun	5
Plur Det, Sing Noun → Plur Det, Plur Noun AND Sing Det, Sing Noun	3
Plur Det, Sing Noun → Sing Det, Sing Noun	1
Count Det, Mass Noun → Mass Det, Mass Noun	2

Figure 7.2: Correcting Agreement Errors in *corpus1*

of either the verb *uses* or the noun *linguist*.[2]

(7.4) *The made-up ungrammaticalities that the **linguist use** will rarely be encountered by the parser.* (A.158)

Another interesting point to note is that in the event of an agreement clash between a singular subject and a plural verb in a sentence, the error can always (in this dataset) be corrected by changing the verb to singular. This is in keeping with a heuristic used by Genthial *et al.* (1994) to automatically correct agreement errors, whereby the error is assumed to be in a word where a letter, e.g. the letter 's', has been omitted. Also, according to this data, the error in an agreement clash is more likely to be located in the rightmost word in the pair of words involved in the clash. However, the dataset is too small for any reliable conclusions to be drawn.

7.2.2 Agreement and Non-Agreement Errors

When an agreement error occurs in a sentence, a constraint between two constituents in the sentence has been violated. Sentences containing agreement errors contain information which is contradictory. For example, in Sentence (7.5) the subject noun phrase is telling us that the agent is the speaker, whereas the verb is telling us that the agent is a third person.

(7.5) ***I awaits*** *your response.* (A.745)

A grammar for English will stipulate that the person and number information provided by the main verb in a sentence must be consistent with the person and number information provided by the subject noun phrase and if this is not the case, an agreement error has occurred.

[2]Another example is Sentence (3.3) from Section 3.4.

Sentences containing other substitution errors, e.g. verb form errors, do not involve an agreement violation since they do not contain contradictory information, e.g. in Sentence (7.6), the information provided by the subject noun phrase *we* and the object noun phrase *Java* does not contradict the information provided by the verb *done*:

(7.6) *Last year we **done** Java.* (A.385)

Where the contradiction does occur, however, is between the grammar and the sentence, because the grammar will expect a finite verb form, e.g. *did*. A similar example is Sentence (7.7) which does involve an error in number in the word *effect* but which does not involve a clash between this word's number information and the number information of another word in the sentence.

(7.7) *Around 1/3 of the slot errors are side **effect** of the card switching errors.* (A.432)

The clash occurs between the grammar's stipulation that a determinerless noun should be plural and the singular determinerless noun in the sentence. Consider also Sentence (7.8):

(7.8) *I may have **mislead** some people with regard to the name of the new socks5 server.* (A.901)

The verb *have* is expecting as a complement a verb phrase whose head verb is in past-participle form. What happens in Sentence (7.8) is that the complement of the verb *have* has as its head the verb *mislead* which is not a past-participle form. This results in a clash between the expectations of the verb *have* and its actual complement. However, what distinguishes this kind of error from an agreement error is that there is no stipulation that the head and one of its complements should share a value for a particular feature.

So, when a substitution error occurs there is a clash between what the grammar says ought to be the case and what is the case.[3] An agreement error is a special kind of substitution error where there is a clash between what the grammar says ought to be the case and what is the case, but this clash arises because the sentences themselves contain two (or more, e.g. det-adj-noun gender agreement in German) pieces of clashing information.

7.3 Feature Structures and Classical Unification

This section introduces the main elements of a typed feature structure system. An understanding of these is necessary in order to understand the extensions to the system which I describe in Sections 7.4 and 7.5.

If feature structures are employed as the representation for linguistic objects, feature structure unification becomes a basic operation in natural language parsing. Carpenter (1992, p.36) defines a basic feature structure (over the set of features, `Feat`, and the set of types, `Type`) to be a tuple consisting of a finite set of nodes, Q, a root node, $\bar{q}$, a partial function δ from a node and a feature to another node, and a total function θ which assigns a type to each node:

Definition 5 ***Feature Structure***

A feature structure *over* `Type` *and* `Feat` *is a tuple* $F = \langle Q, \bar{q}, \theta, \delta \rangle$ *where:*

[3]Exceptions are covert context-sensitive spelling errors where the two confused words are of the same grammatical category. An example from *corpus1* is Sentence (7.9):

(7.9) *Deep correction does, since it involves explaining why some bit of language is the way it is, so prompting* ***leaners*** *to reorganize their cognitive structures.* (A.39)

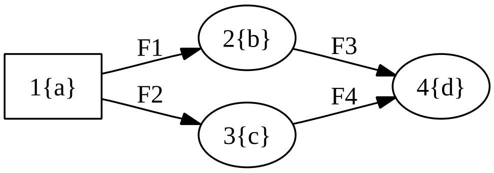


Figure 7.3: A Feature Structure: Graph Notation

$$\boldsymbol{a}\begin{bmatrix} \text{F1:} & \boldsymbol{b}\begin{bmatrix}\text{F3:} & \boxed{1}\ \boldsymbol{d}\end{bmatrix} \\ \text{F2:} & \boldsymbol{c}\begin{bmatrix}\text{F4:} & \boxed{1}\end{bmatrix}\end{bmatrix}$$

Figure 7.4: A Feature Structure: Matrix Notation

- Q *: a finite set of* nodes *rooted at* $\bar{q}$
- $\bar{q} \in Q$ *: the* root *node*
- $\theta : Q \rightarrow \texttt{Type}$ *: a total node* typing *function*
- $\delta : \texttt{Feat} \times Q \rightarrow Q$ *: a partial* feature value *function*

Feature structures can be depicted as directed graphs where the nodes are labelled with a type name and the arcs with the name of a feature. Feature structures can be depicted in a more compact fashion using attribute-value matrix (AVM) notation. Structure-sharing or re-entrancy occurs when two features share the same value. This can be seen in the graph notation when two arcs point to the same node. Tags are used in AVM structures to denote structure-sharing. Two paths whose values are denoted by identical tags point to the same structure. A feature structure is depicted in graph notation in Fig. 7.3 and the same structure is shown in AVM notation in Fig. 7.4. In the graph notation used in this thesis, the type of a node is given in brackets inside the node. A square node, as opposed to an oval one, is used to indicate the root node.

A path is a sequence of features where $\texttt{Path} = \texttt{Feat}*$ represents the set of paths and ϵ represents the empty path. The function δ can be extended as follows so that it deals with paths as well as single features:

- $\delta(\epsilon,\mathrm{q}) = \mathrm{q}$
- $\delta(\mathrm{f}\pi,\mathrm{q}) = \delta(\pi, \delta(\mathrm{f},\mathrm{q}))$

The notation $F@\pi$ is used to denote the value of the feature structure F for some path π for which it is defined. In the feature structure in Fig. 7.3, $F@F1 : F3$ is the atomic feature structure with root node 4. The definition of a path value (Carpenter 1992, p.39) is given below:

Definition 6 ***Path Value***
If $F = \langle Q, \bar{q}, \theta, \delta\rangle$ *and* $\delta(\pi, \bar{q})$ *is defined, then:*

$$F@\pi = \langle Q', \bar{q}', \theta', \delta'\rangle$$

where:

- $\bar{q}' = \delta(\pi, \bar{q})$
- $Q' = \{\delta(\pi', \bar{q}') \mid \pi' \in \texttt{Path}\} = \{\delta(\pi\pi', \bar{q}) \mid \pi' \in \texttt{Path}\}$
- $\theta'(q) = \theta(q)$ *if* $q \in Q'$
- $\delta'(f, q) = \delta(f, q)$ *if* $q \in Q'$

Types are organised into an inheritance hierarchy. A type σ is more specific than another type τ if σ inherits information from τ: τ is more general than σ and is said to subsume σ. This is expressed as $\tau \sqsubseteq \sigma$. If $\tau \sqsubseteq \sigma$, then σ is said to extend τ: τ is a supertype of σ and σ is a subtype of τ. Carpenter (1992) defines a type hierarchy to be a finite bounded complete partial order[4] $\langle \texttt{Type}, \sqsubseteq \rangle$ where $\sqsubseteq$ is the subsumption relation between the types in `Type`. A bounded complete partial order (BCPO) is a partial order where every consistent set of elements has a least upper bound or join. Two elements are consistent or bounded if they have an upper bound. Two types σ, ϕ share an upper bound τ if and only if $\sigma \sqsubseteq \tau$ and $\phi \sqsubseteq \tau$. The join or least upper bound, denoted $\bigsqcup S$, of a set S is defined so that $x \sqsubseteq \bigsqcup S$ for every $x \in S$, and $\bigsqcup S \sqsubseteq y$ for every y such that $x \sqsubseteq y$ for every $x \in S$. The unification of two types in a type hierarchy is defined to be the least upper bound of the types. The unification of a set of types is not defined if the types are not consistent. Carpenter (1992, p.12-13) defines a universal type called $\perp$ such that $\perp \sqsubseteq \tau$ for every type τ. An element $\top$ can be added to the type hierarchy to represent inconsistency.[5] Fig. 7.5 shows a simple hierarchy of types, and Fig 7.6 shows some sample joins defined over this hierarchy. The type hierarchy in Fig. 7.7 is an example of a hierarchy which is not a bounded complete partial order because the types `detnumber` and `nounnumber` are consistent but have no unique least upper bound. In these hierarchies the symbol *bot* represents the most general type $\perp$.

Carpenter (1992, p.41) defines a subsumption relation between feature structures: the feature structure $F1$ subsumes the feature structure $F2$ if all the nodes in $F1$ can be mapped onto the nodes in $F2$ such that the root node is preserved, the types of all the nodes in $F1$ subsume the corresponding types in $F2$, and all the transitions in $F1$ are present in $F2$.

Definition 7 ***Feature Structure Subsumption***

$F = \langle Q, \bar{q}, \theta, \delta \rangle$ subsumes $F' = \langle Q', \bar{q}', \theta', \delta' \rangle, F \sqsubseteq F'$, *if and only if there is*

[4] A *partial order* is a binary relation which is reflexive, transitive and anti-symmetric.

[5] In Fouvry (2003b) and Copestake (2002), the type hierarchy is defined in the opposite direction, and $\top$ is the most general type.

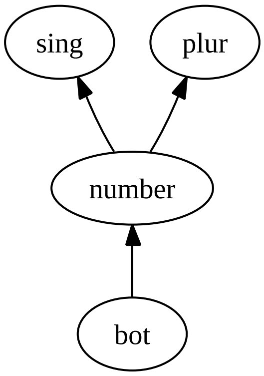


Figure 7.5: A Simple Type Hierarchy

$$\texttt{number} \sqcup \texttt{sing} = \texttt{sing}$$

$$\texttt{sing} \sqcup \texttt{plur} \text{ is undefined}$$

$$\perp \sqcup \texttt{plur} = \texttt{plur}$$

Figure 7.6: Some Sample Joins defined over the Hierarchy in Fig. 7.5

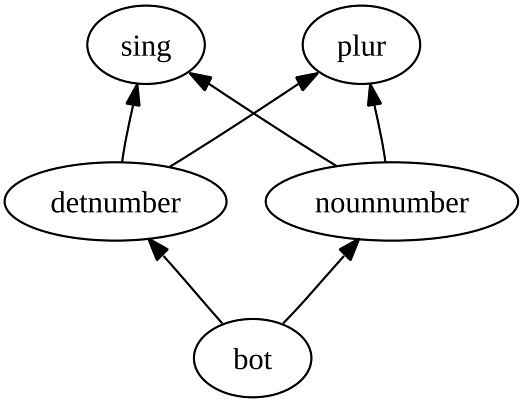


Figure 7.7: A Type Hierarchy which is not a BCPO

a total function $h : Q \to Q'$*, called a* morphism *such that:*

- $h(\bar{q}) = \bar{q}'$
- $\theta(q) \sqsubseteq \theta'(h(q))$ *for every* $q \in Q$
- $h(\delta(f, q)) = \delta'(f, h(q))$ *for every* $q \in Q$ *and feature* f *such that* $\delta(f, q)$ *is defined*

The unification of two feature structures is the least upper bound of the two structures in $\langle \mathcal{F}, \sqsubseteq \rangle$ where $\mathcal{F}$ is the set of feature structures and $\sqsubseteq$ is subsumption between feature structures. The unification of two feature structures takes the information contained in the input structures and combines this information (and only this information) in the output feature structure. Nodes in the input feature structures which are required to be unified are organised into equivalence classes.[6] The least upper bound of the types of all the nodes in an equivalence class is found, and if there is no least upper bound for the types of the nodes in some equivalence class, unification fails.

Definition 8 ***Feature Structure Unification***
Suppose $F_1, F_2 \in \mathcal{F}$ *are the feature structures* $\langle Q_1, \bar{q}_1, \theta_1, \delta_1 \rangle$ *and* $\langle Q_2, \bar{q}_2, \theta_2, \delta_2 \rangle$ *such that* $Q_1 \cap Q_2 = \emptyset$*. An equivalence relation* $\bowtie$ *on* $Q_1 \cup Q_2$ *is defined as the least equivalence relation such that:*

- $\bar{q}_1 \bowtie \bar{q}_2$

[6]An *equivalence relation* is a relation which is transitive, reflexive and symmetric. If $\equiv$ is an equivalence relation over X, then the following notations are used for the *equivalence classes* into which X can be partitioned by $\equiv$:

- $[x]_{\equiv} = \{y \in X | y \equiv x\}$: the equivalence class of x
- $X/_{\equiv} = \{[x]_{\equiv} | x \in X\}$: the quotient set of X modulo $\equiv$.

(Carpenter 1992, p.46)

- $\delta_1(f, q_1) \bowtie \delta_2(f, q_2)$ *if both are defined and* $q_1 \bowtie q_2$

The unification *of* F_1 *and* F_2 *is then defined to be:* $F_3 = \langle Q_3, \bar{q}_3, \theta_3, \delta_3 \rangle$ *where:*

- $Q_3 = (Q_1 \cup Q_2)/_{\bowtie}$
- $\bar{q}_3 = [\bar{q}_1]_{\bowtie}$
- $\theta_3([q]_{\bowtie}) = \bigsqcup\{(\theta_1 \cup \theta_2)(q')\ \text{,}\ q' \bowtie q\}$
- $\delta_3(f, [q]_{\bowtie}) = \begin{cases} [(\delta_1 \cup \delta_2)(f, q)]_{\bowtie} if (\delta_1 \cup \delta_2)(f, q) is\ defined \\ undefined\ otherwise \end{cases}$

if all the joins in the definition of θ_3 *exist.* F_3 *is undefined otherwise.*

Carpenter (1992, p.52-53) defines a description language for talking about feature structures and provides a set of axioms which capture the logical behaviour of feature structures. The only way for inconsistency to result in Carpenter's basic system is if an attempt is made to unify two inconsistent types. This behaviour is captured by the following axiom, in which σ and τ are types in the set of types `Type`:

$$\sigma \wedge \tau \Leftrightarrow \top \quad [if\ \sigma \sqcup \tau\ does\ not\ exist]$$

The "Top Smashing" axiom demonstrates how inconsistency is propagated:

$$\pi : \top \Leftrightarrow \top$$

π indicates a path or sequence of features, and $\pi : \top$ indicates that the path π has the value $\top$, or that the node at the end of the path π is labelled with the type $\top$. So, if a path in a feature structure has a value labelled by an inconsistent type, the type of the whole feature structure is inconsistent.

A traditional unification-based parser will not parse, for example, the noun phrase *this women* because unification will fail when an attempt is made to find a least upper bound for the inconsistent types `sing` and `plur`. What we would like a parser to return is that *this women* is a noun phrase, its head is the plural noun *women*[7], its complement is the singular determiner *this* and there is a number mismatch between the determiner and the noun. In Section 7.4, an operation called *robust agreement unification* is outlined which allows this information to be returned.

7.4 Agreement Errors and Robust Agreement Unification

In this section, I explain, using examples, how the inconsistent information involved in an agreement clash can be represented in a feature structure and how the unification operation can produce feature structures containing such information. Before my approach is presented, other approaches which attempt to encode inconsistent information in a feature structure are discussed.

7.4.1 Inconsistent Agreement Identity

A lexical entry displaying some linguistic properties of the word *this* is illustrated in Fig. 7.8. This feature structure states that *this* is a singular determiner. The feature structure in Fig. 7.9 provides us with similar information for the plural noun *women*. Fig. 7.10 shows a feature structure which states that the head of a noun phrase is a noun, its complement is a

[7] Or the determiner *this*, depending on one's view of noun phrase syntax.

$$\textbf{\textit{det}}\begin{bmatrix} \text{NUM: } \textbf{\textit{sing}} \end{bmatrix}$$

Figure 7.8: The Determiner *this*

determiner and both these constituents share a value for their `NUM` feature. Here, the type `number` is the general type subsuming the types `sing` and `plur`.

When a parser which employs classical unification encounters the noun phrase *this women*, it will attempt to build a feature structure for the whole phrase by unifying the lexical entry for *this* (Fig. 7.8), the lexical entry for *women* (Fig. 7.9) and the feature structure representing a noun phrase (Fig. 7.10). Assuming that the first thing that the parser does is access the lexical entry for the determiner *this*, it will attempt to unify this with the value of the `COMP-DTR` feature in the noun phrase feature structure, yielding the feature structure shown in Fig. 7.11. At this stage, no inconsistency has occurred, and the parser is in the process of building up a noun phrase which is singular. The problem occurs, however, when the lexical entry for the word *women* is accessed – an attempt is made to unify this feature structure into the existing structure (the value of the `HEAD-DTR` feature in Fig. 7.11), yielding the inconsistent feature structure in Fig. 7.12. "Top smashing" ensures that this structure is actually equivalent to $\top$, the inconsistent type. Therefore, a parser employing classical unification will return the unhelpful value $\top$ when confronted with a sentence containing the phrase *this women.* This is clearly an undesirable outcome since we want natural language parsers to recover as much structure as possible from sentences containing errors.

To overcome this problem, Vogel and Cooper (1995) propose a relaxed approach to unification which results in "mildly inconsistent" feature struc-

$$\boldsymbol{noun}\begin{bmatrix} \text{NUM: } \boldsymbol{plur} \end{bmatrix}$$

Figure 7.9: The Noun *women*

$$\boldsymbol{np}\begin{bmatrix} \text{HEAD-DTR:} & \boldsymbol{noun}\begin{bmatrix} \text{NUM: } \boxed{1}\ \boldsymbol{number} \end{bmatrix} \\ \text{COMP-DTR:} & \boldsymbol{det}\begin{bmatrix} \text{NUM: } \boxed{1} \end{bmatrix} \end{bmatrix}$$

Figure 7.10: Determiner-Noun Number Agreement

$$\boldsymbol{np}\begin{bmatrix} \text{HEAD-DTR:} & \boldsymbol{noun}\begin{bmatrix} \text{NUM: } \boxed{1} \end{bmatrix} \\ \text{COMP-DTR:} & \boldsymbol{det}\begin{bmatrix} \text{NUM: } \boxed{1}\,\boldsymbol{sing} \end{bmatrix} \end{bmatrix}$$

Figure 7.11: *this women*: First Stage of Classical Unification

$$np\begin{bmatrix} \text{HEAD-DTR:} & \boldsymbol{noun}\begin{bmatrix} \text{NUM: } \boxed{1}\top \end{bmatrix} \\ \text{COMP-DTR:} & \boldsymbol{det}\begin{bmatrix} \text{NUM: } \boxed{1} \end{bmatrix} \end{bmatrix} = \top$$

Figure 7.12: The NP *this women*: Classical Unification

$$\boldsymbol{np}\begin{bmatrix} \text{HEAD-DTR:} & \boldsymbol{noun}\begin{bmatrix} \text{NUM: } \boxed{1}\ \top(sing, plur) \end{bmatrix} \\ \text{COMP-DTR:} & \boldsymbol{det}\begin{bmatrix} \text{NUM: } \boxed{1} \end{bmatrix} \end{bmatrix} \neq \top$$

Figure 7.13: The NP *this women*: Vogel and Cooper, 1995

tures. An example of a "mildly inconsistent" feature structure is given in Fig. 7.13. In their approach, the inconsistent types `sing` and `plur` are actually unified and the result is the type $\top$(`sing`, `plur`). This means that the number value of the whole noun phrase is both singular and plural. So, in their approach, the feature structures contain inconsistent information, without this resulting in a breakdown in the unification process. This is quite a radical approach which would need to be controlled so that feature structures contain the minimal possible amount of inconsistent information. A disadvantage is that the number information in the feature structure for the noun phrase *this women* will be the same as the number information in the feature structure for the equally ungrammatical, yet different, noun phrase *these woman*.

Fouvry (2003b) provides a fully worked-out account of how robustness can be integrated into Carpenter's typed feature structure system. A brief sketch of this work was presented in the discussion of constraint relaxation in Section 4.3.1. There are essentially two parts to this work: the first part is the extension of the type hierarchy so that any pair of inconsistent types has a join. This means that type unification will never fail. An example of an extended robust type hierarchy is shown in Fig. 7.14: this is an extension of the type hierarchy from Fig. 7.5. Fig 7.15 shows the feature structure for the noun phrase *this women* which would result given the type hierarchy in Fig. 7.14. Note that it is similar to the kind of feature structure produced

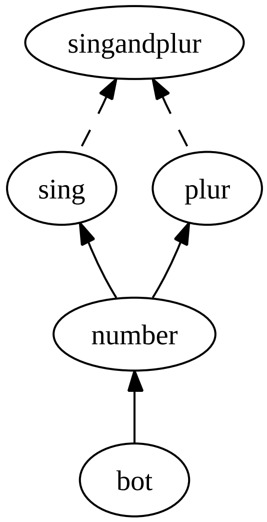


Figure 7.14: Extending the Type Hierarchy in Fig. 7.5: Fouvry, 2000, 2003

$$\boldsymbol{np}\begin{bmatrix} \text{HEAD-DTR:} & \boldsymbol{noun}\begin{bmatrix} \text{NUM: } \boxed{1}\ \textbf{singandplur} \end{bmatrix} \\ \text{COMP-DTR:} & \boldsymbol{det}\begin{bmatrix} \text{NUM: } \boxed{1} \end{bmatrix} \end{bmatrix} \neq \top$$

Figure 7.15: The NP *this women*: Fouvry, 2000,2003

by Vogel and Cooper's approach although in Fouvry's approach the value of the `NUM` feature is a well-defined type according to the type hierarchy.

The second part of Fouvry's work controls the robustness introduced by the first part by associating weights with feature structures. Each feature structure is assigned a weight which is the sum of the weights on the types on each node in the feature structure. A more specific type will have a higher weight than a type that subsumes it. These weights are essentially keeping track of information loss: this is because when a type value from the grammar (called the *reference* value, this is assumed to be correct) is unified with an inconsistent type in the input (the *ground* value), the inconsistent information in the *ground* value is discarded. For example, given the type hierarchy in Fig. 7.5, assume that the type `sing` has weight 2, the type `plur` has weight 2 and the type `number` has weight 1. Assume that `plur` is the reference value and `sing` is the ground value[8], the ground value is changed to be the generalisation of the inconsistent types `sing` and `plur` which is the type `number`. The type `number` has a lower weight than the original type `sing` so information, in this case the value 1, has been lost. However, in the case of agreement errors, the inconsistent type `singandplur` is not discarded because the types `sing` and `plur` are both *ground* values. In doing this, Fouvry is making the same distinction I made in Section 7.2.2 between agreement errors where there is a clash between two words in the input sentence, and other errors where there is a clash between the input and the expectations of the grammar alone. The information loss measure can be used for ranking parses, and for establishing a threshold to limit the amount of parses that are produced for any input sentence. Unfortunately, as Fouvry

[8]This might occur, for example, in a grammar rule for bare plural noun phrases, which stipulates that the head noun must have a plural number value.

himself admits (Fouvry 2003a), his robust typed feature structure system has yet to be evaluated on real ungrammatical text. It would certainly be interesting to investigate whether this information loss measure would select parses for the *corpus1* sentences which reflect their corrections in *corpus2*.

Reuer (2000) describes a constraint relaxation approach to unification-based parsing which forms part of a computer-aided language learning system. This work is carried out within the linguistic framework of Lexical Functional Grammar. In Reuer's approach, two feature structures or f-structures containing inconsistent values for the same feature can be unified producing a set of two feature structures which store the inconsistent values in an ERR feature. This feature is then used to diagnose the error and provide feedback to the user. The relaxed unification is limited by specifying in advance what kinds of inconsistent values can be unified. Fouvry (2003b) points out that if Reuer's approach was adopted in a totally well-typed system such as Carpenter's where the features appropriate for a type must always be present in a feature structure of that type, the feature ERR would always have to be present. Another relaxed unification approach in a system other than Carpenter's is described by Kim (1994). In this approach, inconsistent atomic feature values are unified to produce a *disjunction* of values.

The approach I am proposing is quite conservative since it does not require that two inconsistent nodes actually be unified. I propose an operation called *robust agreement unification* which will produce the structure given in Fig. 7.16. The symbol $\leftrightarrow_\top$ indicates a relation which I will call *inconsistent agreement identity*, and if two nodes in a feature structure are related by *inconsistent agreement identity*, this means that they would be structure-shared (i.e they would be the same node) were it not for the fact that they are labelled with inconsistent types. The advantage of this is that it allows

$$\boldsymbol{np}\begin{bmatrix} \text{HEAD-DTR:} & \boldsymbol{noun}\begin{bmatrix} \text{NUM:} & \boxed{1}\ \boldsymbol{plur} \end{bmatrix} \\ \text{COMP-DTR:} & \boldsymbol{det}\begin{bmatrix} \text{NUM:} & \boxed{2}\ \boldsymbol{sing} \end{bmatrix} \\ \boxed{1} \leftrightarrow_{\top} \boxed{2} & \end{bmatrix} \neq \top$$

Figure 7.16: The NP *this women*: Robust Agreement Unification

the information conveyed by *this women* to be represented whilst still informing us that there is a number agreement violation between the noun and the determiner.

7.4.2 Robust Agreement Unification in Action

In order to illustrate how exactly robust agreement unification operates, I will work step-by-step through another example of a number agreement violation between a determiner and a noun – the noun phrase *these woman.* Fig. 7.17 shows the relevant parts of the lexical entry for the word *these*, and the feature structure in Fig. 7.18 the relevant parts of the lexical entry for the word *woman.* We have already seen the feature structure which enforces determiner-noun number agreement in Fig. 7.10 - this is repeated for convenience in Fig. 7.19

$$\boldsymbol{det}\begin{bmatrix} \text{NUM: } \boldsymbol{plur} \end{bmatrix}$$

Figure 7.17: The Determiner *these*

$$\boldsymbol{noun}\begin{bmatrix} \text{NUM: } \boldsymbol{sing} \end{bmatrix}$$

Figure 7.18: The Noun *woman*

$$\boldsymbol{np}\begin{bmatrix} \text{HEAD-DTR:} & \boldsymbol{noun}\begin{bmatrix}\text{NUM: } \boxed{1}\ \boldsymbol{number}\end{bmatrix} \\ & \\ \text{COMP-DTR:} & \boldsymbol{det}\begin{bmatrix}\text{NUM: } \boxed{1}\end{bmatrix}\end{bmatrix}$$

Figure 7.19: Determiner-Noun Number Agreement

If robust agreement unification is used to unify the feature structure for the determiner *these* (Fig. 7.17) with the value of the `COMP-DTR` feature in the feature structure for the whole noun phrase (Fig. 7.19) the structure shown in Fig. 7.20 is returned. Here the type `plur` unifies with the type `number` to yield the more specific type `plur` as a result. Classically unifying these two feature structures would bring the same result, so robust agreement unification behaves in the same way as classical unification when there is no type clash in the feature structures to be unified.

Finally if we robustly unify the value of the `HEAD-DTR` feature in the feature structure in Fig. 7.20 with the feature structure for the noun *woman* (Fig. 7.18), we get as a result the feature structure shown in Fig. 7.21.

This is what happens in order to obtain this result:

1. A type clash occurs between the type `sing` in the feature structure in Fig. 7.18 and the type `plur` in the feature structure in Fig. 7.20.

2. The structure-sharing constraint between the paths `HEAD-DTR:NUM` and `COMP-DTR:NUM` in the feature structure in Fig. 7.20 is relaxed.

3. The two paths which were to share a value now have separate values, the value of the path `HEAD-DTR:NUM` arising from the current unification with the feature structure for the noun *woman* (see Fig. 7.18), and the value for the path `COMP-DTR:NUM` arising from the previous unification with the feature structure for the determiner *these* (see Fig. 7.17).

$$\boldsymbol{np}\begin{bmatrix} \text{HEAD-DTR:} & \boldsymbol{noun}\begin{bmatrix}\text{NUM: } \boxed{1}\end{bmatrix} \\ \\ \text{COMP-DTR:} & \boldsymbol{det}\begin{bmatrix}\text{NUM: } \boxed{1}\ \boldsymbol{plur}\end{bmatrix}\end{bmatrix}$$

Figure 7.20: The Unification of the Feature Structures in Fig 7.17 and Fig. 7.19

$$\boldsymbol{np}\begin{bmatrix} \text{HEAD-DTR:} & \boldsymbol{noun}\begin{bmatrix}\text{NUM: } \boxed{1}\ \boldsymbol{sing}\end{bmatrix} \\ \text{COMP-DTR:} & \boldsymbol{det}\begin{bmatrix}\text{NUM: } \boxed{2}\ \boldsymbol{plur}\end{bmatrix} \\ \boxed{1} \leftrightarrow_{\top} \boxed{2} & \end{bmatrix}$$

Figure 7.21: The NP *these woman*: Robust Agreement Unification

4. The two nodes labelled with the types `sing` and `plur` are related by the *inconsistent agreement identity* relationship, represented by the symbol $\leftrightarrow_{\top}$.

It is always a structure sharing constraint that is relaxed, to be replaced by *inconsistent agreement identity.* Note also that the feature structure for the ungrammatical phrase *these woman* (Fig. 7.21) differs from the feature structure for the ungrammatical phrase *this women* (Fig. 7.16).

7.4.3 Complications

Propogating Inconsistent Values

If the number information in a phrase is inherited from the head of the phrase, this principle will also apply in the event of an agreement error. This situation is illustrated for the noun phrase *these woman* in the feature structure in Fig. 7.22, assuming again that the head of a noun phrase is

$$
\boldsymbol{np}\begin{bmatrix} \text{HEAD-DTR:} & \boldsymbol{noun}\begin{bmatrix}\text{NUM:} & \boxed{1}\ \boldsymbol{sing}\end{bmatrix} \\ \text{COMP-DTR:} & \boldsymbol{det}\begin{bmatrix}\text{NUM:} & \boxed{2}\ \boldsymbol{plur}\end{bmatrix} \\ \text{NUM:} \ \boxed{1} & \\ \boxed{1} \leftrightarrow_{\top} \boxed{2} & \end{bmatrix}
$$

Figure 7.22: The NP *these woman*: Propagating Values

the noun. The means that there will be one agreement clash for a sentence such as (7.10), the agreement clash within the subject noun phrase, but two agreement clashes for a sentence such as (7.11), one within the subject noun phrase and the second between the subject noun phrase and the verb:

(7.10) *These woman walks.*

(7.11) *These woman walk.*

The feature structure in Fig. 7.23 represents the information in Sentence (7.10), and the structure in Fig. 7.24 the information in Sentence (7.11). According to the table of agreement error corrections in Fig. 7.2, the noun is more likely to be erroneous in a noun phrase agreement error, and so it might be a better reflection of reality if the determiner number information was propagated up to the noun phrase level rather than the noun number information. This would be the case in a grammar which treats determiners as heads in a noun phrase, rather than nouns.

Structure Sharing Constraint Relaxation and Inconsistent Agreement Identity

Consider the feature structure for the pronoun *I* given in Fig. 7.25[9]. This feature structure differs from the feature structures given in Fig. 7.9, Fig. 7.8,

[9] Assume that the type `pronoun` inherits from more general type `noun`.

$$
\boldsymbol{sen}\left[\begin{array}{ll}
\text{HEAD-DTR:} & \boldsymbol{vp}\left[\begin{array}{ll}\text{HEAD-DTR:} & \boldsymbol{verb}\left[\begin{array}{l}\text{NUM: } \boxed{1}\ \boldsymbol{sing}\end{array}\right]\end{array}\right] \\
\text{COMP-DTR:} & \boldsymbol{np}\left[\begin{array}{ll}
\text{HEAD-DTR:} & \boldsymbol{noun}\left[\begin{array}{l}\text{NUM: } \boxed{1}\ \boldsymbol{sing}\end{array}\right] \\
\text{COMP-DTR:} & \boldsymbol{det}\left[\begin{array}{l}\text{NUM: } \boxed{2}\ \boldsymbol{plur}\end{array}\right] \\
\text{NUM: } \boxed{1} & \\
\boxed{1} \leftrightarrow_{\top} \boxed{2} &
\end{array}\right]
\end{array}\right]
$$

Figure 7.23: *These woman walks*

$$
\boldsymbol{sen}\left[\begin{array}{ll}
\text{HEAD-DTR:} & \boldsymbol{vp}\left[\begin{array}{ll}\text{HEAD-DTR:} & \boldsymbol{verb}\left[\begin{array}{l}\text{NUM: } \boxed{3}\ \boldsymbol{plur}\end{array}\right]\end{array}\right] \\
\text{COMP-DTR:} & \boldsymbol{np}\left[\begin{array}{ll}
\text{HEAD-DTR:} & \boldsymbol{noun}\left[\begin{array}{l}\text{NUM: } \boxed{1}\ \boldsymbol{sing}\end{array}\right] \\
\text{COMP-DTR:} & \boldsymbol{det}\left[\begin{array}{l}\text{NUM: } \boxed{2}\ \boldsymbol{plur}\end{array}\right] \\
\text{NUM: } \boxed{1} & \\
\boxed{1} \leftrightarrow_{\top} \boxed{2} &
\end{array}\right] \\
\boxed{1} \leftrightarrow_{\top} \boxed{3} &
\end{array}\right]
$$

Figure 7.24: *These woman walk*

Fig. 7.18 and Fig. 7.17 since it contains an extra feature `INDEX` whose value is a referential index for which the agreement features `PER` and `NUM` are appropriate. Fig. 7.26 represents a feature structure for the third person singular verb *walks* and Fig. 7.27 illustrates a feature structure which enforces person and number agreement between a verb and its subject.

Consider the ungrammatical fragment *I walks* which contains a person agreement violation between the verb *walks* and its subject *I*. For the reasons outlined in section 7.4.1, classical unification will fail when confronted with *I walks*. Robust agreement unification will produce the structure shown in Fig. 7.28. Note that this example is different to the examples seen previously since the nodes which are related by inconsistent agreement identity (i.e. those which are values of the feature `PER` and which are labelled by the inconsistent types `1st` and `3rd`) are not explicitly required to be structure shared – instead their parent nodes are explicitly required to be structure-shared (Fig. 7.27) and it is this structure-sharing constraint which must be relaxed. This has the effect that none of their children are structure-shared, even those whose types are consistent (in this example, the nodes labelled by the type `sing`).

So, in the approach outlined here, the occurrence of an agreement error always involves the relaxation of the structure-sharing constraint which enforces the agreement, along with the relating of the two nodes involved in the agreement clash in an *inconsistent agreement identity*. However, as the above example illustrates, these two actions are not guaranteed to take place at an identical point in the feature structure.

$$\textbf{\textit{pronoun}}\begin{bmatrix}\text{INDEX:} & \textbf{\textit{index}}\begin{bmatrix}\text{NUM: } \textbf{\textit{sing}}\\ \text{PER: } \textbf{\textit{first}}\end{bmatrix}\end{bmatrix}$$

Figure 7.25: The Pronoun *I*

$$\textbf{\textit{verb}}\begin{bmatrix}\text{INDEX:} & \textbf{\textit{index}}\begin{bmatrix}\text{NUM: } \textbf{\textit{sing}}\\ \text{PER: } \textbf{\textit{3rd}}\end{bmatrix}\end{bmatrix}$$

Figure 7.26: The Verb *walks*

$$\textbf{\textit{sen}}\begin{bmatrix}\text{HEAD-DTR:} & \textbf{\textit{vp}}\begin{bmatrix}\text{HEAD-DTR:} & \textbf{\textit{verb}}\begin{bmatrix}\text{INDEX: } \boxed{1}\end{bmatrix}\end{bmatrix}\\ \text{COMP-DTR:} & \textbf{\textit{np}}\begin{bmatrix}\text{HEAD-DTR:} & \textbf{\textit{noun}}\begin{bmatrix}\text{INDEX: } \boxed{1}\end{bmatrix}\end{bmatrix}\end{bmatrix}$$

Figure 7.27: Subject-Verb Agreement

$$\textbf{\textit{sen}}\begin{bmatrix}\text{HEAD-DTR:} & \textbf{\textit{vp}}\begin{bmatrix}\text{HEAD-DTR:} & \textbf{\textit{verb}}\begin{bmatrix}\text{INDEX:} & \textbf{\textit{index}}\begin{bmatrix}\text{PER: } \boxed{2}\,\textbf{\textit{3rd}}\\ \text{NUM: } \textbf{\textit{sing}}\end{bmatrix}\end{bmatrix}\end{bmatrix}\\ \text{COMP-DTR:} & \textbf{\textit{np}}\begin{bmatrix}\text{HEAD-DTR:} & \textbf{\textit{noun}}\begin{bmatrix}\text{INDEX:} & \textbf{\textit{index}}\begin{bmatrix}\text{PER: } \boxed{3}\,\textbf{\textit{1st}}\\ \text{NUM: } \textbf{\textit{sing}}\end{bmatrix}\end{bmatrix}\end{bmatrix}\end{bmatrix}$$

$$\boxed{2} \leftrightarrow_{\top} \boxed{3}$$

Figure 7.28: *I walks*:Robust Agreement Unification

7.5 Robust Agreement Feature Structures and Robust Agreement Unification Defined

In this section, the idea of robust agreement unification introduced in the previous section is formally defined.

7.5.1 Robust Agreement Feature Structures

A definition of the structures that result from robust agreement unification, i.e. *robust agreement feature structures* (RAFs), is provided. This definition is based on Carpenter's original feature structure definition (see Def. 5 in Section 7.3).

Definition 9 ***Robust Agreement Feature Structure***
A robust agreement feature structure *is a quintuple* $F = \langle Q, \bar{q}, \theta, \delta, \leftrightarrow_\top \rangle$, *where* $F = \langle Q, \bar{q}, \theta, \delta \rangle$ *is a feature structure and where* $\leftrightarrow_\top \subseteq Q \times Q$ *is an anti-reflexive and symmetric* inconsistent agreement identity *relation.*

The robust agreement feature structure illustrated in Fig. 7.21 is:

$Q = \{q1, q2, q3, q4, q5\}$
$\bar{q} = q1$
$\delta(head - dtr, q1) = q2,\ \delta(comp - dtr, q1) = q3,\ \delta(num, q2) = q4,\ \delta(num, q3) = q5$
$\theta(q1) = np,\ \theta(q2) = noun,\ \theta(q3) = det,\ \theta(q4) = sing,\ \theta(q5) = plur$
$q4 \leftrightarrow_\top q5$

A RAF F subsumes another RAF F' iff the non-robust part of F subsumes the non-robust part of F' and the inconsistent agreement identity relationships which hold between nodes in F hold between equivalent nodes in F':

Definition 10 ***Robust Agreement Feature Structure Subsumption*** $F = \langle Q, \bar{q}, \theta, \delta \leftrightarrow_{\top} \rangle$ subsumes $F' = \langle Q', \bar{q}', \theta', \delta' \leftrightarrow'_{\top} \rangle, F \sqsubseteq F'$, *if and only if there is a total function* $h : Q \to Q'$, *called a* morphism *such that:*

- $h(\bar{q}) = \bar{q}'$
- $\theta(q) \sqsubseteq \theta'(h(q))$ *for every* $q \in Q$
- $h(\delta(f, q)) = \delta'(f, h(q))$ *for every* $q \in Q$ *and feature* f *such that* $\delta(f, q)$ *is defined*
- *if* $q \leftrightarrow_{\top} q'$ *then* $h(q) \leftrightarrow_{\top} h(q')$

Carpenter's description language is extended to include inconsistent path equations: the symbol $\doteq_{\top}$ is used to equate two paths π_1, π_2 in an *inconsistent agreement equality.* The satisfaction relation between robust agreement feature structures and descriptions containing inconsistent agreement path equations is defined as follows:

$$\langle Q, \bar{q}, \theta, \delta, \leftrightarrow_{\top} \rangle \models \pi_1 \doteq_{\top} \pi_2 \text{ if and only if } \delta(\pi_1, \bar{q}) \leftrightarrow_{\top} \delta(\pi_2, \bar{q})$$

The definition of a robust agreement feature structure has much in common with Carpenter's definition of an *inequated feature structure* (Carpenter 1992, p.112), since in both definitions the original feature structure tuple is augmented with an anti-reflexive and symmetric relation. Carpenter's approach to negation is conservative just as my approach to inconsistency is. According to Carpenter (1992, p.112), "an inequated feature structure consists of an ordinary feature structure with an additional inequality relation between the nodes which are known to be distinct." Similarly, a robust agreement feature structure consists of an ordinary feature structure with an additional *inconsistent agreement identity* relation between the nodes which are known to be inconsistent.

7.5.2 Robust Agreement Unification

Now that a robust agreement feature structure has been defined, the operation of robust agreement unification which operates on and produces robust agreement feature structures can be defined. This definition make uses of a notion called *alphabetic variance* which can be defined as in Def. 11 (Carpenter 1992, p.43). Two RAFs are alphabetic variants if they contain exactly the same information.

Definition 11 ***Alphabetic Variants***
If F and F' are RAFs such that $F \subseteq F'$ and $F' \subseteq F$, then we write $F \sim F'$ and say that they are alphabetic variants.

The definition of robust agreement unification uses as a basis the definition of classical unification, Def. 8, given in Section 7.3. It assumes that any pairs of paths which implement grammatical agreement in one of the input feature structures is known in advance. The pair of paths which implement determiner-noun agreement in the feature structure in Fig. 7.19, for example, are the path `HEAD-DTR|NUM` and the path `COMP-DTR|NUM`.

Definition 12 ***Robust Agreement Unification***
Suppose F_1, F_2 are the RAFs $\langle Q_1, \bar{q}_1, \theta_1, \delta_1, \leftrightarrow^1_\top \rangle$ and $\langle Q_2, \bar{q}_2, \theta_2, \delta_2, \leftrightarrow^2_\top \rangle$ such that $Q_1 \cap Q_2 = \emptyset$. Agree is the set of non-empty agreement path pairs in F_1. An equivalence relation $\bowtie$ on $Q_1 \cup Q_2$ is defined as the least equivalence relation such that:

- $\bar{q}_1 \bowtie \bar{q}_2$
- $\delta_1(f, q_1) \bowtie \delta_2(f, q_2)$ *if both are defined and* $q_1 \bowtie q_2$

The robust agreement unification *of* F_1 *and* F_2 *is then defined to be:* $F_3 = \langle Q_3, \bar{q}_3, \theta_3, \delta_3, \leftrightarrow^3_\top \rangle$ *where:*

- $Q_3 = (Q_1 \cup Q_2)/_{\bowtie}$
- $\bar{q}_3 = [\bar{q}_1]_{\bowtie}$
- $\theta_3([q]_{\bowtie}) = \bigsqcup\{(\theta_1 \cup \theta_2)(q') \,\text{,}\, q' \bowtie q\}$
- $\delta_3(f, [q]_{\bowtie}) = \begin{cases} [(\delta_1 \cup \delta_2)(f, q)]_{\bowtie} if (\delta_1 \cup \delta_2)(f, q) is\ defined \\ \textit{undefined otherwise} \end{cases}$
- $[q]_{\bowtie} \leftrightarrow^3_\top [q']_{\bowtie}$ *if* $q'' \leftrightarrow^1_\top q'''$ *or* $q'' \leftrightarrow^2_\top q'''$ *for some* $q'' \bowtie q$ *and* $q''' \bowtie q'$

If conditions (1) and (2) hold:
$\exists q_1 \in Q_1, q_2 \in Q_2$

1. $q_1 \bowtie q_2$ *and* $\theta_1(q_1) \sqcup \theta_2(q_2) = \top$ *(a type clash has occurred between two nodes in the same equivalence class)*
2. *Conditions (2a) and (2b) hold (an agreement error has occurred):*
   $\exists \pi, \pi', \pi'', \pi''', \sigma \in$ ***Feat****
   (a) $\delta_1(\pi, \bar{q}_1) \in [q_1]_{\bowtie} \wedge \delta_1(\pi', \bar{q}_1) \in [q_1]_{\bowtie} \wedge \delta_2(\pi', \bar{q}_2) \in [q_1]_{\bowtie}$
   (b) $(\pi'', \pi''') \in Agree \wedge \delta_1(\pi''\sigma, \bar{q}_1) = \delta_1(\pi, \bar{q}_1) \wedge \delta_1(\pi'''\sigma, \bar{q}_1) = \delta_1(\pi', \bar{q}_1)$ *where* σ *is a possibly empty path.*

Then the following changes are made to F_3*:*

1. *A new feature structure* $F_4 = \langle Q_4, \bar{q}_4, \theta_4, \delta_4, \leftrightarrow^4_\top \rangle$ *is defined such that* $F_4 \sim F_2@\pi'''$

- $F_3@\pi''' \neq [\delta_1(\pi''', \bar{q}_1)]_{\bowtie}$
- $F_3@\pi''' = F_4$

(the feature structure at the end of path π''' in F_3 now has the value F_4)

2. $F_3@\pi'''\sigma \leftrightarrow^3_\top [q_1]_{\bowtie}$ *(the nodes involved in the type clash are related by inconsistent agreement identity)*

Otherwise F_3 is undefined. The symbol $\bigsqcup_{\leftrightarrow}$ is used to represent the robust agreement unification operator.

An example is required to make this definition clearer. Assume that F_1 is the robust agreement feature structure illustrated in Fig. 7.29 and F_2 is the robust agreement feature structure illustrated in Fig. 7.30. *Agree*, the set of agreement path pairs in F_1 is $\{(\texttt{HEAD-DTR}|\texttt{NUM}), (\texttt{COMP-DTR}|\texttt{NUM})\}$. The individual values in F_1 and F_2 are as follows:

$Q_1 = \{q_a, q_b, q_c, q_d\}$

$\bar{q}_1 = q_a$

$\delta_1(head - dtr, q_a) = q_b,\ \delta_1(comp - dtr, q_a) = q_c,\ \delta_1(num, q_b) = q_d,\ \delta_1(num, q_c) = q_d$

$\theta_1(q_a) = np,\ \theta_1(q_b) = noun,\ \theta_1(q_c) = det,\ \theta_1(q_d) = plur$

$Q_2 = \{q_e, q_f, q_g\}$

$\bar{q}_2 = q_e$

$\delta_2(head - dtr, q_e) = q_f,\ \delta_2(num, q_f) = q_g$

$\theta_2(q_e) = np,\ \theta_2(q_f) = noun,\ \theta_2(q_g) = sing$

Note that no inconsistent agreement identity relationships hold between any pair of nodes in Q_1 or Q_2. The first thing that is done in the robust agreement unification is to organize the nodes in F_1 and F_2 into equivalence classes. The result of this is as follows:

1. $q_a \bowtie q_e$

2. $q_b \bowtie q_f$

3. $q_d \bowtie q_g$

The types in the equivalence class $[q_a]_{\bowtie}$ are unified:

$$\theta_1(q_a) \bigsqcup \theta_2(q_e)$$

$$\texttt{np} \bigsqcup \texttt{np}$$

$$= \texttt{np}$$

yielding as a result the type `np`. The types in the equivalence class $[q_b]_{\bowtie}$ are unified:

$$\theta_1(q_b) \bigsqcup \theta_2(q_f)$$

$$\texttt{noun} \bigsqcup \texttt{noun}$$

$$= \texttt{noun}$$

yielding as a result the type `noun`. The types in the equivalence class $[q_d]_{\bowtie}$ are unified:

$$\theta_1(q_d) \bigsqcup \theta_2(q_g)$$

$$\texttt{sing} \bigsqcup \texttt{plur}$$

$$= \top$$

yielding as a result the inconsistent type $\top$.

A type clash has occurred between two nodes in an equivalence class. The next step is to ascertain whether an agreement error has occurred. The path leading to the type clash, π' in the definition, is the path HEAD-DTR|NUM. There is another path, π, in F_1 which is reentrant with π', namely, the path COMP-DTR|NUM. The pair (HEAD-DTR|NUM,COMP-DTR|NUM) is an element

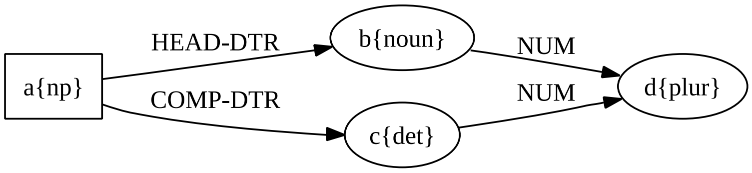


Figure 7.29: Example Input Feature Structure F_1

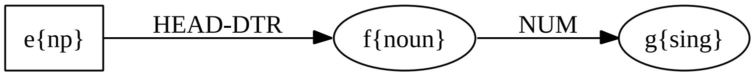


Figure 7.30: Example Input Feature Structure F_2

of *Agree*, the set of agreement path pairs in F_1, so an agreement error has occurred. In this example, the paths π'' and π''' are the same as π and π' respectively, and σ is the empty path ϵ.[10] The following steps are taken:

1. F_4 is defined. It is shown in Fig. 7.31.

2. $\delta_3(\texttt{HEAD|NUM}, [q_a]_{\bowtie}) = F_4$

3. $q_h \leftrightarrow^3_\top [q_d]_{\bowtie}$

The output feature structure F_3 is then built up as follows:

1. $Q_3 = \{[q_a]_{\bowtie}, [q_b]_{\bowtie}, [q_c]_{\bowtie}, [q_d]_{\bowtie}, q_h\}$

2. $\bar{q}_3 = [q_a]_{\bowtie}$

3. $\theta_3([q_a]_{\bowtie}) = np,\ \theta_3([q_b]_{\bowtie}) = noun,\ \theta_3([q_c]_{\bowtie}) = det, \theta_3([q_d]_{\bowtie}) = plur, \theta_3(q_h) = sing$

[10]In an example such as the one described in Section 7.4.3 where the type clash occurs at a point further down in the unificands than the agreement structure-sharing constraint, the paths π'' and π''' would be proper subpaths of the paths π and π' and σ would be a non-empty path.

Figure 7.31: The Feature Structure F_4

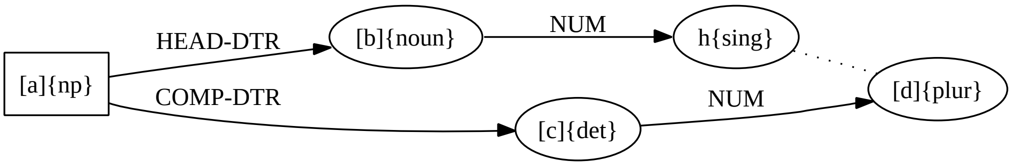


Figure 7.32: Output Feature Structure F_3: $F_1 \bigsqcup_{\leftrightarrow} F_2$ (Figs. 7.29 and 7.30)

4. $\delta_3(head, [q_a]_{\bowtie}) = [q_b]_{\bowtie}$, $\delta_3(comp, [q_a]_{\bowtie}) = [q_c]_{\bowtie}$, $\delta_3(num, [q_c]_{\bowtie}) = [q_d]_{\bowtie}$, $\delta_3(num, [q_b]_{\bowtie}) = q_h$

5. $q_h \leftrightarrow^3_\top [q_d]_{\bowtie}$

The feature structure is shown in Fig. 7.32. Inconsistent agreement identity is indicated by a dotted line, and equivalence classes of nodes indicated by square brackets.

Note again that the robust agreement unification is dependent on the existence of the set of agreement path pairs. This information needs to be read from the grammar being used, and this makes this approach a combination of a constraint relaxation approach to grammar and an error anticipation approach. In this it has more in common with the approach described in Reuer (2000) than the more abstract robustness described in Vogel and Cooper (1995) and Fouvry (2003b). The specification of agreement path pairs is also reminiscent of the feature path specifications used by Douglas (1995) which are compiled into relaxable constraints (see Section 4.3.1).

I mentioned in Section 7.3 that classical unification takes all the information from the input feature structures (unificands) and combines this information in the output feature structure. Classical feature structure unification is monotonic in the sense that the unificands are guaranteed to subsume the unification result: the output feature structure is the minimal extension or least upper bound of the input feature structures. Robust agreement unification operation does not have this property. In Fig. 7.29, for example, the feature structure stipulates that two paths should share a value. However, when the information in this feature structure is combined via robust agreement unification with the information in the feature structure in Fig. 7.30, the resulting feature structure (Fig. 7.32) contains no such stipulation. The relaxation or withdrawal of a structure-sharing constraint in the event of a type clash involves the retraction of information. The definition of subsumption between robust agreement features structures (Def. 10) reflects this: according to this definition, the feature structure in Fig. 7.29 does not subsume the feature structure in Fig. 7.32. Although this property of non-monotonicity is against the spirit of a declarative typed feature structure system, this non-monotonicity only occurs in very specific circumstances and the robust agreement feature structures can be compiled back into normal feature structures once the source of the agreement error has been established.

Robust agreement unification is non-commutative: Def. 12 makes the assumption that the first unificand contains a structure-sharing agreement constraint. The definition is particularly suited to situations like the ones described up until now where lexical feature structures are unified into phrasal feature structures. The unificands shown in Fig. 7.33 show how order-dependence ensues when both inconsistent unificands encode the structure-sharing agreement constraint. According to the definition of robust agreement unification, there are two possible results for these two unificands and

$$
\boldsymbol{np}\begin{bmatrix} \textsc{head-dtr:} & \boldsymbol{noun}\begin{bmatrix} \textsc{num:} \boxed{1}\ \textit{sing} \end{bmatrix} \\ \textsc{comp-dtr:} & \boldsymbol{det}\begin{bmatrix} \textsc{num:} \boxed{1}\ \textit{sing} \end{bmatrix} \end{bmatrix} \sqcup_{\leftrightarrow} \boldsymbol{np}\begin{bmatrix} \textsc{head-dtr:} & \boldsymbol{noun}\begin{bmatrix} \textsc{num:} \boxed{1}\ \textit{plur} \end{bmatrix} \\ \textsc{comp-dtr:} & \boldsymbol{det}\begin{bmatrix} \textsc{num:} \boxed{1}\ \textit{plur} \end{bmatrix} \end{bmatrix}
$$

$$
= \boldsymbol{np}\begin{bmatrix} \textsc{head-dtr:} & \boldsymbol{noun}\begin{bmatrix} \textsc{num:} \boxed{1}\ \textit{plur} \end{bmatrix} \\ \textsc{comp-dtr:} & \boldsymbol{det}\begin{bmatrix} \textsc{num:} \boxed{2}\ \textit{sing} \end{bmatrix} \\ \boxed{1} \leftrightarrow_{\top} \boxed{2} & \end{bmatrix}
$$

$$
\boldsymbol{np}\begin{bmatrix} \textsc{head-dtr:} & \boldsymbol{noun}\begin{bmatrix} \textsc{num:} \boxed{1}\ \textit{plur} \end{bmatrix} \\ \textsc{comp-dtr:} & \boldsymbol{det}\begin{bmatrix} \textsc{num:} \boxed{1}\ \textit{plur} \end{bmatrix} \end{bmatrix} \sqcup_{\leftrightarrow} \boldsymbol{np}\begin{bmatrix} \textsc{head-dtr:} & \boldsymbol{noun}\begin{bmatrix} \textsc{num:} \boxed{1}\ \textit{sing} \end{bmatrix} \\ \textsc{comp-dtr:} & \boldsymbol{det}\begin{bmatrix} \textsc{num:} \boxed{1}\ \textit{sing} \end{bmatrix} \end{bmatrix}
$$

$$
= \boldsymbol{np}\begin{bmatrix} \textsc{head-dtr:} & \boldsymbol{noun}\begin{bmatrix} \textsc{num:} \boxed{1}\ \textit{sing} \end{bmatrix} \\ \textsc{comp-dtr:} & \boldsymbol{det}\begin{bmatrix} \textsc{num:} \boxed{2}\ \textit{plur} \end{bmatrix} \\ \boxed{1} \leftrightarrow_{\top} \boxed{2} & \end{bmatrix}
$$

Figure 7.33: Robust Agreement Unification and Order Dependence

the one that is produced depends upon the order in which the unification is applied. While this kind of order-dependence is not a desirable property, confidence can be derived from the fact that both possible outcomes are still encoding the information that an agreement constraint has been violated and identifying the nodes at the source of the violation.

7.6 Implementing Robust Agreement Unification

The concept of robust agreement unification has been introduced and formally defined. In this section it is implemented and tested using naturalistic agreement errors.

In Section 7.6.1, I describe the process of testing the corrected *corpus2* versions of the *corpus1* agreement error sentences on the ERG. This was done to ensure that the ERG could cope with the grammatical sentences before the ungrammatical versions of these were tested using it. I also tested the ungrammatical sentences to see whether the ERG does actually reject them as expected. In Section 7.6.2, I explain how I adapted the LKB parser so that it uses robust agreement unification.[11] Finally in Section 7.6.3, I describe the results of testing this new version of the LKB on the *corpus1* agreement error sentences with the ERG.

7.6.1 Agreement Errors and the ERG: Part One

In Section 7.2.1, I remarked that there were sixty-eight sentences in *corpus1* containing an agreement error. Sixteen of these could be corrected in two ways, resulting in a set of eighty-four grammatical *corpus2* sentences. An attempt was made to parse the eighty-four grammatical *corpus2* sentences using the ERG and the LKB parser. Without changing the grammar in any way, only eight received a parse. In order to allow the full set to be parsed, appropriate lexical entries were added to the grammar, and, if necessary, the sentences themselves were minimally modified. This modification process usually took the form of paraphrasing the sentence or omitting clauses so that the sentence no longer contained constructions which were problematic for the ERG. A similar approach was adopted when sentences from the BNC were applied to the ERG (Baldwin *et al.* 2004).

The sixty-eight ungrammatical *corpus1* sentences were then parsed. Since the ERG aims to describe grammatical sentences only (see Section 2.3.1 for a discussion of this property), it is expected that none of the ungrammatical sentences should receive a parse. In fact, forty-eight of the sixty-eight ungrammatical sentences do not receive a parse as expected, whereas twenty of them do. In most cases, an agreement violation was not detected because the sentence permits another (usually implausible) reading in which the two items in the agreement violation are no longer related. Sentence (7.12) is

[11] The ERG and the LKB source code were downloaded in June 2004 from *http://lingo.stanford.edu/ftp*.

a quite reasonable example. In this sentence, the violation is between the singular verb *ends* and the plural noun *edges* but a reading is also available where the verb *ends* agrees with the singular noun *chart*.

(7.12) *It returns all* ***edges*** *of the chart that* ***ends*** *at a certain position.* (A.336)

7.6.2 Incorporating Robust Agreement Unification into the LKB

The classical unification provided in Def. 8 proceeds by stepping through the input feature structure graphs, organizing into equivalence classes the nodes that need to be merged and unifying the types of the nodes in a equivalence class. This can be implemented efficiently if the nodes in the unificands are used in the result as representatives of the equivalence classes. For example, if the feature structures in Figs. 7.34 and 7.35 are unified classically, all the nodes in the feature structure in Fig. 7.34 and one of the nodes in the feature structure in Fig. 7.35 can be used in the result. The result of the unification is depicted in Fig. 7.36. The coloured nodes represent the actual feature structure that is returned as a result. The problem with implementing feature structure unification in this way is that it is destructive: some of the structure in the unificands becomes part of the unification result. If feature structures represent partial parses in a parse chart, they will need to be available for re-use and so they cannot be modified as part of the unification process.

The obvious non-destructive graph unification algorithm is one in which a copy of the unificands is made before the unification is carried out. The problem with this is the cost, in terms of both time and space, of copying large feature structures. Tomabechi (1991) proposed a *quasi-destructive* al-

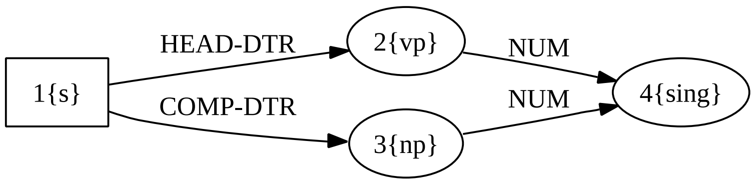


Figure 7.34: Unificand 1: Destructive Unification

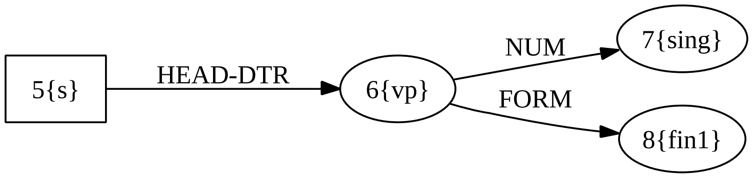


Figure 7.35: Unificand 2: Destructive Unification

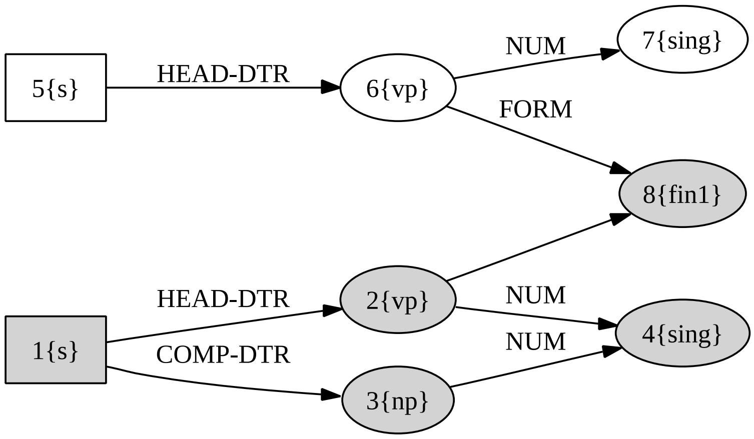


Figure 7.36: Destructive Unification Result

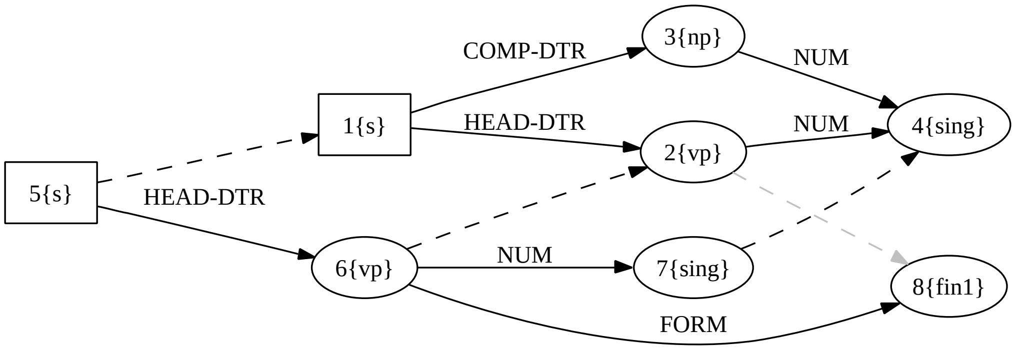


Figure 7.37: Quasi-Destructive Unification: Setting Temporary Pointers

gorithm in which copying only takes place when unification is guaranteed to succeed. This is done by setting temporary pointers between nodes in an equivalence class, and by constructing temporary new arcs. The resulting feature structure is constructed based on these pointers, only if no type clash has occurred. The pointer setting is illustrated in Fig. 7.37 for the unificands in Figs. 7.34 and 7.35, and the result of quasi-destructive unification in Fig. 7.38. In Fig. 7.37, the black dashed lines represent pointers from a node to the representative node in its equivalence class and the grey dashed line represents an arc which needs to be copied from a node to the representative node in its equivalence class. In Fig. 7.38, the node label *1c* indicates that node *1* in the first unificand has been copied.

The LKB uses a variant of Tomabechi's quasi-destructive algorithm to unify feature structures. The LKB version (see Malouf *et al.* (2000) for details) is more efficient than Tomabechi's original algorithm because it employs node-sharing between the unification result and the unificands in certain circumstances. This algorithm was modified so that it employs the robust agreement unification defined in Def. 12 during the parsing process. The

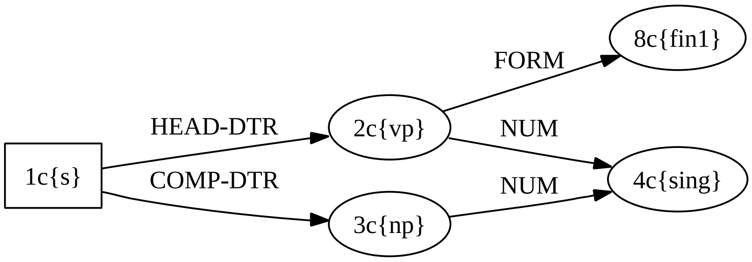


Figure 7.38: Quasi-Destructive Unification Result

modification is shown in Fig. 7.39. For ease of explanation, the diagram in Fig. 7.39 assumes Tomabechi's quasi-destructive algorithm rather than the optimized version used by the LKB but the changes I have made do not affect the differences between the two. Note that the type on node 7 is now *plur* so that I can explain what happens when an agreement error occurs.

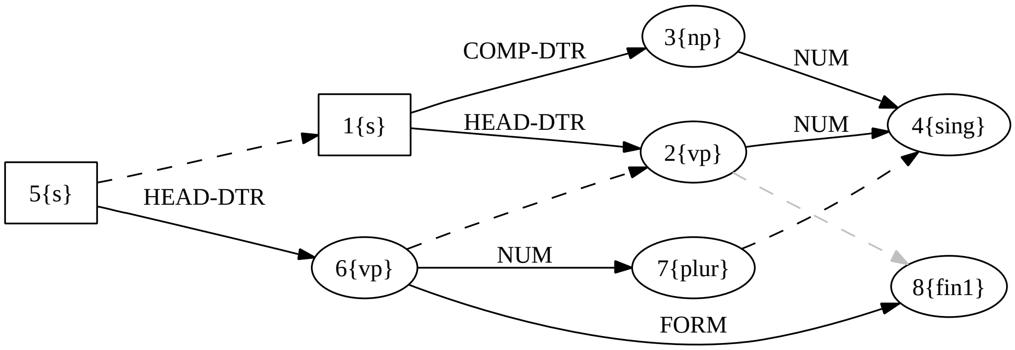
Robust agreement unification is computationally more expensive than classical unification since new structure is always created when an agreement error occurs. Performance figures are provided in the next section. The file *dag.lsp* was changed in order to implement robust agreement unification, and the main changes that were made to the code in this file are shown in Appendix E.[12]

7.6.3 Agreement Errors and the ERG: Part Two

The test data consists of the forty-eight agreement error sentences from *corpus1* which were originally rejected by the LKB, and their corrected equivalents from *corpus2*. When the test data was applied to the modified LKB

[12]The typed feature structures used in the LKB are actually embedded into larger structures known as *typed default feature structures* (Lascarides and Copestake 1999), which allow the use of defaults to be incorporated into the system. This does not interfere with the dag implementation of robust agreement unification.

1. Unification proceeds as per Fig. 7.37.

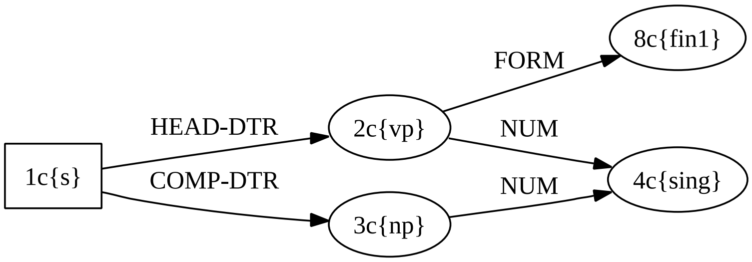


2. A type clash occurs between node 4 and 7.

3. The path leading to the clash is recorded, i.e. HEAD-DTR—NUM. A check is made that this is an agreement error path.

4. A copy of node 7 is made.

5. The resulting feature structure is constructed as per Fig. 7.37.



6. The copy of node 7 is set to be the value of the path HEAD-DTR—NUM.

7. An inconsistent agreement identity pointer is set between nodes 4c and 7c.

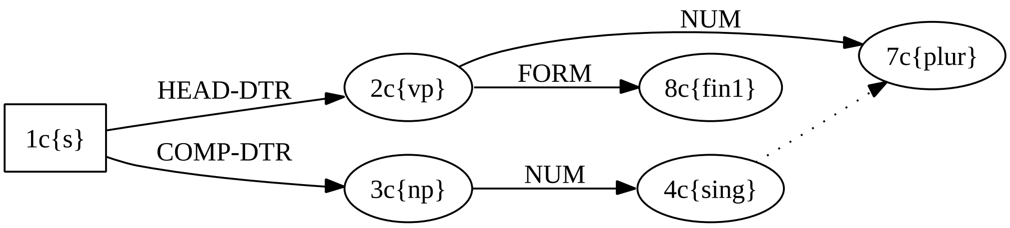


Figure 7.39: Quasi-Destructive Robust Agreement Unification

	Old Lkb		Modified Lkb	
	Grammatical	Ungrammatical	Grammatical	Ungrammatical
Total CPU Time	83558msecs	35849msecs	123264msecs	135648msecs
Mean Edges	1623.1	729.4	2085.4	2075.7
Mean Parses	16.9	0	23.6	21.9

Figure 7.40: LKB Performance Figures for Agreement Error Sentences Before and After Changes

which implements the robust agreement unification algorithm, in forty-five of the forty-eight cases, agreement constraints were relaxed. The other three cases which were still rejected were those which I considered to be cases of determiner-noun number agreement but which were not encoded in this way by the ERG. The three sentences contained the number *two* followed by a singular noun, e.g. Sentence (7.13):

(7.13) *It is important to observe how the syntactic module passes information to the semantic module through* ***two mechanism****.* (A.680)

The ERG encodes *two* as an adjective rather than a determiner, and combinations of *two* followed by a singular noun are ruled out by the non-agreement constraint which stipulutes that the noun in a determinerless noun phrase must have the property of being plural or mass. Thus, robust agreement unification will not apply to these examples and they will not receive a parse.

The table in Fig. 7.40 compares the time taken by the modified LKB to parse the forty-five *corpus1* sentences and their grammatical *corpus2* equivalents with the time taken by the original LKB to parse the same sentences. The table also shows the average number of chart edges and parses per sentence. These figures were obtained by running the original and modified LKB with the ERG grammar in tty mode on CMU Lisp (version 18e) on a Linux machine with 512 MB RAM and two 500MHz Pentium 3 processors.

Looking at the results in Fig. 7.40, it is not surprising to see that the modified LKB takes longer to parse forty-five sentences than the original LKB. The robust agreement unification algorithm is less efficient than the normal unification algorithm because it involves the creation of new structure. Also, the parsing process is slower because the relaxation of agreement constraints results in more parses.

Since robust agreement unification is available for grammatical sentences as well as for ungrammatical sentences, the modified LKB finds more parses for the grammatical sentences than the original LKB. An example is the grammatical *corpus2* sentence (7.14):

(7.14) *Two fire units, an ambulance and an emergency unit* ***were*** *sent to the scene.* (A.592)

The modified LKB finds the parses which are found by the original LKB: in all of these parses *two fire units* is analyzed as a noun phrase. It also relaxes a subject-verb number agreement constraint to find another parse where *two* is a singular noun phrase, *fire* is a plural verb which should agree with two, and *two fire units* is a sentence which is conjoined with another sentence *an ambulance and an emergency unit were sent to the scene.* The modified LKB is able to find these parses because the original LKB accepts the sentences *two fires units* as grammatical. Although it would be better if robust agreement unification was only applied to ungrammatical sentences, the parses where no constraint has been relaxed can be favoured over those achieved by relaxing an agreement constraint.

The modified LKB finds a similar number of parses for the grammatical sentence set as for the ungrammatical sentence set, and this is expected. The numbers are not exactly the same because sometimes the slight differences between the sentence pairs resulted in parses for one sentence in the pair and not for the other. Sentences (7.15) and (7.16) are an example:

(7.15) *The first of* ***these*** *visiting* ***scientist*** *will begin their contracts on the first of January 2004.* (A.859)

(7.16) *The first of* ***these*** *visiting* ***scientists*** *will begin their contracts on the first of January 2004.*

The original (and modified) LKB finds a parse for the grammatical (7.16) in which *these* and *visiting scientists* are separate noun phrases. This parse will not be found for the ungrammatical (7.15) by the original LKB because *visting scientist* can not function as a stand-alone noun phrase, and it will not be found by the modified LKB because it does not involve determiner-noun number disagreement. An example which works in the opposite direction is the sentence pair (7.17) and (7.18):

(7.17) *No real* ***improvements has*** *been done in achieving full robustness for a NLP system.* (A.200)[13]

(7.18) *No real* ***improvement has*** *been done in achieving full robustness for a NLP system.*

The ungrammatical (7.17) receives twice as many analyses from the modified LKB as the grammatical (7.18): the first set of analyses are the same as the analyses for (7.18) but the second reflect the ungrammatical sentence (7.19):

[13] This is actually an abbreviated version of the *corpus1* sentence because the full *corpus2* sentence did not receive a parse.

(7.19) *No, real **improvements has** been done in achieving full robustness for a NLP system.*

This second set of analyses will not (and should not) be found for the grammatical sentence (7.18) because *real improvement* is not a stand-alone noun phrase.

The modified LKB takes longer to parse the ungrammatical sentences than the grammatical sentences even though it produces fewer parses for the ungrammatical sentence set. This is explained by the fact that for most of the parses in the grammatical test set, no agreement constraint is relaxed and no new structure needs to be built.

Overall, the results for the modified LKB are encouraging. Although the robust agreement unification algorithm slows down the parsing process, it correctly parses the agreement error sentences in a reasonable time, without an explosion in spurious parses.

7.7 Inconsistent Agreement Identity and Non-Agreement Errors

In the final section of this chapter, I discuss why robust agreement unification is appropriate for agreement errors and not for other types of substitution errors.

Consider the sentence *We done.* Assuming a grammar which contains a constraint stipulating that the main verb in a sentence must be in finite form, when an attempt is made to classically unify the information contained in this sentence with the information contained in the grammar, a type clash will occur since the main verb in the sentence is in past-participle rather than finite form. If we related the two inconsistent nodes using *inconsistent*

$$\textbf{\textit{sen}}\begin{bmatrix} \text{COMP-DTR: } \textbf{\textit{np}} & \\ \text{HEAD-DTR: } \textbf{\textit{vp}} \begin{bmatrix} \text{VFORM: } \boxed{1}\ \textbf{\textit{pastp}} \\ \text{VFORM: } \boxed{2}\ \textbf{\textit{finite}} \end{bmatrix} \\ \boxed{1} \leftrightarrow_{\top} \boxed{2} \end{bmatrix}$$

Figure 7.41: The sentence *We done*: Invalid Robust Agreement Feature Structure

agreement identity, we would expect the resulting structure to look something like the structure illustrated in Fig. 7.41. But this structure violates the definition of a robust agreement feature structure given in Section 7.5.1 because a node (the one labelled with the type `vp`) and a feature (the feature `VFORM`) have two nodes as their value (the node labelled `pastp` and the one labelled `finite`), which can not be the case if δ is defined to be a function. Looking at this structure, there is no way of knowing at which node the error is located.

The decision to distinguish substitution errors which involve agreement and those which do not, makes sense if we take into account the differences between agreement and non-agreement errors. The clash of values inherent to an agreement error introduces an ambiguity, meaning that both values involved in the clash need to be retained until disambiguating information is available. Although the agreement error sentences in *corpus1* were corrected without ambiguity, this was because the discourse context made the meaning of the sentence clear. Such information is unlikely to be available to a parser which is encountering the sentences out of context. Take, for example, Sentence 7.20[14]:

(7.20) *Where did **this women** learn such arrogance?*

[14] Accessed via *http://sara.natcorp.ox.ac.uk/lookup.html*, March 2000.

Without any further contextual information it is impossible to tell whether this sentence means Sentence (7.21) or Sentence (7.22):

(7.21) *Where did **these women** learn such arrogance?*

(7.22) *Where did **this woman** learn such arrogance?*

The data in Fig. 7.2 suggests that Sentence 7.22 is the more likely correction, but this cannot be said with certainty. Non-agreement substitution errors do not involve the same kind of ambiguity – the "wrong" grammatical form of a word has simply been produced.

7.8 Conclusion

In this chapter, a unification-based constraint relaxation approach specifially designed to handle agreement errors has been presented. The feature structures described in Carpenter (1992) which are the linguistic objects of the HPSG linguistic theory were extended so that they have the potential to provide information about agreement errors using an *inconsistent agreement identity* relation. The resulting structures, *robust agreement feature structures*, can then be combined via *robust agreement unification*, an operation which is the same as classical unification but with the added ability to relax agreement constraints where necessary.

The relaxed unification in this chapter is more restrictive than similar approaches because it only applies to agreement errors and it does not unify inconsistent types. However it has been tested on realistic data using a wide-coverage grammar with encouraging results. The research presented in this chapter could be extended by defining a transformation procedure which transforms robust agreement feature structures back to normal feature structures during an error correction phase.

Chapter 8

Conclusion

In this final chapter of the thesis, I summarize the contributions made by the thesis to the subfield of natural language parsing which is concerned with the parsing of written ungrammatical sentences. This is done in Section 8.1 by returning to the aims of the thesis set out in Chapter 1 and discussing to what extent they have been achieved. Suggestions for further work are peppered throughout the thesis. These suggestions are gathered together and summarized in Section 8.2.

8.1 Thesis Achievements

Each thesis aim is listed, followed by a brief discussion.

1. **Demonstrate that collecting a corpus of ungrammatical written English is a useful thing to do in the context of an investigation into the problem of parsing ungrammatical written English:** The parallel corpus of ungrammatical/grammatical sentences, *corpus1* and *corpus2*, produced as part of this thesis has been used in three ways:

   (a) **To Inform an Error Grammar**: In Chapter 5, *corpus1* and *corpus2* were used to inform an error anticipation approach to the problem of parsing ungrammatical language. The differences between the two corpora were analysed and some of the results of this analysis were used to derive an error grammar from a conventional grammar. In Chapter 2, I endorsed an error grammar approach to parsing by arguing that a parser will be better able to accurately parse ungrammatical language if it has information about its structure. Some support for this claim was provided by the accuracy results for the parser described in Chapter 5: only six of the forty-four sentences in the test data which were expected to receive the correct parse did not and this was because they were not recognized as ungrammatical in the first place. The evaluation results presented in Chapter 6 for the Penn-Treebank-trained parsers (Collins 1999, Charniak 2000), although good, show that an explicit model of ungrammaticality could be of use to such parsers.

(b) **As Test Data**: In Chapter 7, the robust agreement unification operator was tested on sentences from *corpus1* which contain an agreement error. In Chapter 5, naturalistic ungrammatical sentences which were collected in the same way as the *corpus1* sentences were used to test the parser. Naturalistic sentences always represent a more realistic test of a parser than invented ones.

(c) **As Part of an Evaluation Procedure**: Chapter 6 described an evaluation procedure for testing how well the parse of an ungrammatical sentence produced by some parser reflects that sentence's meaning. This procedure has as its basis the parallel *corpus1* and *corpus2*.

These three uses demonstrate that a corpus of ungrammatical written sentences can be fruitfully used in research into the parsing of such sentences. Furthermore, *corpus1* and *corpus2* represent a valuable linguistic resource which can be used in other areas besides parsing.

2. **Provide a methodology which can be used to compile a corpus of ungrammatical English sentences:** In Chapter 3, I described the methodology I used to compile *corpus1* and *corpus2*. This methodology, which, to my knowledge, has not been used previously to

compile a corpus of ungrammatical language, involves the application of one person's grammaticality judgement on sentences encountered while reading. The application of the grammaticality judgements are anchored in a practical definition of the term *ungrammatical*. Unclear cases are omitted so that *corpus1* only includes sentences upon which my linguistic intuition is clear. A questionnaire study, also described in Chapter 3, suggests that the data in *corpus1* and *corpus2* can be relied upon and provides support for the argument that the grammaticality of sentences should be judged in context rather than in isolation.

3. **Investigate the kind of ungrammatical sentences that tend to be produced in the sentences of written English:** In Chapter 5, I presented the results of an analysis of the differences between the sentences in *corpus1* and those in *corpus2*. This analysis found the following frequency ordering among the correction operators which were applied to *corpus1* sentences to arrive at the *corpus2* sentences: composite < delete < insert < substitute. Further classifications were then carried out on each of the four correction types. Some of the error analysis results were employed in Chapter 5, in designing the error grammar for a two-stage parsing algorithm. These error analysis results are also available for use by other researchers.

4. **Provide a procedure for evaluating a parser's ability to handle ungrammatical sentences:** In Chapter 6, a novel and flexible method for evaluating a parser's ability to parse ungrammatical sentence was presented. The test data for this method is a set of ungrammatical sentences (*corpus1* is an example) and the gold standard is the set of parses produced by the parser under evaluation for

the corrected versions of the ungrammatical sentences (e.g. *corpus2*). This method was applied to two broad-coverage probabilistic parsers (Collins 1999, Charniak 2000).

5. **Investigate parsing approaches which are capable of finding analyses for the kind of ungrammatical sentences found in written English:** Two contributions of this thesis are in the actual parsing of ungrammatical written Engish. The first contribution is the two-stage bottom-up error-grammar-based parser described in Chapter 5: this parser embodies various aspects of well-known robust techniques discussed in Chapter 4, but, to my knowledge, it represents an original combination of these elements. Relative to its grammar of grammatical language, this parser will parse ungrammatical sentences which can be corrected by adding or deleting one word or phrase, and the majority (according to the error analysis carried out in the same chapter) of ungrammatical sentences which can be corrected by substituting one word for another. Chapter 7 described an extension to the typed feature structure system underlying the Linguistic Knowledge Base (LKB) (Copestake 2002). The concepts of *inconsistent agreement identity*, *robust agreement feature structures* and *robust agreement unification* which can be used to parse sentences containing agreement errors are original. Initial testing of these concepts, carried out using the LKB parser and the English Resource Grammar (ERG) (Copestake and Flickinger 2000) are encouraging.

8.2 Future Work

There are many ways in which the work presented in the preceding chapters can be extended. The following list is a summary of extensions which I think would be useful:

1. **A Larger Corpus**: At present, *corpus1* contains approximately 20,000 words. Although, the gathering of errors was suspended in order to provide a stable unit to analyse for this thesis research, the development of the corpus is ongoing work. It would be useful if a probabilistic error grammar could be learnt automatically from *corpus1*, or if error correction techniques could be automatically devised based on the differences between *corpus1* and *corpus2*. While machine-learning techniques cannot be applied fruitfully to a small corpus, they could prove more useful if applied to a larger corpus.[1] It would be interesting to investigate to what extent errors could be artificially introduced into a large corpus of grammatical written English based on the differences between the sentences in *corpus1* and *corpus2* (see, for example, Bigert (2004)).

2. **Error Grammar Parsing**: The error grammar approach to parsing described in Chapter 5 could be improved so that fewer error rules are explored but without a significant decrease in accuracy. A discussion of how this could be achieved is contained in Section 5.3.4. It could also be extended so that sentences containing composite errors and more than one error can be parsed. Whether or not the approach will scale-

[1] An attempt was made to use a decision tree to learn how to classify errors based on the differences between *corpus1* and *corpus2* but the results were not encouraging enough to be included in this thesis and I took the decision to suspend this line of investigation until the corpus was extended.

up when used with a very large set of grammar rules and lexical entries also needs to be investigated.

3. **Robust Agreement Unification**: The parsing work presented in Chapter 7 can be extended by defining and implementing an operation which compiles robust agreement feature structures back into normal feature structures if the source of an agreement error is detected. There is potential to improve upon the figures presented in Fig. 7.40 in Section 7.6.3 by implementing robust agreement unification in a more efficient way.

4. **Parser Evaluation**: The parser evaluation method presented in Chapter 6 has been applied to two parsers which produce Penn-Treebank style phrase structure trees (Collins 1999, Charniak 2000). However the method itself has been defined in a flexible way and it would be interesting to apply this method to parsers which return analyses in a different form, e.g the dependency graphs produced by the Minipar parser (Lin 1998). The evaluation results for the probabilistic parsers show that, according to this measure, they are quite successful at parsing ungrammatical language. However, a further line of research for those working within probabilistic parsing would to be adapt these parsers so that the figures presented in Fig. 6.24 of Section 6.4.2 are improved upon.

8.3 Concluding Remarks

After carrying out this thesis work, I am firmly convinced that the only way to advance the field of natural language engineering is to make optimum use of "natural" language. The use of natural language as opposed to invented examples does not preclude the existence of a grammatical/ungrammatical distinction. Even though a clean distinction between the two concepts is not possible, this does not mean that a distinction does not exist. We can make optimum use of natural language by analysing it in a natural way: it is better, for example, to obtain grammaticality judgements on sentence presented in a natural context than on sentences presented in isolation because in their everyday use of language, people simply do not deal with isolated sentences. While the programs developed to process language cannot be expected to possess the same contextual knowledge that people do, there is no reason why this knowledge cannot be put to good use in the development of linguistic resources to be used by computational linguistic applications.

To conclude, this thesis attacks the problem of parsing ungrammatical written English from a variety of empirically motivated perspectives, with the corpus of ungrammatical/grammatical English as a springboard. The corpus, the way the corpus has been developed and the ways that this corpus has been put to use offer a valuable contribution to the field of parsing.

Bibliography

Jan Aarts, 1991. Intuition-based and Observation-based Grammars. In *English Corpus Linguistics*, Karin Aijmer and Bengt Altenberg, (ed), pp. 44–63. Longman Group UK Limited.

Steven Abney, 1990. Rapid Incremental Parsing with Repair. In *Proceedings of the sixth New OED conference: electronic text research*, pp. 1–9, Waterloo, Ontario.

Steven Abney, 1996. Partial Parsing via Finite-State Cascades. In Carroll (1996), pp. 8–15.

Itziar Aduriz, Izaskun Aldezabal, Maxux Aranzabe, Bertol Arrieta, Jose Mari Arriola, Aitziber Atutxa, Arantza Díaz de Ilarraza, Koldo Gojenola, Maite Oronoz, Kepa Sarasola, and Ruben Urizar, 2002. The Design of a Digital Resource to store the Knowledge of Linguistic Errors. In *Proceedings of the 2002 Digital Resources for the Humanities Conference*, Edinburgh.

S. Aït-Mokhtar, J.P. Chanod, and C.Roux, 2002. Robustness Beyond Shallowness: Incremental Deep Parsing. *Natural Language Engineering*, 8-2 (2002), 121–144.

Eric Atwell and Stephen Elliott, 1987. Dealing with Ill-formed English Text. In Garside *et al.* (1987), pp. 120–138.

Timothy Baldwin, John Beavers, Emily M. Bender, Dan Flickinger, Ara Kim, and Stephan Oepen, 2004. Beauty and the Beast: What running a broad-coverage precision grammar over the BNC taught us about the grammar - and the corpus. In Kepser and Reis (2004), pp. 21–26.

Afzal Ballim, Vincenzo Pallotta, and Christian Lieske, 1999. Robust Text Analysis: an Overview. Downloadable from lith-www.epfl.ch/p̃allotta/rota.ps.gz, September 2004.

Afzal Ballim and Graham Russell, 1994. LHIP: Extended DCGs for Configurable Robust Parsing. In *Proceedings of the 15th International Conference on Computational Linguistics*, pp. 501–507, Kyoto, Japan.

E.G. Bard, D. Robertson, and A. Sorace, 1996. Magnitude Estimation of Linguistic Ability. *Language*, 72-1 (1996), 32–68.

Michael Barlow and Suzanne Kemmer, (ed), 2000. *Usage-Based Models of Language*. CSLI Publications, Stanford, California.

Roberto Basili and Fabio Massimo Zanzotto, 2002. Parsing Engineering and Empirical Robustness. *Natural Language Engineering*, 8-2/3 (2002), 97–120.

Markus Becker, Andrew Bredenkamp, Berthold Crysmann, and Judith Klein, 1999. Annotation of Error Types for German News Corpus. In *Proceedings of the ATALA workshop on Treebanks*, Paris.

Emily M. Bender, Dan Flickinger, Stephan Oepen, and Timothy Baldwin, 2004. Arboretum: Using a Precision Grammar for Grammar Checking in CALL. In *Proceedings of the InSIL/ICALL Symposium: NLP and Speech Technologies in Advanced Language Learning Systems*, Venice, Italy.

Thomas Bever, Jerrold Katz, and Terence D. Langendoen, (ed), 1977. *An Integrated Theory of Linguistic Ability*. The Harvester Press Limited: The Language and Thought Series.

Johnny Bigert, 2004. Probabilistic Detection of Context-Sensitive Spelling Errors. In *Proceedings of the 4th International Conference on Language Resources and Evaluation*, Vol. Five, pp. 1633–1636, Lisbon.

Johnny Bigert, Jonas Sjöbergh, Ola Knutsson, and Magnus Sahlgren, 2004. Automatic Evaluation of Parser Robustness: Eliminating Manual Labor and Annotated Resources. Unpublished manuscript.

E. Black, Steve Abney, Dan Flickinger, C. Gdaniec, R. Grishman, P. Harrison, D. Hindle, R. Ingria, Fred Jelinek, J. Klavans, M. Liberman, M. Marcus, S. Roukos, B. Santorini, and T. Strzalkowski, 1991. A Procedure for Quantitatively Comparing the Syntactic Coverage of English Grammars. In *Proceedings of the 1991 DARPA Speech and Natural Language Workshop*, pp. 306–311.

Leonard Bloomfield, 1935. *Language*. Compton Printing Ltd. London and Aylesbury.

Rens Bod, 1998. *Beyond Grammar: An Experience-Based Theory of Language*. CSLI Publications, Stanford.

Paul Boersma, Joost Dekkers, and Jeroen van de Weijer, 2000. Introduction. In *Optimality Theory: Phonology, Syntax and Acquisition*, Joost Dekkers, Frank van der Leeuw, and van de Weijer, (ed). Oxford University Press.

Sabine Brants, Stefanie Dipper, Silvia Hansen, Wolfgang Lezius, and George Smith, 2002. The TIGER Treebank. In *Proceedings of the Workshop on Treebanks and Linguistic Theories*, Sozopol.

Andrew Bredenkamp, Berthold Crysmann, and Mirela Petrea, 2000. Looking for Errors: A Declarative Formalism for Resource-Adaptive Language Checking. In *Proceedings of the 2nd International Conference on Language Resources and Evaluation*, Athens, Greece.

Joan Bresnan and Ronald M. Kaplan, 1982. Lexical-Functional Grammar: a formal system for grammatical representation. In *The Mental Representation of Grammatical Relations*, Joan Bresnan, (ed). MIT Press, Cambridge, Massachusetts.

Ted Briscoe, John Carroll, Jonathon Graham, and Ann Copestake, 2002. Relational Evaluation Schemes. In Carroll *et al.* (2002), pp. 4–8.

Alan Bryman and Duncan Cramer, 2001. *Quantitative Data Analysis with SPSS Release 10 for Windows.* Routledge.

L. Burnard, 2000. User Reference Guide for the British National Corpus. Technical report, Oxford University Computing Services.

James G. Carbonell and Philip J. Hayes, 1983. Recovery Strategies for Parsing Extragrammatical Language. *American Journal of Computational Linguistics*, 9-3–4 (1983), 123—146.

Bob Carpenter, 1992. *The Logic of Typed Feature Structures.* Cambridge University Press.

Bob Carpenter and Gerald Penn, 2001. *The Attribute Logic Engine: User's Guide, Version 3.2.1.* Distributed as part of ALE. http://www.cs.toronto.edu/ gpenn/ale.html. Last accessed, October 2004.

John Carroll, (ed), 1996. *Proceedings of the Workshop on Robust Parsing at the Eighth Summer School in Logic, Language and Information*, 1996.

John Carroll, (ed), 1998a. *Proceedings of Workshop on The Evaluation of Parsing Systems at the First International Conference on Language Resources and Evaluation*, Cognitive Science Research Papers 489. University of Sussex, 1998a.

John Carroll, 1998b. Evaluation: Parsing. *Elsnews*, 7-3 (1998).

John Carroll and Ted Briscoe, 1996. Robust Parsing – A Brief Overview. In Carroll (1996), pp. 1–7.

John Carroll, Anette Frank, Dekang Lin, Detlef Prescher, and Hans Uszkoreit, (ed), 2002. *Proceedings of the "Beyond Parseval - Towards Improved Evaluation Measures for Parsing Systems" Workshop at the 3rd International Conference on Linguistic Resources and Evaluation (LREC)*, Las Palmas, Gran Canaria, 2002.

Jean-Pierre Chanod and Pasi Tapanainen, 1996. A Robust Finite-State Parser for French. In Carroll (1996), pp. 16–25.

Eugene Charniak, 1997. Statistical Techniques for Natural Language Parsing. *AI Magazine*, 18-4 (1997).

Eugene Charniak, 2000. A Maximum-Entropy-Inspired Parser. In *Proceedings of the Annual Meeting of the North American Association for Computational Linguistics*.

Noam Chomsky, 1957. *Syntactic Structures*. Mouton, The Hague.

Noam Chomsky, 1961. Formal discussion: the development of grammar in child language. Reprinted in J.P.B. Allen and P. van Buren, Chomsky: Selected Readings, 1971, Oxford University Press, 129–34.

Noam Chomsky, 1964. Degrees of Grammaticalness. In Fodor and Katz (1964), pp. 384–389. Also appeared as "Some Methodological Remarks on Generative Grammar", Word(17), 1961, 219–239.

Noam Chomsky, 1965. *Aspects of the Theory of Syntax.* The MIT Press.

Stephen Clark and Julia Hockenmaier, 2002. Evaluating a Wide-Coverage CCG Parser. In Carroll *et al.* (2002), pp. 60–66.

Michael Collins, 1997. Three Generative Lexicalized Models for Statistical Parsing. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics*, pp. 16–23, Madrid.

Michael Collins, 1999. *Head-Driven Statistical Models for Natural Language Parsing.* Phd. Thesis, University of Pennsylvania.

Ann Copestake, 2002. *Implementing Typed Feature Structure Grammars.* CSLI Lecture Notes. Cambridge University Press.

Ann Copestake and Dan Flickinger, 2000. An Open-source Grammar Development Environment and Broad-coverage English Grammar using HPSG. In *Proceedings of the 2nd International Conference on Language Resources and Evaluation*, Athens, Greece.

Stephen Crain and Mark Steedman, 1985. On Not Being Led Up The Garden Path: The Use of Context by the Psychological Syntax Processor. In *Natural Language Parsing: Psychological, Computational and Theoretical Perspectives* (Dowty *et al.* 1985), pp. 320–58.

F.J. Damerau, 1964. A Technique for Computer Detection and Correction of Spelling Errors. *Communications of the ACM*, 7-3 (1964), 171–176.

Bruce L. Derwing, 1979. Against Autonomous Linguistics. In Perry (1979a), pp. 163–189.

Teun A. van Dijk, 1976. Acceptability in Context. In Greenbaum (1976a), pp. 39–62.

Shona Douglas, 1995. Robust PATR for Error Detection and Correction. In *Nonclassical Feature Systems*, Andreas Schöter and Carl Vogel, (ed). Centre for Cognitive Science, University of Edinburgh, Working Papers in Cognitive Science, Volume 10.

Shona Douglas and Robert Dale, 1992. Towards Robust PATR. In *Proceedings of the 15th International Conference on Computational Linguistics*, pp. 468–474.

David Dowty, 2000. The Dual Analysis of Adjuncts/Complements in Categorial Grammar. *ZAS Papers in Linguistics*, 17 (2000).

David D. Dowty, Lauri Karttunen, and Arnold M. Zwicky, 1985. *Natural Language Parsing: Psychological, Computational and Theoretical Perspectives*. Cambridge University Press.

J. Earley, 1970. An Efficient Context-Free Parsing Algorithm. *Communications of the ACM*, 6-8 (1970), 451–455.

Fred R. Eckman, (ed), 1977. *Current themes in linguistics: Bilingualism, Experimental Linguistics, and Language Typologies*. Washington: Hemisphere.

Ute Ehrlich and Gerhard Hanrieder, 1996. Robust Speech Parsing. In Carroll (1996), pp. 26–34.

Izumi Emi, Kiyotaka Uchimoto, and Hitoshi Isahara, 2004. The Overview of the SST speech corpus of Japanese Learner English and Evaluation through the Experiment on Automatic Detection of Learners' Errors. In *Proceedings of the 4th International Conference on Language Resources and Evaluation*, Vol. Four, pp. 1435–1439, Lisbon.

Gregor Erbach, 1997. *Bottom-Up Earley Deduction for Preference-Driven Natural Language Processing.* Phd. Thesis, Universität des Saarlandes.

Charles A. Ferguson and Michael Barlow, 1988. Introduction. In *Agreement in Natural Language: Approaches, Theories and Descriptions*, Michael Barlow and Charles A. Ferguson, (ed), pp. 1–22. CSLI Publications, Stanford, California.

Jerry A. Fodor and Jerrold J. Katz, (ed), 1964. *The Structure of Language: Readings in the Philosophy of Language.* Prentice-Hall.

Jennifer Foster, 2000. A Unification Strategy for Parsing Agreement Errors. In *Proceedings of the Student Session of the European Summer School for Logic, Language and Information*, Catherine Piliére, (ed), pp. 77–87.

Jennifer Foster, 2003. Parsing Ill-formed Text using an Error Grammar. In *Proceedings of the Artificial Intelligence/Cognitive Science Conference 2003*, pp. 55–60, Dublin.

Jennifer Foster, 2004. Parsing Ungrammatical Input: An Evaluation Procedure. In *Proceedings of the 4th International Conference on Language Resources and Evaluation*, Vol. Six, pp. 2039–2042, Lisbon.

Jennifer Foster and Carl Vogel, 2004a. Good Reasons for Noting Bad Grammar: Constructing a Corpus of Ungrammatical Language. In Kepser and Reis (2004), pp. 151–152.

Jennifer Foster and Carl Vogel, 2004b. Parsing Ill-formed Text using an Error Grammar. *Artificial Intelligence Review: Special AICS2003 Issue*, 21 (2004), 269–291.

Frederik Fouvry, 2000. Robust Unification for Linguistics. In *Proceedings of the 1st workshop on Robust Methods in Analysis of Natural Language Data*, Lausanne.

Frederik Fouvry, 2003a. Constraint Relaxation with Weighted Feature Structures. In *Proceedings of the 8th International Workshop on Parsing Technologies*, Nancy, France.

Frederik Fouvry, 2003b. *Robust Processing for Constraint-based Grammar Formalisms*. Phd. Thesis, Department of Language and Linguistics, University of Essex.

Nelson W. Francis and Henry Kučera, 1982. *Frequency analysis of English usage. Lexicon and Grammar*. Houghton Mifflin, Boston.

Anette Frank, Louisa Sadler, Josef van Genabith, and Andy Way, 2003. From Treebank Resources to LFG F-Structures: Automatic F-Structure Annotation of Treebank Trees and CFGs extracted from Treebanks. In *Treebanks: Building and Using Syntactically Annotated Corpora*, Anne Abeille, (ed). Kluwer Academic Publishers, Dordrecht/Boston/London, The Netherlands.

Victoria A. Fromkin, (ed), 1980. *Errors in Linguistic Performance: Slips of the Tongue, Ear, Pen and Hand*. Academic Press, New York.

R. Gaizauskas, M. Hepple, and C. Huyck, 1998. Modifying Existing Annotated Corpora for General Comparative Evaluation of Parsing. In Carroll (1998a).

Roger Garside, Geoffrey Leech, and Geoffrey Sampson, (ed), 1987. *The Computational Analysis of English: a Corpus-Based Approach*. Longman, London.

Gerald Gazdar, 2003. Paradigm Merger in Natural Language Processing. ftp://ftp.cogs.susx.ac.uk/pub/nlp/g.gazdar/papers/paradigm.ps , Downloaded 7th May 2003.

Manfred Gehrke, 1983. A Multilevel Approach to Handle Non-Standard Input. In *Proceedings of the 1st Meeting of the European Association of Computational Linguistics*, pp. 183–187, Pisa, Italy.

Damien Genthial, Jacques Courtin, and Jacques Ménézo, 1994. Towards a More User-Friendly Correction. In *Proceedings of the 15th International Conference on Computational Linguistics*, pp. 1083–1088.

Amorey Gethin, 1990. *Antilinguistics: A critical assessment of modern linguistic theory and practice*. Intellect, Oxford.

Sebastian Goeser, 1990. A Linguistic Theory of Robustness. In *Proceedings of the 13th International Conference on Computational Linguistics*, pp. 156–161.

Koldo Gojenola and Maite Oronoz, 2000. Corpus-Based Syntactic Error Detection Using Syntactic Patterns. In *Proceedings of the Student Research Workshop at the Annual Meeting of the North American Association for Computational Linguistics*, Seattle.

Sylviane Granger, 1993. International Corpus of Learner English. In *English Language Corpora: Design, Analysis and Exploitation*, J. Aarts, P. de Haan, and N.Oostdijk, (ed). Rodopi, Amsterdam.

Sidney Greenbaum, (ed), 1976a. *Acceptability in Language.* Mouton Publishers, The Hague.

Sidney Greenbaum, 1976b. Introduction. In *Acceptability in Language* (Greenbaum 1976a), pp. 1–12.

Sidney Greenbaum, 1977. The Linguist as Experimenter. In Eckman (1977), pp. 124–144.

Dennis Grinberg, John Lafferty, and Daniel Sleator, 1995. A Robust Parsing Algorithm for Link Grammars. In *Proceedings of the Fourth International Workshop on Parsing Technologies*, Prague.

Donald Hindle, 1994. A Parser for Text Corpora. In *Computational Approaches to the Lexicon*, B.T.S. Atkins and A. Zampolli, (ed), pp. 103–153. Oxford University Press.

Donald Hindle and Mats Rooth, 1991. Structural Ambiguity and Lexical Relations. In *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*, pp. 229–236.

Jerry R. Hobbs, Douglas Appelt, John Bear, David Israel, Megumi Kameyama, Mark Stickel, and Mabry Tyson, 1997. FASTUS: A Cascaded Finite-state Transducer for Extracting Information from Natural Language Text. In Roche and Schabes (1997), pp. 383–406.

John E. Hopcroft, Rajeev Motwani, and Jeffrey D. Ullman, 2001. *Introduction to Automata Theory, Languages and Computation.* Addison Wesley, 2nd edition.

Peter Ingels, 1996. *A Robust Text Processing Technique Applied to Lexical Error Recovery.* Phd. Thesis, Linkoping University, Sweden.

Carl James, 1998. *Errors in Language Learning and Use: Exploring Error Analysis*. Addison Wesley Longman.

K. Jensen, G.E. Heidorn, L.A. Miller, and Y. Ravin, 1983. Parse Fitting and Prose Fixing: Getting a Hold on Ill-formedness. *American Journal of Computational Linguistics*, 9-3–4 (1983), 147—160.

S. Johansson, G. N. Leech, and H. Goodluck, 1978. *Manual of Information to accompany the Lancaster-Oslo/Bergen Corpus of British English, for use with digital computers*. Department of English, University of Oslo.

Mark Johnson, 1988. *Attribute-Value Logic and the Theory of Grammar*. Center for the Study of Language and Information, Stanford.

Aravind K. Joshi and B. Srinivas, 1994. Disambiguation of Super Parts of Speech (or Supertags): Almost Parsing. In *Proceedings of the 15th International Conference on Computational Linguistics*, pp. 154–160, Kyoto, Japan.

Daniel Jurafsky and James H. Martin, 2000. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Prentice-Hall.

Fred Karlsson, 1990. Constraint Grammar as a Framework for Parsing Running Text. In *Proceedings of the 13th International Conference on Computational Linguistics*, Vol. 3, pp. 168–173.

Lauri Karttunen and Arnold M. Zwicky, 1985. Introduction. In *Natural Language Parsing: Psychological, Computational and Theoretical Perspectives* (Dowty *et al.* 1985), pp. 1–25.

Tsuneaki Kato, 1994. Yet Another Chart-based Technique for Parsing Ill-formed Input. In *Proceedings of the Fourth Conference on Applied Natural Language Processing*, pp. 107–112.

Frank Keller and Ash Asudeh, 2000. Constraints on Linguistic Coreference: An Experimental Investigation of Exempt Anaphors. Manuscript, University of Edinburgh and Stanford University.

Suzanne Kemmer and Michael Barlow, 2000. Introduction: A Usage-Based Conception of Language. In Barlow and Kemmer (2000), pp. vii–xxviii.

Stephan Kepser and Marga Reis, (ed), 2004. *Pre-Proceedings of the International Conference on Linguistic Evidence: Empirical, Theoretical and Computational Perspectives*, Tübingen, 2004. Organized by the Sonderforschungsbereich 441 "Linguistic Data Structures", University of Tübingen, Germany.

Albert Kim, 1994. Graded Unification: A Framework for Interactive Processing. In *Proceedings of the Student Session of the 30th Annual Meeting of the Association for Computational Linguistics.*

Sandra Kübler and Heike Telljohann, 2002. Towards a Dependency-Oriented Evaluation for Partial Parsing. In Carroll *et al.* (2002), pp. 9–16.

Stan C. Kwasny and Norman K. Sondheimer, 1981. Relaxation Techniques for Parsing Ill-Formed Input. In *American Journal of Computational Linguistics*, Vol. 7, pp. 99–108.

William Labov, 1972. *Sociolinguistic Patterns.* University of Pennsylvania Press.

George Lakoff, 1971. Presupposition and Relative Well-formedness. In *Semantics: An interdisciplinary reader in philosophy, linguistics and psychology*, Danny D. Steinberg and Leon A. Jakobovits, (ed), pp. 329–340. Cambridge University Press.

George Lakoff, 1973. Fuzzy Grammar and the Performance/Competence Terminology Game. In *Papers from the ninth regional meeting of the Chicago Linguistic Society*, pp. 52–59.

Robin Lakoff, 1976. You say what you are: Acceptability and Gender-related Language. In Greenbaum (1976a), pp. 73–86.

Sidney Lamb, 2000. Bidirectional Processing in Language and Related Cognitive Systems. In Barlow and Kemmer (2000), pp. 87–120.

Terence D. Langendoen and Thomas G. Bever, 1977. Can a not unhappy person be called a not sad one? In Bever *et al.* (1977), pp. 239–260.

Terence D. Langendoen, Nancy Kalish-Landon, and John Dore, 1977. Dative Questions: A Study in the Relation of Acceptability to Grammaticality of an English Sentence Type. In Bever *et al.* (1977), pp. 195–223.

Alex Lascarides and Ann Copestake, 1999. Default Representation in Constraint-Based Frameworks. *Computational Linguistics*, 25-1 (1999), 55–106.

Alon Lavie and Masaru Tomita, 1993. GLR* - an Efficient Noise-skipping Parsing Algorithm for Context-free Grammars. In *Proceedings of the Third International Workshop on Parsing Technologies*, Tilburg (the Netherlands), Durbuy (Belgium).

Kong Joo Lee, Cheol Jung Kweon, Jungyun Seo, and Gil Chang Kim, 1995. A Robust Parser Based on Syntactic Information. In *Proceedings of the Seventh Conference of the European Association for Computational Linguistics.*

Geoffrey Leech, 1987. General Introduction. In Garside *et al.* (1987), Chapter One, pp. 1–12.

W.J.M. Levelt, J.A.W.M. Van Gent, A.F.J. Haans, and A.J.A. Meijers, 1976. Grammaticality, Paraphrase and Imagery. In Greenbaum (1976a), pp. 87–101.

V.I. Levenshtein, 1966. Binary Codes Capable of Correcting Deletions, Insertions and Reversals. *Cybernetics and Control Theory*, 10-8 (1966), 707–710.

Dekang Lin, 1998. Dependency-based Evaluation of MINIPAR. In Carroll (1998a).

Gordon Lyon, 1974. Syntax-Directed Least-Errors Analysis for Context-Free Languages. *Communications of the ACM*, 17-1 (1974), 3–14.

David M. Magerman and Carl Weir, 1992. Efficiency, Robustness and Accuracy in Picky Chart Parsing. In *Proceedings of the 30th Annual Meeting of the Association for Computational Linguistics.*

Robert Malouf, John Carroll, and Ann Copestake, 2000. Efficient Feature Structure Operations without Compilation. *Journal of Natural Language Engineering (Special Issue on Efficient Processing with HPSG: methods, systems, evaluation)*, 6-1 (2000).

Mitchell Marcus, Grace Kim, Mary Ann Marcinkiewicz, Robert MacIntyre, Ann Bies, Mark Ferguson, Karen Katz, and Britta Schasberger, 1994. The

Penn Treebank: Annotating Predicate Argument Structure. In *Proceedings of the 1994 ARPA Speech and Natural Language Workshop.*

Mitchell P. Marcus, Donald Hindle, and Margaret M. Fleck, 1983. D-Theory: Talking about Talking about Trees. In *Proceedings of 21st Annual Meeting of the Association for Computational Linguistics*, pp. 129–36. ACL.

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz, 1993. Building a Large Annotated Corpus of English: the Penn Treebank. *Computational Linguistics*, 19-2 (1993), 313–330.

Peter H. Matthews, 1993. *Grammatical Theory in the United States from Bloomfield to Chomsky.* Cambridge University Press.

Kathleen F. McCoy and Lisa N. Masterman, 1997. A Tutor for Teaching English as a Second Language for Deaf Users of American Sigh Language. In *Proceedings of Natural Language Processing for Communication Aids, a Workshop at the 35th Annual Meeting of the Association for Computational Linguistics.*

Chris S. Mellish, 1989. Some Chart-Based Techniques for Parsing Ill-formed Input. In *Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics*, pp. 102–109.

Andreas Mengel and Wolfgang Lezius, 2000. An XML-based Representation Format for Syntactically Annotated Corpora. In *Proceedings of the Second International Conference on Language Resources and Evaluation*, pp. 121–126.

Wolfgang Menzel, 1995. Robust Processing of Natural Language. In *Proceedings of the 19th Annual German Conference on Artificial Intelligence*, Lecture Notes in Computer Science 981, pp. 19–34, Bielefeld.

Wolfgang Menzel, 1996. Constraint Satisfaction for Robust Parsing of Spoken Language. In *Proceedings of the Workshop on Non-Standard Constraint Processing at the 1996 European Conference on Artificial Intelligence*, Budapest.

Lance A. Miller, 1982. Natural Language Texts are not Necessarily Grammatical and Unambiguous or even Complete. In *Proceedings of the 20th Annual Meeting of the Association for Computational Linguistics*, pp. 167–168.

B.A. Mohan, 1976. Acceptability Testing and Fuzzy Grammar. In Greenbaum (1976a), pp. 133–148.

Robert Moore and John Dowding, 1991. Efficient Bottom-Up Parsing. In *Proceedings of the February 1991 DARPA Speech and Natural Language Workshop*, Asilomar, California.

Gabriele Musillo and Khalil Sima'an, 2002. Towards Comparing Parsers from Different Linguistic Frameworks: An Information Theoretic Approach. In Carroll *et al.* (2002), pp. 44–50.

Mark-Jan Nederhof, Gosse Bouma, Rob Koeling, and Gertjan van Noord, 1997. Grammatical Analysis in the OVIS spoken-dialogue system. In *Proceedings of the Workshop on Spoken Dialogue Systems at the 35th Annual Meeting of the Association for Computational Linguistics*, Madrid.

Frederick J. Newmeyer, 1983. *Grammatical Theory, its Limits and its Possibilities.* The University of Chicago Press.

Gertjan van Noord, 1996. Robust parsing with the head-corner parsing. In Carroll (1996), pp. 83–92.

Stephan Oepen, Ezra Callahan, Dan Flickinger, Christopher D. Manning, and Kristina Toutanova, 2002. LinGO Redwoods: A Rich and Dynamic Treebank for HPSG. In Carroll *et al.* (2002), pp. 17–22.

Miles Osbourne and Derek Bridge, 1994. Learning Unification-based Grammars using the Spoken English Corpus. In *Second IEE International Colloquium on Grammatical Inference*, pp. 260–270, Alicante.

Carolyn Penstein Rosé and Alon Lavie, 1997. An Efficient Distribution of Labor in a Two Stage Robust Interpretation Process. In *Proceedings of 1997 Conference on Empirical Methods in Natural Language Processing*.

Fernando C.N. Pereira and Stuart M. Shieber, 1987. *Prolog and Natural-Language Analysis*. CSLI Lecture Notes: Number 10. CSLI Publications, Stanford, California.

Thomas A. Perry, (ed), 1979a. *Evidence and Argumentation in Linguistics*. de Gruyter, Berlin.

Thomas A. Perry, 1979b. Introduction. In *Evidence and Argumentation in Linguistics* (Perry 1979a), pp. 1–7.

Carl Pollard and Ivan A. Sag, 1994. *Head-Driven Phrase Structure Grammar*. University of Chicago Press and CSLI Publications.

Alan Prince and Paul Smolensky, 1993. *Optimality Theory*. MIT Press, Cambridge, Massachusetts.

Uwe Quasthoff, 1998. Tools for Automatic Lexicon Maintenance: Acquisition, Error Correction, and the Generation of Missing Values. In *Proceedings of the First International Conference on Language Resources & Evaluation*.

Randolph Quirk and Jan Svartvik, 1966. *Investigating linguistic acceptability.* Mouton, The Hague.

Adwait Ratnaparkhi, 1996. A Maximum Entropy Part-of-Speech Tagger. In *Proceedings of the Empirical Methods in Natural Language Processing Conference*, University of Pennsylvania.

Veit Reuer, 2000. Error-recognition and parsing of syntactically mildly ill-formed natural language. In *Proceedings of the Lexical Functional Grammar 2000 Conference*, Miriam Butt and Tracy Holloway King, (ed), Berkeley, CA.

Stephen D. Richardson and Lisa C. Braden-Harder, 1988. The Experience of Developing a Large Scale Natural Language Text Processing System: CRITIQUE. In *Proceedings of the Second Conference on Applied Natural Language Processing*, pp. 195–202.

Jon D. Ringen, 1977. On Evaluating Data Concerning Linguistic Intuition. In Eckman (1977), pp. 145–160.

Emmanuel Roche and Yves Schabes, (ed), 1997. *Finite-State Language Processing.* The MIT Press, Cambridge, Massachusetts.

Geoffrey Sampson, 1987. Probabilistic Models of Analysis. In Garside *et al.* (1987), pp. 16–29.

Geoffrey Sampson, 1995. *English for the Computer.* Oxford University Press.

Geoffrey Sampson, 2001a. *Empirical Linguistics.* Continuum, New York.

Geoffrey Sampson, 2001b. Evidence against the Grammatical Ungrammatical Distinction. In *Empirical Linguistics* (Sampson 2001a), Chapter 10, pp. 165–180.

Geoffrey Sampson, 2001c. From Central Embedding to Empirical Linguistics. In *Empirical Linguistics* (Sampson 2001a), Chapter 2, pp. 12–23.

Geoffrey Sampson, 2001d. Objective Evidence is All We Need. In *Empirical Linguistics* (Sampson 2001a), Chapter 8, pp. 122–141.

Geoffrey Sampson and Anna Babarczy, 2002. A Test of the Leaf-Ancestor Metric for Parse Accuracy. In Carroll *et al.* (2002).

David Schneider and Kathleen McCoy, 1998. Recognizing Syntactic Errors in the Writing of Second Language Learners. In *Proceedings of the Thirty-Sixth Annual Meeting of the Association for Computational Linguistics and the Seventeenth International Conference on Computational Linguistics*, Vol. 2, Montreal, Canada.

Carson T. Schütze, 1996. *The Empirical Base of Linguistics: Grammaticality Judgments and Linguistic Methodology.* The University of Chicago Press.

Stuart. M. Shieber, 1986. *An Introduction to Unification-Based Approaches to Grammar.* University of Chicago Press, CSLI Lecture Notes Series.

Khalil Sima'an, 2003. Probabilistic Models for NLP? Empirical Validity and Technological Viability. Distributed as part of the foundational course "Probabilistic Models of NLP", 15th European Summer School in Logic, Language and Information, August 2003, Vienna.

Daniel D. Sleator and Davy Temperley, 1993. Parsing English with a Link Grammar. In *Proceedings of the Third International Workshop on Parsing Technologies*, pp. 277–292, Tilburg (the Netherlands), Durbuy (Belgium).

Catherine Snow and Guus Meijer, 1976. On the Secondary Nature of Syntactic Intuitions. In Greenbaum (1976a), pp. 163–177.

Catherine E. Snow, 1975. Linguists as Behavioral Scientists: Towards a Methodology for Testing Linguistic Intuitions. In *Linguistics in the Netherlands 1972–1973*, A. Kraak, (ed), pp. 271–275. Van Gorcum, Assen.

J.P. Stemberger, 1982. Syntactic Errors in Speech. *Journal of Psycholinguistic Research*, 11-4 (1982), 313–45.

S. Smith Stevens, 1956. The Direct Estimation of Sensory Magnitudes – Loudness. *American Journal of Psychology*, -69 (1956), 1–25.

Mark Stevenson and Yorick Wilks, 2000. Large Vocabulary Word Sense Disambiguation. In *Polysemy*, Ravin and Leacock, (ed), pp. 161–178.

Tomek Strzalkowski, 1992. TTP: a Fast and Robust Parser for Natural Language. In *Proceedings of the 15th International Conference on Computational Linguistics*, pp. 198–204.

Hideto Tomabechi, 1991. Quasi-destructive Graph Unification. In *Proceedings of the 29th annual meeting of the Association for Computational Linguistics*, pp. 315–322, Berkeley, California.

Marcus Tomalin, 2002. The Formal Origins of Syntactic Theory. *Lingua*, 112 (2002), 827–848.

Masaru Tomita, 1990. The Generalized LR Parser/Compiler - Version 8.4. In *Proceedings of the 1990 International Conference on Computational Linguistics*, pp. 59–63, Helsinki, Finland.

Gunnel Tottie, 1976. Variation, Acceptability and the Advanced Foreign Learner: Towards a Sociolinguistics without a Social Context. In Greenbaum (1976a), pp. 203–213.

Kristina Toutanova and Christopher D. Manning, 2002. Feature Selection for a Rich HPSG Grammar Using Decision Trees. In *Proceedings of the Sixth Conference on Natural Language Learning*, Taipei, Taiwan.

Lynne Truss, 2003. *Eats, Shoots and Leaves: The Zero Tolerance Approach to Punctuation.* Profile Books Ltd, Great Britain.

Carl Vogel and Robin Cooper, 1995. Robust Chart Parsing with Mildly Inconsistent Feature Structures. In *Nonclassical Feature Systems*, Andreas Schöter and Carl Vogel, (ed). Centre for Cognitive Science, University of Edinburgh, Working Papers in Cognitive Science. Volume 10.

Atro Voutilainen and Pasi Tapanainen, 1993. Ambiguity Resolution in a Reductionistic Parser. In *Proceedings of the 1993 Conference of the European Chapter of the Association for Computational Linguistics.*

Ye-Yi Wang, 2001. Robust Language Understanding in Mipad. http://research.microsoft.com/srg/papers/2001-yeyiwang-eurospeech.pdf, downloaded 15/7/03.

Thomas Wasow, 1985. Postscript. In *Lectures on Contemporary Syntactic Theories: An Introduction to Government-Binding Theory, Generalized Phrase Structure Grammar, and Lexical-Functional Grammar by Peter Sells*, CSLI Lecture Notes No.3, pp. 193–205. CSLI Publications, Stanford, California.

Ralph M. Weischedel and Norman K. Sondheimer, 1983. Meta-Rules as a Basis for Processing Ill-Formed Input. *American Journal of Computational Linguistics*, 9-3–4 (1983), 161–177.

Ralph M. Weishedel and John E. Black, 1980. Responding Intelligently to Unparsable Inputs. *American Journal of Computational Linguistics*, 6-2 (1980).

Fuliang Weng, 1993. Handling Syntactic Extra-grammaticality. In *Proceedings of the Third International Workshop on Parsing Technologies*, pp. 319–332.

Paul Ziff, 1964. On Understanding "Understanding Utterances". In Fodor and Katz (1964), pp. 390–399.

Good Reasons for Noting Bad Grammar: Empirical Investigations into the Parsing of Ungrammatical Written English

Jennifer Foster

A Thesis submitted for the Degree of Doctor of Philosophy

Department of Computer Science
University of Dublin
Trinity College

Vol. II: Appendices

29 October 2004

Appendix A

Corpus

This appendix contains a list of the ungrammatical *corpus1* sentences and the corrected *corpus2* sentences. Each list item has the following format: the *A* sentence is the ungrammatical sentence from *corpus1*, and the *B* sentence (and, in some cases, *C* sentence) is a corrected *corpus2* version of the *A* sentence. The source of the ungrammatical sentence is then provided. If the source is an email, website, non-academic book, text message, sign or draft of my own writing, the date the error was noticed is given. If the source is a newspaper or magazine article, the publication date is provided. If the source is an academic book or paper, it is cited in the usual way and its full reference is given in the bibliography. Names have been changed in all personal emails and letters.

(A.1) *A.* What we see disturbs us every bit as much as it disturbs children, because we have no idea **to** what to do about it. *B.* What we see disturbs us every bit as much as it disturbs children, because we have no idea what to do about it.
(Evening Herald, 27/9/03, p.11)

(A.2) *A.* He said the Customs and Excise intelligent unit's analysis of shipping contents raised **suspicious** about this consignment and a search was made of the container where it was stored at Foster Avenue, Mount Merrion. *B.* He said the Customs and Excise intelligent unit's analysis of shipping contents raised **suspicions** about this consignment and a search was made of the container where it was stored at Foster Avenue, Mount Merrion.
(Evening Herald, 14/10/03, p.19)

(A.3) *A.* When the jury returned after legal argument today Mr. Justice Butler said DNA evidence alone **could** not **to** be relied upon and pointed to the lack of corroborative evidence. *B.* When the jury returned after legal argument today Mr. Justice Butler said DNA evidence alone could not be relied upon and pointed to the lack of corroborative evidence. *C.* When the jury returned after legal argument today Mr. Justice Butler said DNA evidence alone **was** not to be relied upon and pointed to the lack of corroborative evidence.
(Evening Herald, 14/10/03, p.6)

(A.4) *A.* I've never **own** anybody to sleep as long as him. *B.* I've never **known** anybody to sleep as long as him.
(http://www.channel4.com/bigbrother/news/newsstory_t.jsp?id=1742, 9/6/03)

(A.5) *A.* Readers interested in general **references** grammars of English should waste no time in getting hold of Quirk et al. *B.* Readers interested in general **reference** grammars of English should waste no time in getting hold of Quirk et al.
(Jurafsky and Martin 2000, p.355)

(A.6) *A.* Like the Verb class this Pronoun class will interact **will** the Sentence class. *B.* Like the Verb class this Pronoun class will interact **with** the Sentence class.
(student assignment, Feb 2003)

(A.7) *A.* Worldwide, cable modem technology systems revenue plummeted as cable service providers tightened their belts in 2001, dropping by more than 36 per cent in 2001 and **has** volume in terms of unit shipments and overall revenue has only started to climb back up this year. *B.* Worldwide, cable modem technology systems revenue plummeted as cable service providers tightened their belts in 2001, dropping by more than 36 per cent in 2001 and volume in terms of unit shipments and overall revenue has only started to climb back up this year.
(IT's Monday News electronic newsletter, 18/8/03)

(A.8) *A.* why **is** is good to use young people for speech data *B.* why **it** is good to use young people for speech data
(draft of own writing, 15/8/03)

(A.9) *A.* Judge Angela ni Chondain was told that **that** the two girls, aged 15 and 17, were arrested on August 22 last for theft of the goods from the Boots store in the shopping centre, which they admitted. *B.* Judge Angela ni Chondain was told that the two girls, aged 15 and 17, were arrested on August 22 last for theft of the goods from the Boots store in the shopping centre, which they admitted.
(Evening Herald, 11/10/03, p.14)

(A.10) *A.* an as **possible** as **high** contrast to alternative relations *B.* an as **high** as **possible** contrast to alternative relations
(Menzel 1995)

(A.11) *A.* Drop it into the **technicians** lab if you do. *B.* Drop it into the **technicians'** lab if you do.
(personal email, 29/9/03)

(A.12) *A.* As they all got to know each other better, other housemates came of their shells. *B.* As they all got to know each other better, other housemates came **out** of their shells.
(http://www.channel4.co.uk/bigbrother/news/newsstory_t.jsp?id=1375, 31/5/03)

(A.13) *A.* Lexical rules also contribute to the ease of maintenance of a implementation lexicon. *B.* Lexical rules also contribute to the ease of maintenance of **an** implementation lexicon.
(Fouvry 2003b, p.96)

(A.14) *A.* Jon re-entered the House **with on** a mission to cause some trouble. *B.* Jon re-entered the House **on** a mission to cause some trouble. *C.* Jon re-entered the House **with** a mission to cause some trouble.
(http://bigbrother.channel4.com/bigbrother/index_t.html, 24th July 2003)

(A.15) *A.* Mercury promises a satisfying venture your favourite activities. *B.* Mercury promises a satisfying venture **involving** your favourite activities.
(Evening Herald, 14/10/03, p.28)

(A.16) *A.* But as long as **as** local government is a shambles the situation will only get worse. *B.* But as long as local government is a shambles the situation will only get worse.
(Evening Herald, 27/9/03, p.21)

(A.17) *A.* That allows on the one hand **to ignore** relatively bad solutions, but keeps them as a fall-back option in case the more promising solutions lead to nothing. *B.* That allows **us** on the one hand to ignore relatively bad solutions, but keeps them as a fall-back option in case the more promising solutions lead to nothing. *C.* That allows on the one hand relatively bad solutions to **be ignored**, but keeps them as a fall-back option in case the more promising solutions lead to nothing.
(Fouvry 2003a, p.104)

(A.18) *A.* But no one of them is capable **to deal** with robustness as a whole. *B.* But no one of them is capable **of dealing** with robustness as a whole.
(Ballim *et al.* 1999, p.2)

(A.19) *A.* The JavaChart executor is used to run the chart parsing algorithm on compiled grammar, lexicon and minilexicon. *B.* The JavaChart executor is used to run the chart parsing algorithm on **the** compiled grammar, lexicon and minilexicon. *C.* The JavaChart executor is used to run the chart parsing algorithm on **a** compiled grammar, lexicon and minilexicon.
(User Manual for the JavaChart 0.1.3 system, p.7, downloaded May 2003)

(A.20) *A.* Her teacher has not mentioned anything to us should I go and talk to her and maybe get the child assessed? B. Her teacher has not mentioned anything to us**.** Should I go and talk to her and maybe get the child assessed?
(Evening Herald, p.25, 21/10/03)

(A.21) *A.* Michael tried a variety of jobs before settling into **an** retail company. *B.* Michael tried a variety of jobs before settling into **a** retail company.
(Evening Herald, p.22, 21/10/03)

(A.22) *A.* We now turn the question of incorporating the sentential level recovery strategies we have been discussing into the various approaches to parsing mentioned in the introduction. *B.* We now turn **to** the question of incorporating the sentential level recovery strategies we have been discussing into the various approaches to parsing mentioned in the introduction.
(Carbonell and Hayes 1983, p.130)

(A.23) *A.* He will now be free again to regale us with his views on why we would really all be better off re-reading the speeches of Eamon deValera **and stopped** all this Celtic Tiger nonsense. *B.* He will now be free again to regale us with his views on why we would really all be better off re-reading the speeches of Eamon deValera and **stopping** all this Celtic Tiger nonsense. *C.* He will now be free again to regale us with his views on why we would really all be better off re-reading the speeches of Eamon deValera **having** stopped all this Celtic Tiger nonsense.
(Evening Herald, p.14, 25/11/03)

(A.24) *A.* Table 4 examines some of the false alarms, two of them due to expressions similar to dates that are mistaken for dates, other two relate to constructions not taken into account in the design of the partial grammar, and the last one is due to insufficient lexical coverage. *B.* Table 4 examines some of the false alarms, two of them due to expressions similar to dates that are mistaken for dates, **the** other two relate to constructions not taken into account in the design of the partial grammar, and the last one is due to insufficient lexical coverage. *C.* Table 4 examines some of the false alarms, two of them due to expressions similar to dates that are mistaken for dates, **another** two relate to constructions not taken into account in the design of the partial grammar, and the last one is due to insufficient lexical coverage.
(Gojenola and Oronoz 2000)

(A.25) *A.* If you have problems **to get** the sh scripts to run, then either edit the first line of the scripts to point to your sh location or call sh explicitly. *B.* If you have problems **getting** the sh scripts to run, then either edit the first line of the scripts to point to your sh location or call sh explicitly.
(User Manual for the JavaChart 0.1.3 system, p.8, downloaded May 2003)

(A.26) *A.* But as soon as **of** the strangely creepy Jack started chatting up Kate the audience forgot they could see the strings being pulled. *B.* But as soon as the strangely creepy Jack started chatting up Kate the audience forgot they could see the strings being pulled.
(http://www.channel4.com/apps3/bigbrother/news/newsstory_t.jsp?id=14040, July 2002)

(A.27) *A.* LP constraints are associated with a after a separator. *B.* LP constraints are associated with a **rule** after a separator.
(draft of "Enabling Left Recursive DCGs" by Dr. Carl Vogel, March 2003)

(A.28) *A.* Act decisively, you are a force to be reckoned now. *B.* Act decisively, you are a force to be reckoned **with** now.
(Evening Herald, 15/8/03, p.20)

(A.29) *A.* **It's** little prank had gone too far. **B.** Its little prank had gone too far.
(Sleevenotes, "Senor My Friend", Mumbling Deaf Ro, Feb 2003)

(A.30) *A.* Last weekend, the son of a good friend of mine, a scarce commodity for me some would say, got a text message on his mobile phone asking him if he was interested in joining one of the top **club**. *B.* Last weekend, the son of a good friend of mine, a scarce commodity for me some would say, got a text message on his mobile phone asking him if he was interested in joining one of the top **clubs**.
(Evening Herald, 29/9/03, p.62)

(A.31) *A.* One of the most recent deals for SWORD has been with UK insurer AMP, which **choose** the product based on an assessment of available products and services by Ernst & Young. *B.* One of the most recent deals for SWORD has been with UK insurer AMP, which **chose** the product based on an assessment of available products and services by Ernst & Young.
(IT's Monday News electronic newsletter, 29/9/03)

(A.32) *A.* Moreover, the JavaChart **grammar use** so-called closure expressions in rule bodies. *B.* Moreover, the JavaChart grammar **uses** so-called closure expressions in rule bodies. *C.* Moreover, the JavaChart **grammars** use so-called closure expressions in rule bodies.
(User Manual for the JavaChart 0.1.3 system, p.9, downloaded May 2003)

(A.33) *A.* The emphasis this week is on all matters to do with creativity**.** Self-expression and your general level of vitality. *B.* The emphasis this week is on all matters to do with creativity self-expression and your general level of vitality.
(Evening Herald, p.36, 8/10/03)

(A.34) *A.* He believed there was a major issue in the House, but **he** unable to confront Alex directly, he spoke to Tim man to man. *B.* He believed there was a major issue in the House, but unable to confront Alex directly, he spoke to Tim man to man. *C.* He believed there was a major issue in the House, but **as** he **was** unable to confront Alex directly, he spoke to Tim man to man.
(http://www.channel4.com/apps3/bigbrother/news/newsstory_t.jsp?id=12728, July 2002)

(A.35) *A.* Perhaps their **practises** should be looked at and adopted? *B.* Perhaps their **practices** should be looked at and adopted?
(personal email, 11/12/03)

(A.36) *A.* The techniques for tackling this problem on the level of syntactic representations are already present in the baseline processing **describe** in the previous paragraphs since they are an integrated part of Generalized LR Parsing.
B. The techniques for tackling this problem on the level of syntactic representations are already present in the baseline processing **described** in the previous paragraphs since they are an integrated part of Generalized LR Parsing.
(Ballim *et al.* 1999, p.8)

(A.37) *A.* The accommodation is unfurnished, **comprises of** two double bedrooms main ensuite, generous sized livingroom with balcony. *B.* The accommodation is unfurnished, **comprises** two double bedrooms main ensuite, generous sized livingroom with balcony. *C.* The accommodation is unfurnished, **consists** of two double bedrooms main ensuite, generous sized livingroom with balcony.
(estate agent's electronic newsletter, 9/10/03)

(A.38) *A.* On the top floor there are a further two bedrooms and a bathroom ideal for an **au pairs** use. *B.* On the top floor there are a further two bedrooms and a bathroom ideal for an **au pair's** use.
(estate agent's electronic newsletter, 9/10/03)

(A.39) *A.* Deep correction does, since it involves explaining why some bit of language is the way it is, so prompting **leaners** to reorganize their cognitive structures. *B.* Deep correction does, since it involves explaining why some bit of language is the way it is, so prompting **learners** to reorganize their cognitive structures.
(James 1998, p.237)

(A.40) *A.* If it is, because **of** only six elements are contained in the array PERSON, calculation is necessary. *B.* If it is, because only six elements are contained in the array PERSON, calculation is necessary.
(student assignment, Feb 2003)

(A.41) *A.* This **programmes** purports to give us a **peak** beneath the surface of the music scene at some of the quality therein. *B.* This **programme** purports to give us a **peek** beneath the surface of the music scene at some of the quality therein.
(http://thumpedbbs.com/bbs/, post on Thumped forum, 20/02/03)

(A.42) *A.* It should report to the user what the winning numbers are, and prompt **to** the user to enter their six numbers. *B.* It should report to the user what the winning numbers are, and prompt the user to enter their six numbers.
(draft of own writing, 7/1/04)

(A.43) *A.* Giddes is **a** more quietly confident. *B.* Giddes is a more quietly confident **person**. *C.* Giddes is more quietly confident.
(The Irish Times, The Ticket, p.7, 9/10/03)

(A.44) *A.* For Givens, the win was a pleasing end to a campaign which started with two successive defeats, especially as the Irish boss was making a return to the ground of Neuchatel Xamax, the club he played with for seven seasons in the 1980s, captaining them to a Swiss league success in 1987, but **was** pleased him most was the mixture of positive individual performances and a positive result. *B.* For Givens, the win was a pleasing end to a campaign which started with two successive defeats, especially as the Irish boss was making a return to the ground of Neuchatel Xamax, the club he played with for seven seasons in the 1980s, captaining them to a Swiss league success in 1987, but **what** pleased him most was the mixture of positive individual performances and a positive result.
(Evening Herald, p.62, 11/10/03)

(A.45) *A.* If **its** time for Paidi O Se to go is it time for a few more people to go? *B.* If **it's** time for Paidi O Se to go is it time for a few more people to go?
(Evening Herald, p.68, 8/10/03)

(A.46) *A.* Unless, of course, **the** decide to do the show in a Motel 6 instead of a rainforest. *B.* Unless, of course, **they** decide to do the show in a Motel 6 instead of a rainforest. *C.* Unless, of course, the **producers** decide to do the show in a Motel 6 instead of a rainforest.
(http://www.jumptheshark.com/i/imacelebritygetmeoutofhere.htm, May 2003)

(A.47) *A.* To help **understanding** what weighted feature structures look like, there is one depicted in Figure 7.7 on p.236. *B.* To help **understand** what weighted feature structures look like, there is one depicted in Figure 7.7 on p.236. *C.* To help **in** understanding what weighted feature structures look like, there is one depicted in Figure 7.7 on p.236.
(Fouvry 2003b, p.233)

(A.48) *A.* That image received a further boost when British ambassador presented him his MBE. *B.* That image received a further boost when **the** British ambassador presented him his MBE.
(Evening Herald, p.14, 14/8/03)

(A.49) *A.* Clearly, significant employment **has** created in this area for at least five years, bringing economic benefits to west Belfast. *B.* Clearly, significant employment has **been** created in this area for at least five years, bringing economic benefits to west Belfast. *C.* Clearly, significant employment **was** created in this area for at least five years, bringing economic benefits to west Belfast.
(IT's Monday News electronic newsletter, 22/9/03)

(A.50) *A.* The following **statements** declares and creates an array of five dimensions. *B.* The following **statement** declares and creates an array of five dimensions.
(lecture notes, Feb 2003)

(A.51) *A.* When each new element is tied into already-learned data, and is presented so that pieces of new knowledge introduced together are related conceptually, the learning process gains a more significant meaning and new material is **assimilate** more quickly and entirely. *B.* When each new element is tied into already-learned data, and is presented so that pieces of new knowledge introduced together are related conceptually, the learning process gains a more significant meaning and new material is **assimilated** more quickly and entirely.
(McCoy and Masterman 1997)

(A.52) *A.* A quality entry contains a quality value and the reason **of** the entry, which can be either an algorithm or an expert source. *B.* A quality entry contains a quality value and the reason **for** the entry, which can be either an algorithm or an expert source.
(Quasthoff 1998)

(A.53) *A.* Full details of the schedule may be **view** on the IS Services website. *B.* Full details of the schedule may be **viewed** on the IS Services website.
(personal email, 17/10/03)

(A.54) *A.* Rita said **to** that if he wants to give her a ring or anything to ask about the job, then that's grand. *B.* Rita said **too** that if he wants to give her a ring or anything to ask about the job, then that's grand.
(personal email, August 2003)

(A.55) *A.* Surely it would be much better try and get a better thing going between you. *B.* Surely it would be much better **to** try and get a better thing going between you.
(Evening Herald, p.28, 25/11/03)

(A.56) *A.* If you are willing to play with either a dictionary or bablefish, you can find more recent page here. *B.* If you are willing to play with either a dictionary or bablefish, you can find **a** more recent page here.
(personal email, 9/12/03)

(A.57) *A.* But one common thread running **trough** the various assessments is the huge degree of uncertainty caused by events in the Gulf. *B.* But one common thread running **through** the various assessments is the huge degree of uncertainty caused by events in the Gulf.
(Irish Independent Business Section, 27/3/03)

(A.58) *A.* Topics to be covered include techniques for self-regulation to help you plan, monitor and evaluate **you** progress, time management, building self-discipline and finding out more about the university learning process so that you gain control. *B.* Topics to be covered include techniques for self-regulation to help you plan, monitor and evaluate **your** progress, time management, building self-discipline and finding out more about the university learning process so that you gain control.
(electronic newsletter, 15/1/04)

(A.59) *A.* Ferdinand, who getting over a kidney complaint, is not apparently a man to answer his mobile phone, and when this failed, he was sent a text message. *B.* Ferdinand, who **was** getting over a kidney complaint, is not apparently a man to answer his mobile phone, and when this failed, he was sent a text message.
(Evening Herald, p.4, 8/10/03)

(A.60) *A.* I guess you can mail me here as I know I here til friday anyhow! *B.* I guess you can mail me here as I know I **am** here til friday anyhow! *C.* I guess you can mail me here as I know I **will be** here til friday anyhow!
(personal email, 3/3/03)

(A.61) *A.* Recall that on p.170 **is** it was established that only the last feature of the path was relevant. *B.* Recall that on p.170 it was established that only the last feature of the path was relevant.
(Fouvry 2003b, p.174)

(A.62) *A.* Big Brother **had** earlier **reaffirmed** its message to the housemates that reasons for nomination must be valid may have been a factor in the long duration of the Diary Room sessions. *B.* **The fact that** Big Brother had earlier reaffirmed its message to the housemates that reasons for nomination must be valid may have been a factor in the long duration of the Diary Room sessions. *C.* Big Brother earlier **reaffirming** its message to the housemates that reasons for nomination must be valid may have been a factor in the long duration of the Diary Room sessions.
(http://bigbrother.channel4.com/bigbrother/news/newsstory_t.jsp?id=5127, 15/7/03)

(A.63) *A.* Our long term goal is **get** the revenue breakdown to look more like the industry, with over 50 per cent of total revenue coming from international markets. *B.* Our long term goal is **to** get the revenue breakdown to look more like the industry, with over 50 per cent of total revenue coming from international markets. *C.* Our long term goal is **getting** the revenue breakdown to look more like the industry, with over 50 per cent of total revenue coming from international markets.
(IT's Monday News eletronic newsletter, 12th May 2003)

(A.64) *A.* Will be declaring their undying love for each other? *B.* Will **they** be declaring their undying love for each other? *C.* Will **the pair** be declaring their undying love for each other?
(http://www.channel4.com/apps3/bigbrother/news/newsstory_t.jsp?id=8198, July 2002)

(A.65) *A.* But then maybe they weren't paid **being** a million euros for their gig. *B.* But then maybe they weren't **being** paid a million euros for their gig.
(Irish Independent Review, p.6, 16/8/03)

(A.66) *A.* I ask this not be facetious, but simply because Pilger's once-astounding television documentaries have, over the years, seen him morph from an angry, crusading journalist into some sort of Antipodean Grinch. *B.* I ask this not **to** be facetious, but simply because Pilger's once-astounding television documentaries have, over the years, seen him morph from an angry, crusading journalist

into some sort of Antipodean Grinch.
(Evening Herald, p.23, 27/9/03)

(A.67) *A.* Was **you** supervisor annoyed? *B.* Was **your** supervisor annoyed?
(personal email, August 2002)

(A.68) *A.* **Its** the 'yellow stick it' syndrome where people write down their user name and password or worse still, stick them to the monitor. *B.* **It's** the 'yellow stick it' syndrome where people write down their user name and password or worse still, stick them to the monitor.
(IT's Monday News electronic newsletter, 20/10/03)

(A.69) *A.* Pickthorne gives examples of these bugs made by **leaners** of maths from elementary to college level. *B.* Pickthorne gives examples of these bugs made by **learners** of maths from elementary to college level.
(James 1998, p.59-60)

(A.70) *A.* The accommodation **comprises of** spacious living/dining room with stunning seaviews, separate fully equipped kitchen/breakfastroom. *B.* The accommodation **comprises** spacious living/dining room with stunning seaviews, separate fully equipped kitchen/breakfastroom. *C.* The accommodation **consists of** spacious living/dining room with stunning seaviews, separate fully equipped kitchen/breakfastroom.
(estate agent's electronic newsletter, 9/10/03)

(A.71) *A.* The reader should not be misled by rather negative tone of this section. *B.* The reader should not be misled by **the** rather negative tone of this section.
(Gazdar 2003, p.8)

(A.72) *A.* Of course, the fact that **comedians** work, by their very nature, brings them to clubs and pubs where drink is constantly on offer is a factor. *B.* Of course, the fact that **comedians'** work, by their very nature, brings them to clubs and pubs where drink is constantly on offer is a factor.
(Evening Herald, p.12, 27/9/03)

(A.73) *A.* Analyzing these operations will tell us the ways in which ill-formed sentences tend to deviate in structure from well-formed ones and this information can be used to derive a grammar of ill-formed of well-formed language so that a parser can deal with ungrammatical language in an appropriate. *B.* Analyzing these operations will tell us the ways in which ill-formed sentences tend to deviate in structure from well-formed ones and this information can be used to derive a grammar of ill-formed **language from a grammar** of well-formed language so that a parser can deal with ungrammatical language in an appropriate manner.
(draft of own writing, 31/8/03)

(A.74) *A.* Different dialog states need different **grammar** too. *B.* Different dialog states need different **grammars** too. *C.* Different dialog states need **a** different grammar too.
(Wang 2001)

(A.75) *A.* While Owens was reluctant to give specific sales targets for the coming year he did say that the company was looking **to for** new wins to account for 60 per cent of total revenue by year end, which means quite a few private sector enterprise customers can expect a phone call from BIC over the next few months. *B.* While Owens was reluctant to give specific sales targets for the coming year he did say that the company was looking **for** new wins to account for 60 per cent of total revenue by year end, which means quite a few private sector enterprise customers can expect a phone call from BIC over the next few months. *C.* While Owens was reluctant to give specific sales targets for the coming year he did say that the company was looking **to** new wins to account for 60 per cent of total revenue by year end, which means quite a few private sector enterprise customers can expect a phone call from BIC over the next few months.
(IT's Monday News electronic newsletter, 2/6/03)

(A.76) *A.* Here are also a couple of short accessible references people can read for background **it** they aren't familiar with the ideas yet. *B.* Here are also a couple of short accessible references people can read for background **if** they aren't familiar with the ideas yet.
(personal email, 19/11/03)

(A.77) *A.* But **he's** be all but alone this week in Renards because most of the club's other celebrity regulars are down in South Africa performing a concert for former South African President Nelson Mandela. *B.* But **he'll** be all but alone this week in Renards because most of the club's other celebrity regulars are down in South Africa performing a concert for former South African President Nelson Mandela. *C.* But **he's going to** be all but alone this week in Renards because most of the club's other celebrity regulars are down in South Africa performing a concert for former South African President Nelson Mandela.
(Evening Herald, p.28, 28/11/03)

(A.78) *A.* The event led to the incongruous **site** of large record company bouncers and PR ladies dominating and taking over the entrance as if the diplomatic mission was a nightclub, rather than a sophisticated diplomatic outpost. *B.* The event led to the incongruous **sight** of large record company bouncers and PR ladies dominating and taking over the entrance as if the diplomatic mission was a nightclub, rather than a sophisticated diplomatic outpost.
(Evening Herald, p.3., 27/11/03)

(A.79) *A.* Step 4 in the recovery algorithm means that only one error rule is **every** considered during any one parse attempt, with the result that this algorithm will not handle sentences containing two or more errors. *B.* Step 4 in the recovery algorithm means that only one error rule is **ever** considered during any one parse attempt, with the result that this algorithm will not handle sentences containing two or more errors.
(draft of own writing, 5/9/03)

(A.80) *A.* They are looking **or** longer life cycles from their product and all at a fraction of the cost. *B.* They are looking **for** longer life cycles from their product and all at a fraction of the cost.
(IT's Monday news electronic newsletter, 28th July 2003)

(A.81) *A.* We have a couple of orders going out to Dell so we could get one for **your** in that batch or Mike K. could do the ordering. *B.* We have a couple of orders going out to Dell so we could get one for **you** in that batch or Mike K. could do the ordering.
(personal email, 14/10/03)

(A.82) *A.* modifications to the expansion **that** neither **transform** the grammar with loss of strong equivalence, nor **involving** depth bounded search **nor** a change in parsing paradigm from top-down depth-first search *B.* modifications to the expansion that neither transform the grammar with loss of strong equivalence, nor **involve** depth bounded search **or** a change in parsing paradigm from top-down depth-first search *C.* modifications to the expansion neither **transforming** the grammar with loss of strong equivalence, nor **involve** depth bounded search **or** a change in parsing paradigm from top-down depth-first search
(draft of "Enabling Left Recursive DCGs" by Dr. Carl Vogel, March 2003)

(A.83) *A.* Of course he is shot down immediately, and after a tongue lashing is warned to stop **been** so childish. *B.* Of course he is shot down immediately, and after a tongue lashing is warned to stop **being** so childish.
(Evening Herald, p.41, 21/10/03)

(A.84) *A.* This **in** not the UK. B. This **is** not the UK.
(personal email, 4/4/03)

(A.85) *A.* Far from looking as though their brains were in the process of being **picked** however, the gang who kept running with the game genuinely appeared to be enjoying themselves. *B.* Far from looking as though their brains were in the process of being **pickled** however, the gang who kept running with the game genuinely appeared to be enjoying themselves.
(http://www.channel4.com/bigbrother/news/newsstory.jsp?id=1578, June 2003)

(A.86) *A.* Technology is eliminating or commoditising many jobs that are based on holding and deseminating information and this **effects** most if not not all of the information oriented careers that parents and career teachers look to for job security and, of course, above average salaries for students. *B.* Technology is eliminating or commoditising many jobs that are based on holding and deseminating information and this **affects** most if not all of the information oriented careers that parents and career teachers look to for job security and, of course, above average salaries for students.
(personal email, 8/12/03)

(A.87) *A.* A complete disambiguation of structural descriptions has first been described for Constraint Dependency Grammar and simply requires to replace the monadic categories of CG by pairs consisting of the relation name and the exactly specified modifee. *B.* A complete disambiguation of structural descriptions has first been described for Constraint Dependency Grammar and simply requires **us** to replace the monadic categories of CG by pairs consisting of the relation name and the exactly specified modifee.*C.* A complete disambiguation of structural descriptions has first been described for Constraint Dependency Grammar and simply requires **that** the monadic categories of CG **be replaced** by pairs consisting of the relation name and the exactly specified modifee.
(Menzel 1995)

(A.88) *A.* I do explain things but he has to be convinced and to discover these things **things** himself. *B.* I do explain things but he has to be convinced and to discover these things himself.
(Evening Herald, p.89, 1/12/03)

(A.89) *A.* Cabin Fever should have come with a Government health warning that you might be asked to **fun** for Fine Gael if you win. *B.* Cabin Fever should have come with a Government health warning that you might be asked to **run** for Fine Gael if you win.
(Evening Herald, p.24, 12/11/03)

(A.90) *A.* This was explained as being due to the sale of a number of businesses, including Horizon's Cisco Systems training division, which resulted in costs **beings** halved. *B.* This was explained as being due to the sale of a number of businesses, including Horizon's Cisco Systems training division, which resulted in costs **being** halved.
(IT's Monday News electronic newsletter, 8/9/03)

(A.91) *A.* Want to **saving** money?*B.* Want to **save** money? *C.* Want to **start** saving money?
(header of junk mail, July 2002)

(A.92) *A.* The tests of the first 500 members of the defence forces resulted in two being discharged after **the** they tested positive. *B.* The tests of the first 500 members of the defence forces resulted in two being discharged after they tested positive.
(Evening Herald, p.8, 11/8/03)

(A.93) *A.* Anyway, then he mailed me back **with** in sarcastic tones telling me what it meant and all. *B.* Anyway, then he mailed me back in sarcastic tones telling me what it meant and all.
(personal email, 3/9/02)

(A.94) *A.* You feel a little vulnerable, and **its** up to family members to give support. *B.* You feel a little vulnerable, and **it's** up to family members to give support.
(Evening Herald, p.36, 8/10/03)

(A.95) *A.* The plan is to increase employee numbers to 11 over **next three years** and generate revenues of STG 6 million during that time. *B.* The plan is to increase employee numbers to 11 over **the next three years** and generate revenues of STG 6 million during that time. *C.* The plan is to increase employee numbers to 11 over **three years** and generate revenues of STG 6 million during that time.
(IT's Monday News electronic newsletter, 17/11/03)

(A.96) *A.* Does it make **sentence** to take a sentence out of context? *B.* Does it make **sense** to take a sentence out of context?
(draft of own writing, 28/7/03)

(A.97) *A.* He expects telcos and mobile operators to dominate the market with the current new breed of service providers either being acquired or coming to dominate niche markets that the big players consider **to** small to service. *B.* He expects telcos and mobile operators to dominate the market with the current new breed of service providers either being acquired or coming to dominate niche markets that the big players consider **too** small to service.
(IT's Monday News electronic newsletter, 1/9/03)

(A.98) *A.* I'd be worried about is the way she'll glaringly completely contradict herself in everything she does without realising it. *B.* **What** I'd be worried about is the way she'll glaringly completely contradict herself in everything she does without realising it.
(http://www.channel4.com/bigbrother/news/newsstory_t.jsp?id=2882, 19th June 2003)

(A.99) *A.* The hope must also be, that within Paisley's Democratic Unionists **their** lurks some reasonable, realistic consensus politicians. *B.* The hope must also be, that within Paisley's Democratic Unionists **there** lurks some reasonable, realistic consensus politicians.
(Evening Herald, p.14, Monday 1/12/03)

(A.100) *A.* In the context in which the sentence occurred it is clear that the word **to that needs** to be added is they. *B.* In the context in which the sentence occurred it is clear that the word **that needs** to be added is they. *C.* In the context in which the sentence occurred it is clear that the word **to be added** is they.
(draft of own writing, 22/10/03)

(A.101) *A.* If you want to find out what's really going, it's a lot harder. *B.* If you want to find out what's really going **on**, it's a lot harder.
(http://www.thumped.com/bbs/showthread.php?+=9918, 11/11/03)

(A.102) *A.* Discourse information considered here includes patterns of various **kind** repeated frequently in the same text.*B.* Discourse information considered here includes patterns of various **kinds** repeated frequently in the same text.
(Ballim *et al.* 1999)

(A.103) *A.* Asks Curly for a divorce and **tell** him he had a daughter. *B.* Asks Curly for a divorce and **tells** him he had a daughter.
(Evening Herald, p.17, 29/9/03)

(A.104) *A.* So the only way we could feel at ease was by going abroad together to secret **hideaway**. *B.* So the only way we could feel at ease was by going abroad together to **a** secret hideaway. *C.* So the only way we could feel at ease was by going abroad together to secret **hideaways.**
(News of the World, p.7, 9/3/03)

(A.105) *A.* There is no reason **the** believe these represent different types of information, rather than different ways of conceptualising and encoding it. *B.* There is no reason **to** believe these represent different types of information, rather than different ways of conceptualising and encoding it.
(Stevenson and Wilks 2000)

(A.106) *A.* Start tying up the **lose** ends in your life. *B.* Start tying up the **loose** ends in your life.
(Evening Herald, p. 28, 18/8/03)

(A.107) *A.* It was one part Del Boy, one part Minder and the rest Fair City, with Roy **form** Corrie parachuted in as the groundsman. *B.* It was one part Del Boy, one part Minder and the rest Fair City, with Roy **from** Corrie parachuted in as the groundsman.
(Evening Herald, p.69, 15/8/03)

(A.108) *A.* When it seems necessary to make assumptions of independence between the modules, these **assumption** are stated explicitly and rigorously. *B.* When it seems necessary to make assumptions of independence between the modules, these **assumptions** are stated explicitly and rigorously.
(Sima'an 2003, p.11)

(A.109) *A.* Maybe you threw the odd hairy eyelash at one of **you** dad's mates, or had an innocent flirt with some young student teacher who was easy prey. *B.* Maybe you threw the odd hairy eyelash at one of **your** dad's mates, or had an innocent flirt with some young student teacher who was easy prey.
(U Magazine, November 2003)

(A.110) *A.* Delegates were asked what **where** the biggest threats to any company's security system, and the answer from delegates was apathy and users. *B.* Delegates were asked what **were** the biggest threats to any company's security system, and the answer from delegates was apathy and users.
(IT's Monday News electronic newsletter, 15/9/03)

(A.111) *A.* Finally, a research-based research-oriented design model will be presented, which not only allows **to solve** the above mentioned problem, but which can also be used for language courseware development and evaluation. *B.* Finally, a research-based research-oriented design model will be presented, which not only allows **us** to solve the above mentioned problem, but which can also be used for language courseware development and evaluation. *C.* Finally, a research-based research-oriented design model will be presented, which not only allows the above mentioned problem to **be solved**, but which can also be used for language courseware development and evaluation.
(DCLRS electronic newsletter, 12/11/03)

(A.112) *A.* An e-mail with FTP details will be **send** to you shortly. *B.* An e-mail with FTP details will be **sent** to you shortly.
(http://gate.ac.uk/cgi-bin/gate-reg-new, 5/11/03)

(A.113) *A.* Also later on in it, she **others** a Cosmo from the bar, and you can see the back of my head at the other side of the bar and a glimpse of Johnny who I'm talking **too**. B. Also later on in it, she **orders** a Cosmo from the bar, and you can see the back of my head at the other side of the bar and a glimpse of Johnny who I'm talking **to**.
(http://thumpedbbs.com/bbs/, post on Thumped forum, 20th February 2003)

(A.114) *A.* In order to express the tendency that the above substrings occur with a specific structure, regardless the instantiations for VB and IN, we want to have a generalization over the verbs and prepositions but not over the noun information. *B.* In order to express the tendency that the above substrings occur with a specific structure, regardless **of** the instantiations for VB and IN, we want to have a generalization over the verbs and prepositions but not over the noun information.
(Bod 1998, p.62)

(A.115) *A.* **Got** that sounds so shit. *B.* **God** that sounds so shit.
(personal email, 4/4/03)

(A.116) *A.* This task seems to be the combination of the first two **task** where the generation phase is replaced by the answer computation and its translation into a natural language sentence. *B.* This task seems to be the combination of the first two **tasks** where the generation phase is replaced by the answer computation and its translation into a natural language sentence.
(Ballim *et al.* 1999, p.7)

(A.117) *A.* It allows to yield an at least vague interpretation even in cases of extremely distorted input.*B.* It allows **us** to yield an at least vague interpretation even in cases of extremely distorted input.*C.* It allows an at least vague interpretation **to be yielded** even in cases of extremely distorted input.
(Menzel 1995)

(A.118) *A.* **Implementation** the algorithm *B.* Implementation **of** the algorithm *C.* **Implementing** the algorithm
(student assignment, Feb 2003)

(A.119) *A.* I had to laugh at the various newspapers assembling lists of people who Leeds are interested in the job. *B.* I had to laugh at the various newspapers assembling lists of people who Leeds are interested in **for** the job.
(Evening Herald, p.74, 12/11/03)

(A.120) *A.* You'll have to tell me all about your weekend then but **we we'll** make sure that we get a few games in too. *B.* You'll have to tell me all about your weekend then but **we'll** make sure that we get a few games in too. *C.* You'll have to tell me all about your weekend then but **we will** make sure that we get a few games in too.
(personal email, 6/8/03)

(A.121) *A.* Industry analysts see the deal as **a** worrying for Intel and very good news for AMD. *B.* Industry analysts see the deal as worrying for Intel and very good news for AMD. *C.* Industry analysts see the deal as a worrying **one** for Intel and very good news for AMD.
(IT's Monday News electronic newsletter, 10/11/03)

(A.122) *A.* It's probably safe bet to say that the Bush administration and associated religious crazies scare the bejaysus out of all of us. *B.* It's probably **a** safe bet to say that the Bush administration and associated religious crazies scare the bejaysus out of all of us.
(Evening Herald, p.23, 27/9/03)

(A.123) *A.* HPSG succeeds integrating several of them elegantly into a uniform description. *B.* HPSG succeeds **in** integrating several of them elegantly into a uniform description.
(Fouvry 2003b, p.72)

(A.124) *A.* For example, the users may have different and changing contact **list**, therefore the parser should dynamically modify the rule for the contacts of different users. *B.* For example, the users may have different and changing contact **lists**, therefore the parser should dynamically modify the rule for the contacts of different users. *C.* For example, the users may have **a** different and changing contact list, therefore the parser should dynamically modify the rule for the contacts of different users.
(Wang 2001)

(A.125) *A.* Please **not** that Jane Smith is not entitled to statutory preserved benefit since the requirements of the 2002 Pensions Act are not satisfied. *B.* Please **note** that Jane Smith is not entitled to statutory preserved benefit since the requirements of the 2002 Pensions Act are not satisfied.
(personal letter, March 2003)

(A.126) *A.* points the compiled grammar file *B.* points **to** the compiled grammar file *C.* points **out** the compiled grammar file
(User Manual for the JavaChart 0.1.3 system, p.8, downloaded May 2003)

(A.127) *A.* talk **to** later anyhow *B.* talk **to you** later anyhow *C.* talk later anyhow
(personal email, 12/1/04)

(A.128) *A.* The command **have** only optional arguments. *B.* The command **has** only optional arguments.
(User Manual for the JavaChart 0.1.3 system, p.7, downloaded May 2003)

(A.129) *A.* She first noticed his change of heart when she was slagging **of** the new Conservative candidate in their area as an Aryan android who'd never seen a working-class person, much less a dole queue. *B.* She first noticed his change of heart when she was slagging **off** the new Conservative candidate in their area as an Aryan android who'd never seen a working-class person, much less a dole queue.
(It's Friday magazine, p.23, Evening Herald, 26/9/03)

(A.130) *A.* This **programs** calculates the net annual salary when the gross annual salary is 25,000.50 euro and the tax rate is 22%. *B.* This **program** calculates the net annual salary when the gross annual salary is 25,000.50 euro and the tax rate is 22%.
(lecture notes, October 2003)

(A.131) *A.* Jeff wants me to change a lot of **thing** so they're politically correct. *B.* Jeff wants me to change a lot of **things** so they're politically correct.
(personal email, Feb 2003)

(A.132) *A.* **There is** long time **ago** I **am** very interested in Welsh language. *B.* **For a** long time I **have been** very interested in **the** Welsh language.
a letter quoted in James (1998, p.268)

(A.133) *A.* how information **much** can be lost *B.* how **much** information can be lost
(Fouvry 2003a, p.110)

(A.134) *A.* The abstracts for last **years** talks can be found from the postgrad website. *B.* The abstracts for last **years'** talks can be found from the postgrad website.
(personal email, 23/10/03)

(A.135) *A.* Me and james always seem to have a **have a** clash of wills or something and i always seem to be having them with john too. *B.* Me and james always seem to have a clash of wills or something and i always seem to be having them with john too.
(personal email, Jan 2003)

(A.136) *A.* Topics to be covered include techniques for self-regulation to help you plan, monitor and evaluate **you** progress, time management, building self-discipline and finding out more about the university learning process so that you gain control. *B.* Topics to be covered include techniques for self-regulation to help you plan, monitor and evaluate **your** progress, time management, building self-discipline and finding out more about the university learning process so that you gain control.
(electronic newsletter, 14/10/03)

(A.137) *A.* In this perspective a naive solution towards robustness may be to improve accuracy **a** first stages of the pipeline in order to minimize the overall inaccuracy.*B.* In this perspective a naive solution towards robustness may be to improve accuracy **at** first stages of the pipeline in order to minimize the overall inaccuracy.
(Ballim *et al.* 1999, p.14)

(A.138) *A.* The structure of the **methods** getter and setter methods apply for the right subtree. *B.* The structure of the getter and setter methods apply for the right subtree.
(student assignment, 2/4/03)

(A.139) *A.* We started off the day with a speed boat ride up the river then we were dropped **of** so we could Kayak back down. *B.* We started off the day with a speed boat ride up the river then we were dropped **off** so we could Kayak back down.
(personal email, 19/4/03)

(A.140) *A.* So, I am writing **these** missive **for asking** whether you promote courses of Welsh (at the University College of North Wales, or by correspondence, like the Open University of London) for non-english speakers. *B.* So, I am writing **this** missive **to ask** whether you promote courses of Welsh (at the University College of North Wales, or by correspondence, like the Open University of London) for non-english speakers.
a letter quoted in James (1998, p.268)

(A.141) *A.* If the structure of the sentence **obey's** the structural rules of the language in question then the sentence will be accepted. *B.* If the structure of the sentence **obeys** the structural rules of the language in question then the sentence will be accepted.
(draft of own writing, 18/8/03)

(A.142) *A.* Exchanging linguistic information among different linguistic levels has been shown to be worthwhile in extending NLP **system** in a robust way. *B.* Exchanging linguistic information among different linguistic levels has been shown to be worthwhile in extending NLP **systems** in a robust way.*C.* Exchanging linguistic information among different linguistic levels has been shown to be worthwhile in extending **an** NLP system in a robust way.
(Ballim *et al.* 1999, p.8)

(A.143) *A.* In the first three minutes you **loose** all your inhibitions. *B.* In the first three minutes you **lose** all your inhibitions.
(Evening Herald, p.17, 27/9/03)

(A.144) *A.* If they do not, or if their feelings are neutral unusual with them they relate to fellow workers only as leaders. *B.* If they do not, or if their feelings are neutral **-** unusual with them **-** they relate to fellow workers only as leaders.
(http://www.crystalinks.com/astrology3.html, Feb 2003)

(A.145) *A.* In addition to the suggested inflection class we find existing inflected forms of low frequency which do **nor** fit into inflection class. *B.* In addition to the suggested inflection class we find existing inflected forms of low frequency which do **not** fit into the inflection class.
(Quasthoff 1998)

(A.146) *A.* And, as he made clear in last week's speech to the American Enterprise Institute, it would be an earnest of America's commitment to the democratic transformation of the Middle East, which would help to solve the Israel-Palestine problem. *B.* And, as he made clear in last week's speech to the American Enterprise Institute, it would be an earnest **sign** of America's commitment to the democratic transformation of the Middle East, which would help to solve the Israel-Palestine problem.
(The Economist, p.52, 8/3/03)

(A.147) *A.* The question is, at the end of eight episodes, **with** The Clinic itself be in need of life-saving resuscitation? *B.* The question is, at the end of eight episodes, **will** The Clinic itself be in need of life-saving resuscitation?
(Evening Herald, 10/10/03, It's Friday supplement, p.10)

(A.148) *A.* Could Rhona Williams please contact **from** office ASAP. *B.* Could Rhona Williams please contact **front** office ASAP.
(personal email, 6/11/03)

(A.149) *A.* At **one least** argument in one parameter list differs in its type from the corresponding argument in the other parameter list. *B.* At **least one** argument in one parameter list differs in its type from the corresponding argument in the other parameter list.
(lecture notes, Feb 2003)

(A.150) *A.* Even if things pick up, as they might, the government parties are getting the **wrap** for mistakes they have patently made themselves. *B.* Even if things pick up, as they might, the government parties are getting the **rap** for mistakes they have patently made themselves.
(Evening Herald, p.10, 27/9/03)

(A.151) *A.* Parlour games are a simple and fun way of passing the time, unless **you're** name is Tania that is. *B.* Parlour games are a simple and fun way of passing the time, unless **your** name is Tania that is.
(http://www.channel4.com/bigbrother/, June 2003)

(A.152) *A.* The reading material **of one** of the corpus compiler was the source for the error data. *B.* The reading material of the corpus compiler was the source for the error data.
(draft of own writing, 1/9/03)

(A.153) *A.* Free-scoring youngsters don't always develop into senior internationals of class, as Irish football history has shown, but the strikers in Ireland's U21 team are certainly giving senior team boss Brian Kerr some food for thought as the latest **player** to come off the conveyor belt. B. Free-scoring youngsters don't always develop into senior internationals of class, as Irish football history has shown, but the strikers in Ireland's U21 team are certainly giving senior team boss Brian Kerr some food for thought as the latest **players** to come off the conveyor belt.
(Evening Herald, p.62, 11/10/03)

(A.154) *A.* For **these type** of **sentences** it is necessary to first know what the sentence means before its structure becomes clear. *B.* For these **types** of sentences it is necessary to first know what the sentence means before its structure becomes clear. *C.* For **this** type of **sentence** it is necessary to first know what the sentence means before its structure becomes clear.
(draft of own writing, 18/8/03)

(A.155) *A.* The degree to which the most probable analyses generated by the system match with the test set analyses is a measure **for** the accuracy of the system. *B.* The degree to which the most probable analyses generated by the system match with the test set analyses is a measure **of** the accuracy of the system.
(Bod 1998, p.9)

(A.156) *A.* All aspects of English surface and logical grammar which are definite enough to be susceptible **of** explicit annotation are indicated. *B.* All aspects of English surface and logical grammar which are definite enough to be susceptible **to** explicit annotation are indicated.
(Sampson 2001a, p.41)

(A.157) *A.* The letters are revealed by ex-royal butler Paul Burrell in a new book in which he also claims Princess Diana predicted her death by saying there was plot to tamper with her car. *B.* The letters are revealed by ex-royal butler Paul Burrell in a new book in which he also claims Princess Diana predicted her death by saying there was **a** plot to tamper with her car.
(Evening Herald, p.3, 21/10/03)

(A.158) *A.* The made-up ungrammaticalities that the **linguist use** will rarely be encountered by the parser. *B.* The made-up ungrammaticalities that the linguist **uses** will rarely be encountered by the parser. *C.* The made-up ungrammaticalities that the **linguists** use will rarely be encountered by the parser.
(draft of my own writing, 20/8/03)

(A.159) *A.* The latest news is that movie bosses in Hollywood **is** blaming some poor box office on youngsters texting their friends to say just how bad a film is. *B.* The latest news is that movie bosses in Hollywood **are** blaming some poor box office on youngsters texting their friends to say just how bad a film is.
(Evening Herald, p. 14, 18/8/03)

(A.160) *A.* Will you be in **you** office later on this evening? *B.* Will you be in **your** office later on this evening?
(personal email, 12/1/04)

(A.161) *A.* Cam are you still speaking to Jon now you what he thought of you? *B.* Cam are you still speaking to Jon now you **know** what he thought of you?
(http://www.channel4.com/community/showcards/B/Big_Brother_4_-_Cameron_and_Ray.html, 31/7/03)

(A.162) *A.* Various **report** last week suggested that Amazon.com is in talks with IDA Ireland with a view to setting up a European operations centre in Dublin. *B.* Various **reports** last week suggested that Amazon.com is in talks with IDA Ireland with a view to setting up a European operations centre in Dublin.
(IT's Monday News electronic newsletter, 8/12/03)

(A.163) *A.* John bought the 'this is just a modern rock song' ep last week and **its** really good. *B.* John bought the 'this is just a modern rock song' ep last week and **it's** really good.
(personal email, 3/12/02)

(A.164) *A.* Judge John Murray said that having regard to the gravity of the offences and notwithstanding matters which had been put forward in mitigation **of** behalf of Byrne, the CCA was satisfied that the sentences imposed in the Circuit Court were not only lenient but unduly lenient. *B.* Judge John Murray said that having regard to the gravity of the offences and notwithstanding matters which had been put forward in mitigation **on** behalf of Byrne, the CCA was satisfied that the sentences imposed in the Circuit Court were not only lenient but unduly lenient.
(Evening Herald, p.18, 21/10/03)

(A.165) *A.* Misspellings **and other form or other form** of ill-formedness *B.* Misspellings and other forms of ill-formedness *C.* Misspellings or other forms of ill-formedness
(Ballim *et al.* 1999, p.6)

(A.166) *A.* They **where** only trying to rob stuff we found out the next day as they had managed to lift stuff from the room next to us without disturbing anyone. *B.* They **were** only trying to rob stuff we found out the next day as they had managed to lift stuff from the room next to us without disturbing anyone.
(personal email, 18/5/03)

(A.167) *A.* Even given the context in which the above sentence occurs there **is** still **remains** some ambiguity. *B.* Even given the context in which the above sentence occurs there still **remains** some ambiguity. *C.* Even given the context in which the above sentence occurs there **is** still some ambiguity.
(draft of own writing, 22/10/03)

(A.168) *A.* **Their** is the clg@cs.tcd.ie address too, but I'm not sure if they're on it yet. *B.* **There** is the clg@cs.tcd.ie address too, but I'm not sure if they're on it yet.
(personal email, 3/11/03)

(A.169) *A.* The excitement over SAP's preliminary numbers stem from its license revenue outlook that **look** to best even the most optimistic analyst expectations of EUR 416 million, as polled by Reuters. *B.* The excitement over SAP's preliminary numbers stems from its license revenue outlook that **looks** to best even the most optimistic analyst expectations of EUR 416 million, as polled by Reuters.
(IT's Monday News electronic newsletter, 13/10/03)

(A.170) *A.* In the rest of this paper we will describe this parser and our approach to **evaluate** its performance. *B.* In the rest of this paper we will describe this parser and our approach to **evaluating** its performance. *C.* In the rest of this paper we will describe this parser and our approach to **the evaluation of** its performance.
(Ehrlich and Hanrieder 1996, p.27)

(A.171) *A.* Having taken over the 10 year old restaurant in May, the weekend DART cancellation is the last thing he needed facing into the **quite** winter period. *B.* Having taken over the 10 year old restaurant in May, the weekend DART cancellation is the last thing he needed facing into the **quiet** winter period.
(Evening Herald, p.12, 8/10/03)

(A.172) *A.* The Department of Transport is set to **set to** make more motoring offences liable for penalty points on a phased basis. *B.* The Department of Transport is set to make more motoring offences liable for penalty points on a phased basis.
(Evening Herald, p.6, 12/11/03)

(A.173) *A.* All these characteristics made more difficult to create a proper error classification in Basque. *B.* All these characteristics made **it** more difficult to create a proper error classification in Basque. (Aduriz *et al.* 2002)

(A.174) *A.* Microsoft cannot collect information about a **users** system without first notifying the user. *B.* Microsoft cannot collect information about a **user's** system without first notifying the user. (IT's Monday News electronic newsletter, 15/9/03)

(A.175) *A.* It is convenient to use the setter method from the syntaxClass, making it **more clearer** which subtrees are set to what nodes. *B.* It is convenient to use the setter method from the syntaxClass, making it **clearer** which subtrees are set to what nodes. *C.* It is convenient to use the setter method from the syntaxClass, making it **more clear** which subtrees are set to what nodes. (student assignment, 2/4/03)

(A.176) *A.* This might be well **know**. *B.* This might be well **known**. (http://www.thumped.com/bbs/showthread.php?+=9918, 11/11/03)

(A.177) *A.* While portal technology is useful **is** will only ever have a niche market because it does not change business processes. *B.* While portal technology is useful **it** will only ever have a niche market because it does not change business processes. (IT's Monday News electronic newsletter, 8/12/03)

(A.178) *A.* Professional working, smoking or non smoking, good humoured, pleasant and tidy individuals required to share house with two other **professional's** one male & one female. *B.* Professional working, smoking or non smoking, good humoured, pleasant and tidy individuals required to share house with two other **professionals** one male & one female. (estate agent's electronic newsletter, 10/10/03)

(A.179) *A.* Almost 50% of the errors which can be corrected by replacing one word (henceforth known as replacement errors involve a change in syntactic category. *B.* Almost 50% of the errors which can be corrected by replacing one word (henceforth known as replacement errors**)** involve a change in syntactic category. (draft of own writing, 23/10/03)

(A.180) *A.* Total revenues are expected **to** about EUR 1.650 billion. *B.* Total revenues are expected to **be** about EUR 1.650 billion. *C.* Total revenues are expected **of** about EUR 1.650 billion. (IT's Monday News electronic newsletter, 13/10/03)

(A.181) *A.* Tango Telecom's decision to take a direct route to market to back up its existing indirect sales strategy has paid off with **of** two lucrative deals. *B.* Tango Telecom's decision to take a direct route to market to back up its existing indirect sales strategy has paid off with two lucrative deals. (IT's Monday news electronic newsletter, 3/11/03)

(A.182) *A.* He is on way up. *B.* He is on **the** way up. (Evening Herald, p.23, 21/10/03)

(A.183) *A.* Jonathon Wrather, who played wicked Joe Carter in Coronation St is big band crooner, Harry Connick Jnr, and Dave Neilson who plays Roy **or** corner-cafe fame appears as Roy Orbison. *B.* Jonathon Wrather, who played wicked Joe Carter in Coronation St is big band crooner, Harry Connick Jnr, and Dave Neilson who plays Roy **of** corner-cafe fame appears as Roy Orbison. (Evening Herald, p.31, 23/8/03)

(A.184) *A.* During one of several conversations that showed how difficult it is to be fed a bit of information from the outside world and not be able to deal with it**.** Nush told Scott, as they kept a safe distance from each other on the bed, how she found not having seen her boyfriend for long periods of time before she came into the house. *B.* During one of several conversations that showed how difficult it is to be fed a bit of information from the outside world and not be able to deal with it Nush told Scott, as they kept a safe distance from each other on the bed, how she found not having seen her boyfriend for long periods of time before she came into the house. (http://bigbrother.channel4.com/bigbrother/news/newsstory_t.jsp?id=5118, 14th July 2003)

(A.185) *A.* Shares in Europe's largest telecoms carrier traded up 1.25 per cent at EUR 13.75 after the targets were **announce** this morning. *B.* Shares in Europe's largest telecoms carrier traded up 1.25 per cent at EUR 13.75 after the targets were **announced** this morning. (IT's Monday News electronic newsletter, 17/11/03)

(A.186) *A.* What **do** do when one leaves Man U? *B.* What **to** do when one leaves Man U? *C.* What does one do when one leaves Man U? (Evening Herald, p. 20, 25/11/03)

(A.187) *A.* This **is** offering **includes** all the great products and features described for Office XP Small Business plus two more two popular business applications: PowerPoint 2002 and Access 2002. *B.* This offering **includes** all the great products and features described for Office XP Small Business

plus two more two popular business applications: PowerPoint 2002 and Access 2002. *C.* This **is** offering all the great products and features described for Office XP Small Business plus two more two popular business applications: PowerPoint 2002 and Access 2002.
(http://www.euro.dell.com/countries/ie/enu/dhs/local/topics/gen_006_inspn_guide.htm, 18/8/03)

(A.188) *A.* Whitespace **consists** of either spaces, blank lines or line breaks **are** ignored. *B.* Whitespace **consisting** of either spaces, blank lines or line breaks **is** ignored. *C.* Whitespace **which** consists of either spaces, blank lines or line breaks **is** ignored.
(ALE User Guide, Version 3.2.1. December 2001)

(A.189) *A.* I'm not talking about players but **lets** look at the executive in there. *B.* I'm not talking about players but **let's** look at the executive in there.
(Evening Herald, p.68, 8/10/03)

(A.190) *A.* Melissa P probably **with out** significant other *B.* Melissa P probably **without** significant other.
(personal email, 26th July 2002)

(A.191) *A.* Also, **Its** kind of a fun class to teach! *B.* Also, **It's** kind of a fun class to teach!
(personal email, Aug 2002)

(A.192) *A.* What it sells: fashion clothing, belts, bags, **umbrella**, photography, fine are, ceramics and jewellery. *B.* What it sells: fashion clothing, belts, bags, **umbrellas**, photography, fine are, ceramics and jewellery.
(Evening Herald, 14/10/03, p.46)

(A.193) *A.* One of them then inadvertently threw Gos the worst insult a young man can have, by saying that he was the only male housemate that **a she** could risk having a cuddle with because, in his case, it would be clear that there was nothing sexual in it. *B.* One of them then inadvertently threw Gos the worst insult a young man can have, by saying that he was the only male housemate that **she** could risk having a cuddle with because, in his case, it would be clear that there was nothing sexual in it. *C.* One of them then inadvertently threw Gos the worst insult a young man can have, by saying that he was the only male housemate that **a girl** could risk having a cuddle with because, in his case, it would be clear that there was nothing sexual in it.
(http://www.channel4.com/bigbrother/news/newsstory_t.jsp?id=2144, 11/6/03)

(A.194) *A.* And i'd like to see john brown **whose** playing with them, since i've heard such good things about him. *B.* And i'd like to see john brown **who's** playing with them, since i've heard such good things about him.
(personal email, May 2003)

(A.195) *A.* When the Crystal Sales office last October 2002 it was envisaged that it would expand the sales operation to service new business and its existing customers. *B.* When the Crystal Sales office **opened** last October 2002 it was envisaged that it would expand the sales operation to service new business and its existing customers.
(IT's Monday News electronic newsletter, 11/8/03)

(A.196) *A.* As it was Sunday the river was lined with Thai families having picnics in the bamboo picnic huts at the **rivers** edge and children swimming in the river. *B.* As it was Sunday the river was lined with Thai families having picnics in the bamboo picnic huts at the **river's** edge and children swimming in the river.
(personal email, 13/10/03)

(A.197) *A.* Therefore, alternative evaluation methods suggest an application-related assessment, which may, e.g. for a Natural Language Interface to a database management system, give an indication how many correct queries are generated or how many the concepts/propositions contained in an utterance are picked up.*B.* Therefore, alternative evaluation methods suggest an application-related assessment, which may, e.g. for a Natural Language Interface to a database management system, give an indication **of** how many correct queries are generated or how many **of** the concepts/propositions contained in an utterance are picked up.
(Ballim *et al.* 1999, p.14)

(A.198) *A.* I just don't know **have** it will work. *B.* I just don't know **how** it will work.
(http://bigbrother.channel4.com/bigbrother/news/newsstory_t.jsp?id=5165, 16/7/03)

(A.199) *A.* I do think I benefited from the utter lack of decent development **environment** as compared to now. *B.* I do think I benefited from the utter lack of **a** decent development environment as compared to now. *C.* I do think I benefited from the utter lack of decent development **environments** as compared to now.
(personal email, 11/12/03)

(A.200) *A.* Even if this statement dates back to 1996, during these last two years no real **improvements has** been done in achieving full robustness for a NLP system. *B.* Even if this statement dates back to 1996, during these last two years no real improvements **have** been done in achieving full robustness for a NLP system. *C.* Even if this statement dates back to 1996, during these last two years no real **improvement** has been done in achieving full robustness for a NLP system.
(Ballim *et al.* 1999, p.2)

(A.201) *A.* Her and dave seem to getting along really well. *B.* Her and dave seem to **be** getting along really well.
(personal email, May 2003)

(A.202) *A.* In many people's eyes Haughey represents the old Ireland where sharp practices were the norm, and while he **has** still **has** pockets of support, most people now accept that what he once considered acceptable is now beyond the pale in our more open and critical society. *B.* In many people's eyes Haughey represents the old Ireland where sharp practices were the norm, and while he still **has** pockets of support, most people now accept that what he once considered acceptable is now beyond the pale in our more open and critical society. *C.* In many people's eyes Haughey represents the old Ireland where sharp practices were the norm, and while he **has** still pockets of support, most people now accept that what he once considered acceptable is now beyond the pale in our more open and critical society.
(Evening Herald, p. 14,14/8/03)

(A.203) *A.* Where do sexual myths come? *B.* Where do sexual myths come **from**?
(Evening Herald, p.16, 11/10/03)

(A.204) *A.* A third journalist has been killed as a result of US attacks on two media bases in Baghdad today after a Spanish TV station confirmed that one of its cameramen **were** killed. *B.* A third journalist has been killed as a result of US attacks on two media bases in Baghdad today after a Spanish TV station confirmed that one of its cameramen **was** killed.
(http://media.guardian.co.uk/broadcast/story/0,7493,932363,00.html, April 2003)

(A.205) *A.* Are people really capable **to** understanding them? *B.* Are people really capable **of** understanding them?
(draft of own writing, 30th July 2003)

(A.206) *A.* From another perspective increasing knowledge about the context can be extracted by a dialogue management module in order to improve robustness at syntactic or semantic level.*B.* From another perspective increasing knowledge about the context can be extracted by a dialogue management module in order to improve robustness at **a** syntactic or semantic level. *C.* From another perspective increasing knowledge about the context can be extracted by a dialogue management module in order to improve robustness at **the** syntactic or semantic level.
(Ballim *et al.* 1999, p.13)

(A.207) **A.** Why **is do** they appear only in this one particular section of the beach and not elsewhere?*B.* Why **do** they appear only in this one particular section of the beach and not elsewhere?*C.* Why **is it** they appear only in this one particular section of the beach and not elsewhere?
(Derwing 1979, p.170)

(A.208) *A.* The group's game of charades this afternoon was **the** worth the entrance fee alone, if only to witness Jade being attacked by a large rubber whale. *B.* The group's game of charades this afternoon was worth the entrance fee alone, if only to witness Jade being attacked by a large rubber whale.
(http://www.channel4.com/apps3/bigbrother/news/newsstory_t.jsp?id=14303, July 2002)

(A.209) *A.* **are** sure why not *B.* **aah** sure why not
(personal email, 16/1/04)

(A.210) *A.* **Does** your circles overlap? *B.* **Do** your circles overlap?
(student assignment, 22/12/04)

(A.211) *A.* When an array exists each **elements** would be installed in index which indicates **as** a number. *B.* When an array exists each **element** would be installed in index which indicates a number. *C.* When an array exists **all** elements would be installed in index which indicates a number.
(student assignment, Feb 2003)

(A.212) *A.* Yoga brings peace and vitality to **you** life. *B.* Yoga brings peace and vitality to **your** life.
(sign in golf shop in Howth, 21/9/03)

(A.213) *A.* Points at entry reflect more than an **individuals** intellectual ability and in UCD we strive to address the initial disadvantage of students entering with low points. *B.* Points at entry reflect more than an **individual's** intellectual ability and in UCD we strive to address the initial disadvantage of students entering with low points.
(Evening Herald, Going to College supplement, p.2-3, 19/8/03)

(A.214) *A.* Reports **suggest** that networking equipment manufacturer 3Com is to close its manufacturing plant at Ballycoolin, Blanchardstown, with the loss of 600 jobs had still not been confirmed by the company by close of business Friday. *B.* Reports **suggesting** that networking equipment manufacturer 3Com is to close its manufacturing plant at Ballycoolin, Blanchardstown, with the loss of 600 jobs had still not been confirmed by the company by close of business Friday. *C.* Reports **that** suggest that networking equipment manufacturer 3Com is to close its manufacturing plant at Ballycoolin, Blanchardstown, with the loss of 600 jobs had still not been confirmed by the company by close of business Friday.
(IT's Monday News electronic newsletter, 8/9/03)

(A.215) *A.* Oh, of course I'd checked the cupboards on the way up the stairs and there was **on** one in any of them. *B.* Oh, of course I'd checked the cupboards on the way up the stairs and there was **no** one in any of them.
("Bridget Jones's Diary" by Helen Fielding, p.174)

(A.216) *A.* I was just chatting Mary, Anna and Laura, and thinking about summer pints at the Pav. *B.* I was just chatting **to** Mary, Anna and Laura, and thinking about summer pints at the Pav.
(personal email, 7/7/03)

(A.217) *A.* Desperate see the outcome of the weekly task? *B.* Desperate **to** see the outcome of the weekly task?
(http://www.channel4.co.uk/bigbrother/news/newsstory_t.jsp?id=1279, 28/5/03)

(A.218) *A.* In one segment, the former 'NSync star got the audience to shout 'Happy New Year' for a television show **to the** broadcast in his home state of Tennessee next week. *B.* In one segment, the former 'NSync star got the audience to shout 'Happy New Year' for a television show **to be** broadcast in his home state of Tennessee next week. *C.* In one segment, the former 'NSync star got the audience to shout 'Happy New Year' for a television show broadcast in his home state of Tennessee next week.
(Irish Independent, p.13, 22/12/03)

(A.219) *A.* But the trend is definitely upwards, and the hope is we might **we might** see a return to the great copywriting standard of Hollywood's past. *B.* But the trend is definitely upwards, and the hope is we might see a return to the great copywriting standard of Hollywood's past.
(Evening Herald, 10/10/03, It's Friday supplement, p.4)

(A.220) *A.* It will continue to screen popular western movies **and** on Tuesdays but on top of this plans to air critically acclaimed films such as The Exorcist, Excalibur and All The President's Men. *B.* It will continue to screen popular western movies on Tuesdays but on top of this plans to air critically acclaimed films such as The Exorcist, Excalibur and All The President's Men.
(Evening Herald, p.3, 27/9/03)

(A.221) *A.* The Cotton Club is certainly a lavish spectacle with director Francis Ford Coppola recreating spirit of 1920s Harlem. *B.* The Cotton Club is certainly a lavish spectacle with director Francis Ford Coppola recreating **the** spirit of 1920s Harlem.
(Evening Herald, p.25, 23/8/03)

(A.222) *A.* I actually think I've got a better chance of winning Pop Idol than I had **to** winning You're A Star in Ireland. *B.* I actually think I've got a better chance of winning Pop Idol than I had **of** winning You're A Star in Ireland.
(Evening Herald, p.3, 25/9/03)

(A.223) *A.* If you are not truly seamless **that** you cannot manage properly. *B.* If you are not truly seamless **then** you cannot manage properly.
(IT's Monday News electronic newsletter, 8/12/03)

(A.224) *A.* In fact, Collins further decomposes the probability of **a** generating left and right **modifies**. *B.* In fact, Collins further decomposes the probability of generating left and right **modifiers**.(draft of paper on probabilistic parsing by Deirdre Hogan, May 2003)

(A.225) *A.* **Their** is no bad feeling which is good. *B.* **There** is no bad feeling which is good.
(Turn newsletter, 6/2/03)

(A.226) *A.* This was **was** reflected in the degree results over 70pc of all honours graduates obtain First or Second Class Honours degrees, placing UCD ninth in the rating of all 110 UK Universities behind Oxford and Cambridge, but ahead of Imperial College, Glasgow and University College London, said a UCD spokesperson. *B.* This was reflected in the degree results**.** Over 70pc of all honours graduates obtain First or Second Class Honours degrees, placing UCD ninth in the rating of all 110 UK Universities behind Oxford and Cambridge, but ahead of Imperial College, Glasgow and University College London, said a UCD spokesperson.
(Evening Herald, Going to College supplement, p.2, 19/8/03)

(A.227) *A.* The **skip** of normal rule symbols adversely increases the constituent space. *B.* The **skipping** of normal rule symbols adversely increases the constituent space.
(Wang 2001)

(A.228) *A.* Start tying up the **lose** ends in your life. *B.* Start tying up the **loose** ends in your life.
(Evening Herald, 14/10/03, p.28)

(A.229) *A.* This may be no more than saying an IT degree travels, but if it does then **its** worth saying. *B.* This may be no more than saying an IT degree travels, but if it does then **it's** worth saying.
(personal email, 8/12/03)

(A.230) *A.* Their potential must be exploited as **as** efficiently as possible, preferably by ensuring a broadcaster will guarantee them copious hours of free publicity with their own TV shows. *B.* Their potential must be exploited as efficiently as possible, preferably by ensuring a broadcaster will guarantee them copious hours of free publicity with their own TV shows.
(The Observer, 27/7/03, p.25)

(A.231) *A.* What components we need for this system? *B.* What components **do** we need for this system?
(15th Esslli, slides from foundational course on Probabilistic Models of NLP, 1/10/03)

(A.232) *A.* And if you can get me any more hands of known **of** unknown origin that would be a plus. *B.* And if you can get me any more hands of known **or** unknown origin that would be a plus.
(personal email, 19/1/04)

(A.233) *A.* We sometimes distinguish between initial suffixes, as those **who** may occur directly after a word stem, and subsequent suffixes, as those **who** may not. *B.* We sometimes distinguish between initial suffixes, as those **which** may occur directly after a word stem, and subsequent suffixes, as those **which** may not.
(User Manual for the JavaChart 0.1.3 system, p.10, downloaded May 2003)

(A.234) *A.* If one of the boys goes **of** into a corner with one of the girls, seeking privacy, it will annoy all the others, who are, in a sense, being rejected by the new 'couple'. *B.* If one of the boys goes **off** into a corner with one of the girls, seeking privacy, it will annoy all the others, who are, in a sense, being rejected by the new 'couple'.
(http://www.channel4.com/bigbrother/news/newsstory_t.jsp?id=2144, 11th June 2003)

(A.235) *A.* The more the annotation scheme allows **to compress** a large corpus of utterances from the language, the more informative this annotation scheme is. *B.* The more the annotation scheme allows a large corpus of utterances from the language **to be compressed**, the more informative this annotation scheme is. *C.* The more the annotation scheme allows **one** to compress a large corpus of utterances from the language, the more informative this annotation scheme is.
(Musillo and Sima'an 2002)

(A.236) *A.* As with the previous elections, **ensure** the involvement of the community, the new FoLLI president will be elected directly by the members. *B.* As with the previous elections, **to** ensure the involvement of the community, the new FoLLI president will be elected directly by the members. *C.* As with the previous elections, **ensuring** the involvement of the community, the new FoLLI president will be elected directly by the members.
(FoLLI electronic newsletter, 19/11/03)

(A.237) *A.* According to Isabelle Roux-Buisson, head of the EMEA Imaging and Printing Group for HP, the company has another USD 300 million of business in the pipeline as enterprise customers **looks** to cut costs in every part of their organisation. *B.* According to Isabelle Roux-Buisson, head of the EMEA Imaging and Printing Group for HP, the company has another USD 300 million of business in the pipeline as enterprise customers **look** to cut costs in every part of their organisation.
(IT's Monday News electronic newsletter, 13/10/03)

(A.238) *A.* In the interest of customer archives and to ensure the story does not get lost in email inboxes, I have decided to send the story out in next **weeks** issue. *B.* In the interest of customer archives and to ensure the story does not get lost in email inboxes, I have decided to send the story out in next **week's** issue.
(IT's Monday news electronic newsletter, 4/11/03)

(A.239) *A.* I'd hate to be stuck at home thinking I **waited** my youth, when I could be out at work. *B.* I'd hate to be stuck at home thinking I **wasted** my youth, when I could be out at work.
(Evening Herald, p.41, 26/9/03)

(A.240) *A.* Wednesday is **find** with me. *B.* Wednesday is **fine** with me.
(personal text message, 23/9/03)

(A.241) *A.* Many of those I have told my story can't believe I disliked the Mediterranean island. *B.* Many of those I have told my story **to** can't believe I disliked the Mediterranean island. *C.* Many of those **to whom** I have told my story can't believe I disliked the Mediterranean island.
(Evening Herald, p.27, 11/8/03)

(A.242) *A.* This section identifies three additional critical points one might of the system. *B.* This section identifies three additional critical points one might **have** of the system.
(draft of "Enabling Left Recursive DCGs" by Dr. Carl Vogel, March 2003)

(A.243) *A.* We can order **then** directly from the web. *B.* We can order **them** directly from the web.
(personal email, 15/5/03)

(A.244) *A.* The word replacement error rules are, however, a more direct reflection of the analysis carried out on this operation since there is an error rule connective for each of the categories into which **these kind** of error **corrections** can be divided. *B.* The word replacement error rules are, however, a more direct reflection of the analysis carried out on this operation since there is an error rule connective for each of the categories into which **this** kind of error **correction** can be divided. *C.* The word replacement error rules are, however, a more direct reflection of the analysis carried out on this operation since there is an error rule connective for each of the categories into which these **kinds** of error correction can be divided.
(draft of own writing, 30/10/03)

(A.245) *A.* I want to express myself desirious of **maintain** correspondence with all Welshmen interested in doing it. *B.* I want to express myself desirious of **maintaining** correspondence with all Welshmen interested in doing it.
a letter quoted in James (1998, p.268)

(A.246) *A.* All the principles together ensure that the information in the rules is passed on to the **the** desired location in the analysis. *B.* All the principles together ensure that the information in the rules is passed on to the desired location in the analysis.
(Fouvry 2003b, p.94)

(A.247) *A.* Propylon established a loosely coupled messaging hub as the core of the system**.** which enables the BizTalk servers in the two agencies to communicate. *B.* Propylon established a loosely coupled messaging hub as the core of the system which enables the BizTalk servers in the two agencies to communicate.
(IT's Monday News eletronic newsletter, 7th April 2003)

(A.248) *A.* Once a random number **has selected**, the index in the array **has** decided and the one verb **has** chosen. *B.* Once a random number **is selected**, the index in the array **is** decided and the one verb **is** chosen. *C.* Once a random number **has been selected**, the index in the array **is** decided and the one verb **is** chosen.
(student assignment, Feb 2003)

(A.249) *A.* Within parsing, the original need for corpora of naturalistic sentences arose because grammars **were generally assigned** more than one sentential structure to a sentence and information about sentence frequency was needed in order to choose the most likely structure. *B.* Within parsing, the original need for corpora of naturalistic sentences arose because grammars **generally assigned** more than one sentential structure to a sentence and information about sentence frequency was needed in order to choose the most likely structure. *C.* Within parsing, the original need for corpora of naturalistic sentences arose because grammars **were generally assigning** more than one sentential structure to a sentence and information about sentence frequency was needed in order to choose the most likely structure.
(draft of own writing, 2/12/03)

(A.250) *A.* We could guess that what is meant is the sentence "He didn't tell me" but our linguistic knowledge tells us that, unlike "hit" in the previous context, "didn't tell" **and "tell"** can't be omitted from the sentence in this context. *B.* We could guess that what is meant is the sentence "He didn't tell me" but our linguistic knowledge tells us that, unlike "hit" in the previous context, "didn't tell" can't be omitted from the sentence in this context.
(draft of own writing, 18/8/03)

(A.251) *A.* Any even thought of tensions **rise** fears of incest. *B.* Any even thought of tensions **raises** fears of incest.
(personal email, 14/9/02)

(A.252) *A.* **Chomskys** books are really really boring though, aren't they? *B.* **Chomsky's** books are really really boring though, aren't they?
(http://www.thumped.com/bbs/showthread.php?+=9918, 11/11/03)

(A.253) *A.* Due to mix up of some numbers and names, we ask that you keep your winning information very confidential till your **claims** has been processed and your prize/money remitted to you. *B.* Due to **the** mix up of some numbers and names, we ask that you keep your winning information very confidential till your **claim** has been processed and your prize/money remitted to you. *C.* Due to **a** mix up of some numbers and names, we ask that you keep your winning information

very confidential till your **claim** has been processed and your prize/money remitted to you.
(spam email from 16/4/03)

(A.254) *A.* Agnes Lyon, the elderly woman who may have **lay** dead in her home for up to a year, has been buried in a private ceremony by her family. *B.* Agnes Lyon, the elderly woman who may have **lain** dead in her home for up to a year, has been buried in a private ceremony by her family.
(Evening Herald, p. 6,14/8/03)

(A.255) *A.* Once **allow** the scientist to invent the facts he uses for hypothesis testing, **and** the door is wide open to self-fulfilling predictions. *B.* Once **we** allow the scientist to invent the facts he uses for hypothesis testing, **then** the door is wide open to self-fulfilling predictions. *C.* Once the scientist **is allowed** to invent the facts he uses for hypothesis testing, **then** the door is wide open to self-fulfilling predictions.
(Sampson 2001a, p.124)

(A.256) *A.* This would **no** in itself be objectionable if they did not pretend that those judgments were data in favor of the theory. *B.* This would **not** in itself be objectionable if they did not pretend that those judgments were data in favor of the theory.
(transcription error, original sentence in Snow and Meijer (1976))

(A.257) *A.* One of the most important arguments for declarative **approaches** to NLP is that **it** is monotonic. *B.* One of the most important arguments for declarative approaches to NLP is that **they** are monotonic. *C.* One of the most important arguments for **a** declarative **approach** to NLP is that it is monotonic.
(Fouvry 2003a, p.112)

(A.258) *A.* C and W didn't **seen** to realise that there are only a handful of companies who require a truly global telecoms provider. *B.* C and W didn't **seem** to realise that there are only a handful of companies who require a truly global telecoms provider.
(IT's Monday News electronic newsletter, 9/6/03)

(A.259) *A.* Of course, since these rules rely on **the use** the CAT feature, their effect could be approximated in the old format by simply enumerating all the various instantiations of the rule. *B.* Of course, since these rules rely on the use **of** the CAT feature, their effect could be approximated in the old format by simply enumerating all the various instantiations of the rule. *C.* Of course, since these rules rely on **using** the CAT feature, their effect could be approximated in the old format by simply enumerating all the various instantiations of the rule.
(Jurafsky and Martin 2000, p.436)

(A.260) *A.* Any body category of a grammar does not occur in the head of some rule, is interpreted as a word category. *B.* Any body category of a grammar **that** does not occur in the head of some rule, is interpreted as a word category.
(User Manual for the JavaChart 0.1.3 system, p.12, downloaded May 2003)

(A.261) *A.* But it suffers **of** the problem **to be** biased and constrained by canonical approaches to NLP.*B.* But it suffers **from** the problem **of being** biased and constrained by canonical approaches to NLP.
(Ballim *et al.* 1999, p.2)

(A.262) *A.* When equipment has come to the end of **it's** life Niamh should also be informed before disposing of it. B. When equipment has come to the end of **its** life Niamh should also be informed before disposing of it.
(personal email, 15/10/03)

(A.263) *A.* Last night Tania, Nush and her were in bed when she spoke of **how** trying to make peace with Jon **and** "he comes back with a negative comment." *B.* Last night Tania, Nush and her were in bed when she spoke of trying to make peace with Jon and "he comes back with a negative comment." *C.* Last night Tania, Nush and her were in bed when she spoke of how trying to make peace with Jon "he comes back with a negative comment."
(http://www.channel4.com/bigbrother/news/newsstory_t.jsp?id=1733, 9/6/03)

(A.264) *A.* Modern technology had enabled some hugely driven clubs to make approaches to budding underage players – the mobile phone text and **it** not against the FAI rules. *B.* Modern technology had enabled some hugely driven clubs to make approaches to budding underage players – the mobile phone text and **it's** not against the FAI rules. *C.* Modern technology had enabled some hugely driven clubs to make approaches to budding underage players – the mobile phone text and it **is** not against the FAI rules.
(Evening Herald, p.62, 29/9/03)

(A.265) *A.* For example, if there is some sort of verb spanning position x through position y in the input and some **some** sort of noun phrase spanning position y through position z the skeletal chart

would record that there is a sequence of type v_np ending at point z. *B.* For example, if there is some sort of verb spanning position x through position y in the input and some sort of noun phrase spanning position y through position z the skeletal chart would record that there is a sequence of type v_np ending at point z.
(Moore and Dowding 1991)

(A.266) *A.* Breen has been an ever-present for Ireland since Brian Kerr took at the start of this year, but he'll have a new defensive partner for Saturday as Kenny Cunningham is suspended. *B.* Breen has been an ever-present **force** for Ireland since Brian Kerr took **over** at the start of this year, but he'll have a new defensive partner for Saturday as Kenny Cunningham is suspended.
(Evening Herald, p.72, 8/10/03)

(A.267) *A.* Before you reach for your email (or, if **your** work for a record label, your lawyer) and accuse me of condoning piracy, please read all the way to the end of the page. *B.* Before you reach for your email (or, if **you** work for a record label, your lawyer) and accuse me of condoning piracy, please read all the way to the end of the page.
(PC Advisor magazine, p.14, Issue 98, October 2003)

(A.268) *A.* Could you tell me what this spec **this** since I was hoping if possible to buy my laptop at the same time as these laptops. *B.* Could you tell me what this spec **is** since I was hoping if possible to buy my laptop at the same time as these laptops.
(personal email, 15/9/04)

(A.269) *A.* And afterwards I felt a bit bad as he saw my facial expressions, but I guess **knew he** beforehand that I wouldn't agree with a word of **his it**. *B.* And afterwards I felt a bit bad as he saw my facial expressions, but I guess **he knew** beforehand that I wouldn't agree with a word of **his**. *C.* And afterwards I felt a bit bad as he saw my facial expressions, but I guess **he knew** beforehand that I wouldn't agree with a word of **it**.
(personal email, 14/9/02)

(A.270) *A.* But this is not to say that Malta is without **it's** charms. *B.* But this is not to say that Malta is without **its** charms.
(Evening Herald, p.27, 11/8/03)

(A.271) *A.* She and Arkie been together for seven years and, although they're still not married, he did propose to her in January 2002. B. She and Arkie **have** been together for seven years and, although they're still not married, he did propose to her in January 2002.
(Now magazine, p.7, 22/10/03)

(A.272) *A.* It is **walker** who **has** the biggest grievance and the superior right. *B.* It is **the** walker who has the biggest grievance and the superior right. *C.* It is **walkers** who **have** the biggest grievance and the superior right.
(Evening Herald, p.15, 26/9/03)

(A.273) *A.* But they also **was** voiced their real concerns for their safety in the future without a fire escape. *B.* But they also voiced their real concerns for their safety in the future without a fire escape. *C.* But also their real concerns **were** voiced for their safety in the future without a fire escape.
(Evening Herald, p.6, 15/8/03)

(A.274) *A.* Minds are like **umbrella's**, they only work when they are open. *B.* Minds are like **umbrellas**, they only work when they are open.
(Evening Herald, p.26, 27/9/03)

(A.275) *A.* Meanwhile back in the real world it was left to St.Pat's and Bohs restore your faith in what all of this is supposed to be about. *B.* Meanwhile back in the real world it was left to St.Pat's and Bohs **to** restore your faith in what all of this is supposed to be about.
(Evening Herald, p.86, 9/10/03)

(A.276) *A.* In response to the vast array of security threats caused by holes in Windows technology, Microsoft boss Steve Ballmer has outlined a series of initiatives **aimed** improving the company's response to future threats. *B.* In response to the vast array of security threats caused by holes in Windows technology, Microsoft boss Steve Ballmer has outlined a series of initiatives aimed **at** improving the company's response to future threats. *C.* In response to the vast array of security threats caused by holes in Windows technology, Microsoft boss Steve Ballmer has outlined a series of initiatives **with the aim of** improving the company's response to future threats.
(IT's Monday News electronic newsletter, 13/10/03)

(A.277) *A.* But the present system is fairer now **then** the one that preceded it. *B.* But the present system is fairer now **than** the one that preceded it.
(Evening Herald, p. 14, 18/8/03)

(A.278) *A.* Promises to better than the last last-ever episode. *B.* Promises to **be** better than the last last-ever episode.
(The Irish Times, Ticket Supplement, p.9, 24/12/03)

(A.279) *A.* ColdFusion MX has been certified as Java Verified through Sun's Java Verification Programme, which means sun **in** satisfied that it is portable across different implementations of J2EE. *B.* ColdFusion MX has been certified as Java Verified through Sun's Java Verification Programme, which means sun **is** satisfied that it is portable across different implementations of J2EE.
(IT's Monday electronic newsletter, 30th June 2003)

(A.280) *A.* This class also contains a static method called IndexOfVerb **made** use of the Random method. *B.* This class also contains a static method called IndexOfVerb **which makes** use of the Random method.
(student assignment, Feb 2003)

(A.281) *A.* A new object **has** created. *B.* A new object **was** created.
(student assignment, Feb 2003)

(A.282) *A.* For instance, while Plumtree customers can share portlets with BEA **customer**, there is no way to do that unless they have a forum for exchanging. *B.* For instance, while Plumtree customers can share portlets with BEA **customers**, there is no way to do that unless they have a forum for exchanging.
(IT's Monday News electronic newsletter, 17/11/03)

(A.283) *A.* The Irish singer who got through to the final ten in Pop Idol UK last weekend flew home to be his daughter Chloe for her birthday this weekend. *B.* The Irish singer who got through to the final ten in Pop Idol UK last weekend flew home to be **with** his daughter Chloe for her birthday this weekend. *C.* The Irish singer who got through to the final ten in Pop Idol UK last weekend flew home to be **there** for his daughter Chloe for her birthday this weekend.
(Evening Herald, p.3, 25/9/03)

(A.284) *A.* The problem, however, in devising head word recognition procedures for an existing tree-bank, has been exemplified by the various versions of the head recognition procedure developed **by** for the WSJ tree-bank. *B.* The problem, however, in devising head word recognition procedures for an existing tree-bank, has been exemplified by the various versions of the head recognition procedure developed for the WSJ tree-bank.
(Musillo and Sima'an 2002)

(A.285) *A.* If it was initialised the value last element of the array will be lost. *B.* If it was initialised the value **of the** last element of the array will be lost.
(lecture notes, Jan 2003)

(A.286) *A.* The Cardiff midfielder has been red hot form for his club since the season began and has looked assured and confident in training over the last four days. *B.* The Cardiff midfielder has been **in** red hot form for his club since the season began and has looked assured and confident in training over the last four days.
(Evening Herald, p.76, 10/10/03)

(A.287) *A.* Two hours and ten minutes is a long time for any movie to try and engage our attention but **when** one as doggedly uninteresting as this it seems like an aeon. *B.* Two hours and ten minutes is a long time for any movie to try and engage our attention but when one **is** as doggedly uninteresting as this it seems like an aeon. *C.* Two hours and ten minutes is a long time for any movie to try and engage our attention but **with** one as doggedly uninteresting as this it seems like an aeon.
(Evening Herald, p.22, 21/8/03)

(A.288) *A.* Four different methods of parsing are integrated, namely, HSPG parser, probabilistic CFG LR-parser, chunk parser and a fall-back HMM-based dialogue act recognizer. *B.* Four different methods of parsing are integrated, namely, **a** HSPG parser, **a** probabilistic CFG LR-parser, **a** chunk parser and **a** fall-back HMM-based dialogue act recognizer.
(Ballim *et al.* 1999, p.11)

(A.289) *A.* Intuitive principles as the Backoff principle *B.* Intuitive principles **such** as the Backoff principle
(Sima'an 2003, p.8)

(A.290) *A.* The price of bottled water in Irish shops can vary as much as **much as** 150pc. *B.* The price of bottled water in Irish shops can vary as much as 150pc.
(Irish Independent, p.7, 10/7/03)

(A.291) *A.* a forum for presentation, discussion and case collections of the use of LRs in **in** Linguistics and Computational Linguistics *B.* a forum for presentation, discussion and case collections of the use of LRs in Linguistics and Computational Linguistics
(LREC electronic newsletter, 5/1/04)

(A.292) *A.* The toString method is also used to print out **in** the bracketed structure. *B.* The toString method is also used to print out the bracketed structure.
(student assignment, 2/4/03)

(A.293) *A.* The rest of the talks are taken in turn by postgrads **give** an update on their research progress. *B.* The rest of the talks are taken in turn by postgrads **giving** an update on their research progress. *C.* The rest of the talks are taken in turn by postgrads **who** give an update on their research progress.
(personal email, 23/10/03)

(A.294) *A.* What the attributes of objects of this type? *B.* What **are** the attributes of objects of this type?
(lecture notes, 25/11/03)

(A.295) *A.* HCS Solutions has moved into the Calyx headquarters at the North Ring Business Park in Santry, Dublin and HCS chief executive, Bernard Hamill will take up senior role in business development in Calyx. *B.* HCS Solutions has moved into the Calyx headquarters at the North Ring Business Park in Santry, Dublin and HCS chief executive, Bernard Hamill will take up **a** senior role in business development in Calyx.
(IT's Monday News electronic newsletter, 20/10/03)

(A.296) *A.* Everybody who's anybody will be at **there** gig in Whelans on the 13th August. B. Everybody who's anybody will be at **their** gig in Whelans on the 13th August.
(Turn electronic newsletter, 7/8/03)

(A.297) *A.* Those **is** the know say she has a voice that could melt mountains. *B.* Those **in** the know say she has a voice that could melt mountains.
(Irish Times Saturday magazine, p.3, 26/4/03)

(A.298) *A.* Critics who hate the way British football is now so dominated by money will **site** the fact that the biggest and most commercially minded clubs can buy a large chunk of the success with their big fat cheque books. *B.* Critics who hate the way British football is now so dominated by money will **cite** the fact that the biggest and most commercially minded clubs can buy a large chunk of the success with their big fat cheque books.
(Evening Herald, p. 14, 18/8/03)

(A.299) *A.* If you feel your machine and/or files are at risk, **so** a call can be logged. *B.* If you feel your machine and/or files are at risk, a call can be logged.
(http://www.cs.tcd.ie/userinfo/faq/virushelp.html, 20/11/03)

(A.300) *A.* The semantic classes define the conceptual structures of the application that are independent **to** linguistic structures. *B.* The semantic classes define the conceptual structures of the application that are independent **of** linguistic structures.
(Wang 2001)

(A.301) *A.* We **where** watching TV one night in the TV room which was basically a large sheltered balcony when 4/5 possums decided to join us. *B.* We **were** watching TV one night in the TV room which was basically a large sheltered balcony when 4/5 possums decided to join us.
(personal email, 18/5/03)

(A.302) *A.* O'Driscoll then showed **this** class, when picking up a loose ball at halfway and dancing through the Romanian defence before being dragged down yards from the line. *B.* O'Driscoll then showed **his** class, when picking up a loose ball at halfway and dancing through the Romanian defence before being dragged down yards from the line.
(Evening Herald, p.64, 11/10/03)

(A.303) *A.* People's brains **have** can process 15 and 50 operations per second, and can hold about seven things in their mind at one time. *B.* People's brains can process **between** 15 and 50 operations per second, and can hold about seven things in their mind at one time.
(personal email, 5/11/03)

(A.304) *A.* After spending about 20 mins pumping them up we set **of** down the river. *B.* After spending about 20 mins pumping them up we set **off** down the river.
(personal email, 19/4/03)

(A.305) *A.* We extend the definition of feature attributes to include references to any **to** occurrences of literals, not just categories. *B.* We extend the definition of feature attributes to include references to any occurrences of literals, not just categories.
(User Manual for the JavaChart 0.1.3 system, p.9, downloaded May 2003)

(A.306) *A.* Implemented systems sometimes **concern** with more than one of these issues, but in the very average of cases robustness is considered as an add-on to existing systems. *B.* Implemented systems sometimes **are concerned** with more than one of these issues, but in the very average of cases robustness is considered an add-on to existing systems. *C.* Implemented systems sometimes

concern **themselves** with more than one of these issues, but in the very average of cases robustness is considered an add-on to existing systems.
(Ballim *et al.* 1999, p.14)

(A.307) *A.* Language is not **an** barrier because most overseas engineers speak good English. *B.* Language is not **a** barrier because most overseas engineers speak good English.
(IT's Monday News electronic newsletter, 9th June)

(A.308) *A.* Entertainer Pat Shortt will be seen his own half-hour production, based on his popular rural characters, in a prime-time slot after Christmas week. *B.* Entertainer Pat Shortt will be seen **in** his own half-hour production, based on his popular rural characters, in a prime-time slot after Christmas week.
(Evening Herald, p.11, 2/12/03)

(A.309) *A.* That's right, Michael, you concentrate on the 'big' issues and let the pederasts use this country as a bolt-hole **in where** they can run and hide. B. That's right, Michael, you concentrate on the 'big' issues and let the pederasts use this country as a bolt-hole **where** they can run and hide. C. That's right, Michael, you concentrate on the 'big' issues and let the pederasts use this country as a bolt-hole in **which** they can run and hide.
(Evening Herald, p.21, 23/8/03)

(A.310) *A.* It is perplexing how little sign of a conflict of loyalties Mary Robinson was feeling when expressing bitter disappointment at **they** way she felt underdeveloped countries were being let down yet again by practically everybody. *B.* It is perplexing how little sign of a conflict of loyalties Mary Robinson was feeling when expressing bitter disappointment at **the** way she felt underdeveloped countries were being let down yet again by practically everybody.
(Evening Herald, p.15, 25/9/03)

(A.311) *A.* Tania **tied** in vain to explain what she'd meant. *B.* Tania **tried** in vain to explain what she'd meant.
(http://www.channel4.co.uk/bigbrother/news/newsstory_t.jsp?id=1240)

(A.312) *A.* Many of the 60,000 students who got their Leaving Cert results will have **woke** up this morning with more than just a hangover. *B.* Many of the 60,000 students who got their Leaving Cert results will have **woken** up this morning with more than just a hangover.
(Evening Herald, p.12, 14/8/03)

(A.313) *A.* It is so wierd, they are all so aware of what is going on in **each of others** lives. *B.* It is so wierd, they are all so aware of what is going on in **each other's** lives. *C.* It is so wierd, they are all so aware of what is going on in **each of the other's** lives.
(personal email, 17/9/02)

(A.314) *A.* The whole day was a bit of a tourist trap to be honest you couldn't move without being followed round the shops with people trying to sell you stuff. *B.* The whole day was a bit of a tourist trap to be honest**.** You couldn't move without being followed round the shops with people trying to sell you stuff.
(personal email, 6/10/04)

(A.315) *A.* Sentences with yes-no-question structure are often (though not always) used to ask questions (hence the name), and begin with **a** auxiliary verb. *B.* Sentences with yes-no-question structure are often (though not always) used to ask questions (hence the name), and begin with **an** auxiliary verb.
(Jurafsky and Martin 2000, p.333)

(A.316) *A.* The evidence on which the type of grammar is based is the corpus of sentences from which the grammar is derived. *B.* The evidence on which the **second** type of grammar is based is the corpus of sentences from which the grammar is derived.
(draft of my own writing, 12/8/03)

(A.317) *A.* It working fine now. *B.* It is working fine now. *C.* It works fine now.
(personal email, 20/12/02)

(A.318) *A.* Of course, dealing this kind of rule requires some changes to our parsing scheme. *B.* Of course, dealing **with** this kind of rule requires some changes to our parsing scheme.
(Jurafsky and Martin 2000, p.436)

(A.319) *A.* Psycholinguistic theories are based on an idealized concept of language performance and/or competence, even when statistics methods are introduced to explain phenomena which are hardly understood **my** means of a formal theory.*B.* Psycholinguistic theories are based on an idealized concept of language performance and/or competence, even when statistics methods are introduced to explain phenomena which are hardly understood **by** means of a formal theory.
(Ballim *et al.* 1999, p.2)

(A.320) *A.* It was really the **prefect** end to our holiday. *B.* It was really the **perfect** end to our holiday.
(personal email, 19/4/03)

(A.321) *A.* Finite-state transducers should appeal to the linguist looking for precise and natural **description** of complex syntactic structures. *B.* Finite-state transducers should appeal to the linguist looking for precise and natural **descriptions** of complex syntactic structures. *C.* Finite-state transducers should appeal to the linguist looking for **a** precise and natural description of complex syntactic structures.
(Roche and Schabes 1997, p.241)

(A.322) *A.* There is still a cost to the customer in terms of their **employees** time. *B.* There is still a cost to the customer in terms of their **employees'** time.
(IT's Monday News electronic newsletter, 1/9/03)

(A.323) *A.* And for now the FA are playing down the **players** action. *B.* And for now the FA are playing down the **players'** action.
(Evening Herald, p.71, 8/10/03)

(A.324) *A.* In this case two neighbours separately applied for the exemption on a joint development which will **the** provide 10 new apartments to the rear of the site as well as five apartments in their existing two semi-detached houses. *B.* In this case two neighbours separately applied for the exemption on a joint development which will provide 10 new apartments to the rear of the site as well as five apartments in their existing two semi-detached houses.
(Irish Independent Property, p.2., 11/7/03)

(A.325) *A.* But many companies are now adopting techniques such collaborative computing, which do not need to as tightly integrated so they save a lot of money by buying a volume server. *B.* But many companies are now adopting techniques such **as** collaborative computing, which do not need to **be** as tightly integrated so they save a lot of money by buying a volume server.
(IT's Monday News electronic newsletter, 1/12/03)

(A.326) *A.* Hard-hitting TV mob boss Tony Soprano has teamed with **with** loveable rogue Del Boy in the latest version of the Oxford English Dictionary. *B.* Hard-hitting TV mob boss Tony Soprano has teamed **up** with loveable rogue Del Boy in the latest version of the Oxford English Dictionary. *C.* Hard-hitting TV mob boss Tony Soprano has teamed with loveable rogue Del Boy in the latest version of the Oxford English Dictionary.
(Evening Herald, p.7, 21/8/03)

(A.327) *A.* Eve O'Brien is **a** 26 **and** from Dublin, where she is assistant manager of a deli. *B.* Eve O'Brien is a 26 **year old** from Dublin, where she is assistant manager of a deli. *C.* Eve O'Brien is 26 and from Dublin, where she is assistant manager of a deli.
(Evening Herald, p.17, 23/8/03)

(A.328) *A.* The body of this for loop checks if value i.e. the number of a particular pronoun is equal to the number of **the number of** a verb in the verb array at position i. *B.* The body of this for loop checks if value i.e. the number of a particular pronoun is equal to the number of a verb in the verb array at position i.
(student assignment, Feb 2003)

(A.329) *A.* The studio audience is asked to decide whether **is** is truly possible to reach out to the other side. *B.* The studio audience is asked to decide whether **it** is truly possible to reach out to the other side.
(Evening Herald, p.31, 23/8/03)

(A.330) *A.* On-going **dialogues** between the user and the simulated computer system **is** recorded.*B.* On-going dialogues between the user and the simulated computer system **are** recorded. *C.* On-going **dialogue** between the user and the simulated computer system is recorded.
(Ballim *et al.* 1999, p.5)

(A.331) *A.* And issues as the domain of language use and context tend to clearly influence the speed of processing. *B.* And issues **such** as the domain of language use and context tend to clearly influence the speed of processing.
(Sima'an 2003, p.5)

(A.332) *A.* You can get these us by mailing help@cs.tcd.ie requesting security cd. *B.* You can get these **from** us by mailing help@cs.tcd.ie requesting security cd.
(personal email, 26/11/03)

(A.333) *A.* If you want to see what happens, try sending a message to yourself with **the a** Subject: line set to "Your financial future". *B.* If you want to see what happens, try sending a message to yourself with **the** Subject: line set to "Your financial future". *C.* If you want to see what happens, try sending a message to yourself with **a** Subject: line set to "Your financial future".
(personal email, 5/11/03)

(A.334) *A.* The change can be made at different **granularity**, from the entire grammar to every single rule. *B.* The change can be made at different **granularities**, from the entire grammar to every single rule. *C.* The change can be made at **a** different granularity, from the entire grammar to every single rule.
(Wang 2001)

(A.335) *A.* We are especially interested theoretical work in the framework of HPSG and in developing a large scale NLP system for German. *B.* We are especially interested **in** theoretical work in the framework of HPSG and in developing a large scale NLP system for German.
(electronic newsletter, 17/4/03)

(A.336) *A.* returns all edges of the chart that **ends** at a certain position *B.* returns all edges of the chart that **end** at a certain position
(User Manual for the JavaChart 0.1.3 system, p.4, downloaded May 2003)

(A.337) *A.* it can be **search** for sentences which have an entry for error type but no entry for number of errors per sentence. *B.* it can be **searched** for sentences which have an entry for error type but no entry for number of errors per sentence.
(Becker *et al.* 1999)

(A.338) *A.* It is nonsense - a fact which anyone who has seen film of the original incident, and the two which have put him into disciplinary trouble again, the gratuitous fly kick at Manchester United's Eric Djemba-Djemba and the bringing down of Everton's Thomas Gravesen **-** is bound to confirm. *B.* It is nonsense - a fact which anyone who has seen film of the original incident, and the two which have put him into disciplinary trouble again, the gratuitous fly kick at Manchester United's Eric Djemba-Djemba and the bringing down of Everton's Thomas Gravesen is bound to confirm.(Evening Herald, p.62, 23/8/03)

(A.339) *A.* The error typology took as its basis traditional grammar books and **an** collection of cases considered grammatically difficult.*B.* The error typology took as its basis traditional grammar books and **a** collection of cases considered grammatically difficult.
(Becker *et al.* 1999)

(A.340) *A.* Makes the interesting point that if you compare his chunking judgements with those of his parser and then his judgements with another human's **the** there is very little difference between the two differences. *B.* Makes the interesting point that if you compare his chunking judgements with those of his parser and then his judgements with another human's there is very little difference between the two differences.
(draft of own writing, Jan 2004)

(A.341) *A.* He's **is** Albufeira. *B.* He's **in** Albufeira.
("Bridget Jones's Diary" by Helen Fielding, p.282)

(A.342) *A.* Since linguists were the first in using computer technology for automatic language processing, it becomes clear the motivations **of** this choice.*B.* Since linguists were the first in using computer technology for automatic language processing, it becomes clear the motivations **for** this choice.
(Ballim *et al.* 1999, p.11)

(A.343) *A.* Put that in your pipe **an** smoke it. *B.* Put that in your pipe **and** smoke it.
(Evening Herald, 21/11/03, p.15)

(A.344) *A.* Mobile phone users have been warned to treat any unsolicited text messages telling them **them** they have won a prize with caution. *B.* Mobile phone users have been warned to treat any unsolicited text messages telling them they have won a prize with caution.
(Evening Herald, p.18, 27/11/03)

(A.345) *A.* Germs simply abound in populations Let's just say they are in the zillions. *B.* Germs simply abound in populations**.** Let's just say they are in the zillions.
(Irish Independent, p.14, 23/4/03)

(A.346) *A.* Let **of** know what time you are calling. *B.* Let **me** know what time you are calling.
(personal text message, 23/9/03)

(A.347) *A.* This is used to control the for loop and **creating** the Verb Array. *B.* This is used to control the for loop and **create** the Verb Array. *C.* This is used to control the for loop and **for creating** the Verb Array.
(student assignment, Feb 2003)

(A.348) *A.* The second defence seems to be **supported up** by an experiment carried out by Snow and Meijer which ellicited grammaticality judgements from a group of linguists and a group of non-linguists and which found greater inter-subject agreement and intra-subject consistency among the linguist group. *B.* The second defence seems to be **supported** by an experiment carried out by Snow and Meijer which ellicited grammaticality judgements from a group of linguists

and a group of non-linguists and which found greater inter-subject agreement and intra-subject consistency among the linguist group. *C.* The second defence seems to be **backed** up by an experiment carried out by Snow and Meijer which ellicited grammaticality judgements from a group of linguists and a group of non-linguists and which found greater inter-subject agreement and intra-subject consistency among the linguist group.
(draft of own writing, 26/8/03)

(A.349) *A.* That is, you can compile the files one by one or all at once, **a** you like. *B.* That is, you can compile the files one by one or all at once, **as** you like. *C.* That is, you can compile the files one by one or all at once, **if** you like.
(User Manual for the JavaChart 0.1.3 system, p.7, downloaded May 2003)

(A.350) *A.* The hope now is that we might see a return to the standard set in the golden age of taglines, the late Seventies**.** when Hollywood came up with some of the most famous phrases in popular culture. *B.* The hope now is that we might see a return to the standard set in the golden age of taglines, the late Seventies when Hollywood came up with some of the most famous phrases in popular culture.
(Evening Herald, 10/10/03, It's Friday supplement, p.4)

(A.351) *A.* **It's** purpose is to create a syntax tree. *B.* **Its** purpose is to create a syntax tree.
(student assignment, 2/4/03)

(A.352) *A.* noun phrases **contains** adjectives *B.* noun phrases **containing** adjectives *C.* noun phrases **that** contain adjectives
(lecture notes)

(A.353) *A.* In line with recent comments by Bruce Claflin, 3Com president and chief executive officer, that the company will provide resellers with higher margins **that** Cisco, O'Connor said the channel at the high end would be limited through certification. *B.* In line with recent comments by Bruce Claflin, 3Com president and chief executive officer, that the company will provide resellers with higher margins **than** Cisco, O'Connor said the channel at the high end would be limited through certification.
(IT's Monday News electronic newsletter, 14th July 2003)

(A.354) *A.* Anyhow I was really angry, she crossed one **to** many lines there. *B.* Anyhow I was really angry, she crossed one **too** many lines there.
(personal email, 10/10/02)

(A.355) *A.* Third, Prof. Nagao has contributed service to the field in ways **to** numerous to list. *B.* Third, Prof. Nagao has contributed service to the field in ways **too** numerous to list.
(ACL Newsletter email, 31st July 2003)

(A.356) *A.* To connect to the departmental network, it **need** to be running windows Xp Professional or Linux. *B.* To connect to the departmental network, it **needs** to be running windows Xp Professional or Linux.
(personal email, 26/11/03)

(A.357) *A.* **In** may also occur around errors, but relatively far less frequently. *A.* **It** may also occur around errors, but relatively far less frequently.
(Atwell and Elliott 1987)

(A.358) *A.* Take the largest chunk sentence-like material within a text string and consider it to be central. *B.* Take the largest chunk **of** sentence-like material within a text string and consider it to be central.
(Fouvry 2003b, p.45)

(A.359) *A.* It can be shown that my proposal about the cognitive "trade-off" **make** some further formal limitation on adjunct meanings beyond that which is implicit in the standard semantic interpretation of the CG category A.*B.* It can be shown that my proposal about the cognitive "trade-off" **makes** some further formal limitation on adjunct meanings beyond that which is implicit in the standard semantic interpretation of the CG category A.
(Dowty 2000, p.23)

(A.360) *A.* The aim of a natural language parser is take a sequence of words from a particular language and to group these words together in ways that reflect possible meanings of the word sequence. *B.* The aim of a natural language parser is **to** take a sequence of words from a particular language and to group these words together in ways that reflect possible meanings of the word sequence.
(draft of own writing, 11/8/03)

(A.361) *A.* For people travelling to the show, below **is** the names and numbers of the most reasonable hostels etc. *B.* For people travelling to the show, below **are** the names and numbers of the most reasonable hostels etc.
(Turn electronic newsletter, 7/8/03)

(A.362) *A.* Commuters looked on in horror **was** windows on the jeep were damaged during the incident. *B.* Commuters looked on in horror **as** windows on the jeep were damaged during the incident.
(Evening Herald, p. 16, 13/8/03)

(A.363) *A.* Expect the unexpected strange events could occur. *B.* Expect the unexpected**.** Strange events could occur.
(Evening Herald, p.26, 10/10/03)

(A.364) *A.* Is our youth really in **in** such a state of disrepair? *B.* Is our youth really in such a state of disrepair?
(The Observer, p.16, 7/9/03)

(A.365) *A.* Formally this is conditional probability of a given expansion given the left hand size non-terminal A. *B.* Formally this is **the** conditional probability of a given expansion given the left hand side non-terminal A.
(Jurafsky and Martin 2000, p.449)

(A.366) *A.* They would like **to** media to give them space. **B.** They would like **the** media to give them space.
(Sunday Independent, p.2., 9/2/03)

(A.367) *A.* Uncertainty with respect to how the analyses-space must be explored, and with respect to what to expect on basis of what have been seen, is yet another problem for the linguistic approach. *B.* Uncertainty with respect to how the analyses-space must be explored, and with respect to what to expect on **the** basis of what have been seen, is yet another problem for the linguistic approach.
(Sima'an 2003, p.5)

(A.368) *A.* But it depends on how rent you pay. *B.* But it depends on how **much** rent you pay.
(Irish Independent Business Section, p.8, 15/5/03)

(A.369) *A.* It's really really weird business at **time** and I think comedians often reflect that in their own personality. *B.* It's **a** really really weird business at **times** and I think comedians often reflect that in their own personality.
(Evening Herald, p.13, 27/9/03)

(A.370) *A.* Compared to **syntactic** valid structures, the set of syntactically incorrect sentences can be considered almost infinite. *B.* Compared to **syntactically** valid structures, the set of syntactically incorrect sentences can be considered almost infinite.
(Gojenola and Oronoz 2000)

(A.371) *A.* Attempts for this seem to have not gotten **to** far.*B.* Attempts for this seem to have not gotten **too** far.
(Ballim *et al.* 1999, p.5)

(A.372) *A.* He shot his victim once in the leg he stood on a local green before driving off. *B.* He shot his victim once in the leg **as** he stood on a local green before driving off.
(Evening Herald, p. 6,14/8/03)

(A.373) *A.* In fact they said they would coming and going from parliament all week. B. In fact they said they would **be** coming and going from parliament all week.
(Evening Herald, 29/10/03, p.4)

(A.374) *A.* Gallagher spent 16 **years'** working in the software and manufacturing industries where he was called upon to develop data collection, management and presentation software to enable companies to better manage their manufacturing processes. *B.* Gallagher spent 16 **years** working in the software and manufacturing industries where he was called upon to develop data collection, management and presentation software to enable companies to better manage their manufacturing processes.
(IT's Monday News electronic newsletter, 20/10/03)

(A.375) *A.* However several attempts **has** been carried out in order to approximate a robust behaviour.*B.* However several attempts **have** been carried out in order to approximate a robust behaviour.
(Ballim *et al.* 1999, p.2)

(A.376) *A.* This ordering allows a form of suppletion, whereby special forms such as those for the irregular plural of goose and key **to be** listed explicitly.*B.* This ordering allows a form of suppletion, whereby special forms such as those for the irregular plural of goose and key **are** listed explicitly. *C.* This ordering allows a form of suppletion, whereby special forms such as those for the irregular plural of goose and key **are to be** listed explicitly.
(ALE User Guide, Version 3.2.1. December 2001, p.48)

(A.377) *A.* And sheila has just **begin** mentioning angela again, so I take it she is back in her life. *B.* And sheila has just **begun** mentioning angela again, so I take it she is back in her life.
(personal email, 10/10/2002)

(A.378) *A.* How this kind of fault-tolerant behaviour **can** be reproduced in **a** artificial system by means of computational techniques? *B.* How **can** this kind of fault-tolerant behaviour be reproduced in **an** artificial system by means of computational techniques?
(Ballim *et al.* 1999, p.2)

(A.379) *A.* The 57-year-old former Sergeant Major, who faces three counts of kidnapping and murder, has been on hunger strike since July and is deemed **to** weak to appear in court in person. *B.* The 57-year-old former Sergeant Major, who faces three counts of kidnapping and murder, has been on hunger strike since July and is deemed **too** weak to appear in court in person.
(Evening Herald, p.4, 14/10/03)

(A.380) *A.* Chapter 6 describes the implementation that resulted **of** the experiment. *B.* Chapter 6 describes the implementation that resulted **from** the experiment.
(downloaded as 2001-001.pdf from stp.ling.uu.se/ matsd/thesis/arch/, 24th July 2003)

(A.381) *A.* More than a million credit cards have been introduced since 1999, and the figure now stands at two and a half credit cards, an astonishingly high figure for a country with our population. *B.* More than a million credit cards have been introduced since 1999, and the figure now stands at two and a half **million** credit cards, an astonishingly high figure for a country with our population.
(Evening Herald, 29/10/03, p.14)

(A.382) *A.* In what kind of written communication did the ungrammatical sentence occur **occur**? *B.* In what kind of written communication did the ungrammatical sentence occur?
(draft of own writing, 12/12/03)

(A.383) *A.* However since the corpora does **provided** insight into systematic linguistic deviance, it must be seen as being relevant to a more general theory of language. *B.* However since the corpora does **provide** insight into systematic linguistic deviance, it must be seen as being relevant to a more general theory of language.
(draft of own writing, 26/11/03)

(A.384) *A.* A queue is a list of elements with the property that elements can only be added to end of queue and can only be removed from the front. *B.* A queue is a list of elements with the property that elements can only be added to **the** end of **the** queue and can only be removed from the front. *C.* A queue is a list of elements with the property that elements can only be added to **the** end of **a** queue and can only be removed from the front.
(Lecture notes, April 2003)

(A.385) *A.* Last year we **done** java. *B.* Last year we **did** java.
(personal email, 2/10/03)

(A.386) *A.* Within 15 minutes of the evacuation, staff **was** allowed to return to the building. *B.* Within 15 minutes of the evacuation, staff **were** allowed to return to the building.
(http://www.nandotimes.com/nation/story/440233p-3522836c.html)

(A.387) *A.* **If** that **being** the case, O Se implied at his first press conference, the change should happen from the top down. *B.* If that **is** the case, O Se implied at his first press conference, the change should happen from the top down. *C.* That being the case, O Se implied at his first press conference, the change should happen from the top down.
(Evening Herald, p.68, 8/10/03)

(A.388) *A.* His suit by Irish design company Oakes, as were his **give** groomsmen and little brother Adam (13) who wore purple suits. *B.* His suit **was** by Irish design company Oakes, as were those of his **five** groomsmen and little brother Adam (13) who wore purple suits. *C.* His suit **was designed** by Irish design company Oakes, as were those of his **five** groomsmen and little brother Adam (13) who wore purple suits.
(Evening Herald, p.3, 11/8/03)

(A.389) *A.* Each **of text line** ends with a Newline character, which serves to separate it from the following line. *B.* Each **line of text** ends with a Newline character, which serves to separate it from the following line. C. Each **text line** ends with a Newline character, which serves to separate it from the following line.
(Emacs tutorial, 24/12/03)

(A.390) *A.* Madden has projected sales figures of 2 million for TechNiche's first year **and** business and claims this projection is 'conservative'. *B.* Madden has projected sales figures of 2 million for TechNiche's first year **of** business and claims this projection is 'conservative'.
(IT's Monday News electronic newsletter, 4/8/03)

(A.391) *A.* Enclosed please find updated sales list for **you** kind attention. *B.* Enclosed please find updated sales list for **your** kind attention.
(electronic newsletter from Jacob Walsh Estates Sales List, 15/8/03)

(A.392) *A.* Our computer kept having some error popup every time it rebooted but now **its** gone. *B.* Our computer kept having some error popup every time it rebooted but now **it's** gone.
(spam email about bug in doctor software for PCs, 31/12/03)

(A.393) *A.* But if you're in a car and the woman doesn't have central locking and she doesn't lean over to unlock **you** side, what does that tell you? *B.* But if you're in a car and the woman doesn't have central locking and she doesn't lean over to unlock **your** side, what does that tell you?
(http://www.channel4.co.uk/bigbrother/news/newsstory_t.jsp?id=1288, May 2003)

(A.394) *A.* A natural consequences *B.* A natural consequence
(lecture notes, Jan 2003)

(A.395) *A.* It is being headed up by Walter Dickson of Access Accounting who is the driving behind the alliance, which also includes ABT Systems, ADC Barcode, Coretime, East Pier and RGC Technologies. *B.* It is being headed up by Walter Dickson of Access Accounting who is the driving **force** behind the alliance, which also includes ABT Systems, ADC Barcode, Coretime, East Pier and RGC Technologies.
(IT's Monday News eletronic newsletter, 7th April 2003)

(A.396) *A.* This **makes** that only the generalisation of both values remains. B. This **means** that only the generalisation of both values remains.
(Fouvry 2003a, p.105)

(A.397) *A.* When 24/7 split, Brian then decided he **world** give Irish Popstars a go. *B.* When 24/7 split, Brian then decided he **would** give Irish Popstars a go.
(Evening Herald, p.5, 29/9/03)

(A.398) *A.* The proposed traffic lay-out has more in common with the Tallaght bypass than with **the any** current thinking on sustainable urban design. *B.* The proposed traffic lay-out has more in common with the Tallaght bypass than with **any** current thinking on sustainable urban design. *C.* The proposed traffic lay-out has more in common with the Tallaght bypass than with **the** current thinking on sustainable urban design.
(Evening Herald, p.18, 9/12/03)

(A.399) *A.* We might **found** out what was going on every six months, during which time researchers could have followed a path we didn't want them to take, wasting both **or** time and our money. *B.* We might **find** out what was going on every six months, during which time researchers could have followed a path we didn't want them to take, wasting both **our** time and our money. *C.* We might **have** found out what was going on every six months, during which time researchers could have followed a path we didn't want them to take, wasting both **our** time and our money.
(IT's Monday News electronic newsletter, 21st July 2001)

(A.400) *A.* McGuire likens this a Y2K event for POS systems and believes it will be a 'significant driver of sales growth in the medium term'. *B.* McGuire likens this **to** a Y2K event for POS systems and believes it will be a 'significant driver of sales growth in the medium term'.
(IT's Monday News electronic newsletter, 21/4/03)

(A.401) *A.* No I don't think Jon spoilt **anyone** chances of winning. *B.* No I don't think Jon spoilt **anyone's** chances of winning.
(http://www.channel4.com/community/showcards/B/Big_Brother_4_-_Scott_and_Steph.html, 30/7/03)

(A.402) *A.* Judge ni Chonduin imposed a four-month suspended sentence for the assault on the elderly man **with** took all other charges into consideration. *B.* Judge ni Chonduin imposed a four-month suspended sentence for the assault on the elderly man **which** took all other charges into consideration.
(Evening Herald, p.14, 11/10/03)

(A.403) *A.* I think we have decided just to go to the gig 'cos **its** too much hassle trying to sell the tickets, thanks for asking people though. *B.* I think we have decided just to go to the gig 'cos **it's** too much hassle trying to sell the tickets, thanks for asking people though.
(personal email, 2/11/02)

(A.404) *A.* Secondly, we gave an historical sketch of the recent developments of EA, up to and beyond its brief official heydey, suggesting that EA **as** never been abandoned, but has rather lain in the doldrums perhaps awaiting the signal to ply the main. *B.* Secondly, we gave an historical sketch of the recent developments of EA, up to and beyond its brief official heydey, suggesting that EA **has** never been abandoned, but has rather lain in the doldrums perhaps awaiting the signal to ply the main.
(James 1998, p.19)

(A.405) *A.* All of the entrants must not own a house, be married and have to be over 18 years old but the good news is if they win the house it can sold for hard cash. *B.* All of the entrants must not own a house, be married and have to be over 18 years old but the good news is if they win the house it can **be** sold for hard cash.
(Evening Herald, p.9, 11/10/03)

(A.406) *A.* On the basis of (i) these annotated examples and (ii) written documentation, new utterances should be possible to **analyse** with a high degree of consistency. *B.* On the basis of (i) these annotated examples and (ii) written documentation, **it should be possible** to analyse new utterances with a high degree of consistency. *C.* On the basis of (i) these annotated examples and (ii) written documentation, new utterances should be possible to **be analysed** with a high degree of consistency.
(Roche and Schabes 1997, p.290)

(A.407) *A.* Such databases can **be** to some extent **be** improved by reference to non-corpus sources such as the native speaker's knowledge of the language.*B.* Such databases can be to some extent improved by reference to non-corpus sources such as the native speaker's knowledge of the language.*C.* Such databases can to some extent be improved by reference to non-corpus sources such as the native speaker's knowledge of the language.
(Leech 1987)

(A.408) *A.* Go **to** Lisa's news archives. *B.* Go **through** Lisa's news archives.
(draft of own writing, 8/10/04)

(A.409) *A.* More **that** 4,600 people in Northern Ireland are employed in call and contact centres working for a wide range of organisations from computer companies and banks to the BBC. *B.* More **than** 4,600 people in Northern Ireland are employed in call and contact centres working for a wide range of organisations from computer companies and banks to the BBC.
(IT's Monday News electronic newsletter, 8/12/03)

(A.410) *A.* Can **you** let you know in a few days how i am getting on with it. *B.* Can let you know in a few days how i am getting on with it.
(personal email, 10/12/03)

(A.411) *A.* People seem to promise much more **then** they can deliver. *B.* People seem to promise much more **than** they can deliver.
(Evening Herald, p.30, 23/9/03)

(A.412) *A.* He said it doesn't any difference to what we do now because since 1996, we have had a policy of mandatory reporting of all abuse allegations, so the idea of anyone who speaks of this being excommunicated is nonsense, Mr. Bailey said. *B.* He said it doesn't **make** any difference to what we do now because since 1996, we have had a policy of mandatory reporting of all abuse allegations, so the idea of anyone who speaks of this being excommunicated is nonsense, Mr. Bailey said.
(Evening Herald, p. 8, 18/8/03)

(A.413) *A.* I forgot about Peter's stand in the Museum Building and he has asked if you could come at 9.30 and do a couple of **yours**. *B.* I forgot about Peter's stand in the Museum Building and he has asked if you could come at 9.30 and do a couple of **hours**.
(personal email, 8/12/03)

(A.414) *A.* The document describes game systems as **among of** the fastest growing segments of the consumer electronics market. *B.* The document describes game systems as **among** the fastest growing segments of the consumer electronics market. *C.* The document describes game systems as **one** of the fastest growing segments of the consumer electronics market.
(IT's Monday News electronic newsletter, 10/11/03)

(A.415) *A.* Taken on its own, this sentence is well-formed but having read and understood everything which preceded this sentence in the manual in which it appeared, it becomes apparent that is in fact ill-formed and that word build should be replaced by built. *B.* Taken on its own, this sentence is well-formed but having read and understood everything which preceded this sentence in the manual in which it appeared, it becomes apparent that **it** is in fact ill-formed and that **the** word build should be replaced by built.
(draft of own writing, 31/8/03)

(A.416) *A.* John Denver is a Post Doc working here in the Centre for Digital Video Processing and he's giving a talk soon on user interface design and evaluation which hopefully be useful for CALL students, or it can be from within the group like yesterday's seminar when we had a Latex tutorial from Robert Schneider and Jeff Mangum. *B.* John Denver is a Post Doc working here in the Centre for Digital Video Processing and he's giving a talk soon on user interface design and evaluation which **will** hopefully be useful for CALL students, or it can be from within the group

like yesterday's seminar when we had a Latex tutorial from Robert Schneider and Jeff Mangum. *C.* John Denver is a Post Doc working here in the Centre for Digital Video Processing and he's giving a talk soon on user interface design and evaluation which hopefully **will** be useful for CALL students, or it can be from within the group like yesterday's seminar when we had a Latex tutorial from Robert Schneider and Jeff Mangum.
(personal email, 23/10/03)

(A.417) *A.* Firstly, the aim was to develop **an** modular NLP software architecture for quickly developing different kinds of CL and grammar checking applications. *B.* Firstly, the aim was to develop **a** modular NLP software architecture for quickly developing different kinds of CL and grammar checking applications.
(Becker *et al.* 1999)

(A.418) *A.* Also a new array is created purpose of placing the tokens. *B.* Also a new array is created **for the** purpose of placing the tokens.
(student assignment, Feb 2003)

(A.419) *A.* He also expresses satisfaction at the **regulators** recent decision to impose a three-month period during which the incumbent cannot attempt to win back customers lost to other operators. *B.* He also expresses satisfaction at the **regulator's** recent decision to impose a three-month period during which the incumbent cannot attempt to win back customers lost to other operators.
(IT's Monday News electronic newsletter, 20/10/03)

(A.420) *A.* **Lose** floppy disks seem strikingly less secure than backedup filesystems. *B.* **Lost** floppy disks seem strikingly less secure than backedup filesystems.
(personal email, 8/4/03)

(A.421) *A.* Here's something that's being done to encourage young people to develop their creativity and **its** gone. *B.* Here's something that's being done to encourage young people to develop their creativity and **it's** gone.
(Evening Herald, p.40, 23/9/03)

(A.422) *A.* TC is a content-based management task and has much in common with **it's** neighbour Information Retrieval, borrowing and applying much of the basic IR methods. *B.* TC is a content-based management task and has much in common with **its** neighbour Information Retrieval, borrowing and applying much of the basic IR methods.
(DCLRS electronic newsletter, 21/10/03)

(A.423) *A.* We are now vastly different to the way we **there** then. *B.* We are now vastly different to the way we **were** then.
(Irish Times, 7th Feb 2003, Business This Week Supplement, p. 24)

(A.424) *A.* The jury of 10 men and two women agreed with the prosecution that Milroy-Sloan was a "cunning fantasist" who was out **for find** fame and fortune. *B.* The jury of 10 men and two women agreed with the prosecution that Milroy-Sloan was a "cunning fantasist" who was out **to find** fame and fortune. *C.* The jury of 10 men and two women agreed with the prosecution that Milroy-Sloan was a "cunning fantasist" who was out **for** fame and fortune.
(Evening Herald, p.39, 10/10/03)

(A.425) *A.* The bronzed Bermondsey belle has openly revealed she is a bottle blonde, and **roots** as the roots started to burst through, her barnet was in need of some urgent TLC. *B.* The bronzed Bermondsey belle has openly revealed she is a bottle blonde, and as the roots started to burst through, her barnet was in need of some urgent TLC.
(http://www.channel4.com/apps3/bigbrother/news/newsstory_t.jsp?id=12201, July 2002)

(A.426) *A.* After the controversy over the Gerry Ryan phone in, Dunne was suspended on full pay pending the outcome of a **a** civil service disciplinary inquiry. *B.* After the controversy over the Gerry Ryan phone in, Dunne was suspended on full pay pending the outcome of a civil service disciplinary inquiry.
(Evening Herald, p.6, 21/8/03)

(A.427) *A.* See you after **the my** teeth situation. *B.* See you after **my** teeth situation. *C.* See you after **the** teeth situation.(personal email, 23/6/03)

(A.428) *A.* If you have 30 minutes to spare for me and you don't mind helping me with **this experiments**, would you please let me know your available time for these experimental sessions? *B.* If you have 30 minutes to spare for me and you don't mind helping me with **these experiments**, would you please let me know your available time for these experimental sessions? *C.* If you have 30 minutes to spare for me and you don't mind helping me with **this experiment**, would you please let me know your available time for these experimental sessions?
(personal email, 14/1/04)

(A.429) *A.* Microsoft is facing an increasingly aggressive competitor in Oracle as it seeks to block its **rivals** expansion plans into the ERP market. *B.* Microsoft is facing an increasingly aggressive competitor in Oracle as it seeks to block its **rival's** expansion plans into the ERP market.
(IT's Monday News electronic newsletter, 1/12/03)

(A.430) *A.* The attribute and atomic value domains must have **an** well-defined identity operation. *B.* The attribute and atomic value domains must have **a** well-defined identity operation.
(User Manual for the JavaChart 0.1.3 system, p.3, downloaded May 2003)

(A.431) *A.* A Thematic Session provides a good occasion to focus **peoples** with the same special interest, and let them meet each other at a specific time-space to discuss and exchange ideas. *B.* A Thematic Session provides a good occasion to focus **people** with the same special interest, and let them meet each other at a specific time-space to discuss and exchange ideas.
(ACL electronic newsletter, 10/10/03)

(A.432) *A.* Around 1/3 of the slot errors are side **effect** of the card switching errors. *B.* Around 1/3 of the slot errors are **a** side effect of the card switching errors. *C.* Around 1/3 of the slot errors are side **effects** of the card switching errors.
(Wang 2001)

(A.433) *A.* Alex was then left alone muttering to himself in the kitchen as Kate and Jade lay in the garden in the early morning sun, as the IT commented, "I'm going to stay out here all day." *B.* Alex was then left alone muttering to himself in the kitchen as Kate and Jade lay in the garden in the early morning sun, as the IT **worker** commented, "I'm going to stay out here all day."
(http://www.channel4.com/apps3/bigbrother/news/newsstory_t.jsp?id=13158, July 2002)

(A.434) *A.* Hypothesize a new feature, test its success on actual data and then decide whether to include it in the model or not on basis of the results. *B.* Hypothesize a new feature, test its success on actual data and then decide whether to include it in the model or not on **the** basis of the results.
.(Sima'an 2003, p.14)

(A.435) *A.* The 807 errors which have been collected to date were added to the corpus over a period of **a** 15 months. *B.* The 807 errors which have been collected to date were added to the corpus over a period of 15 months.
(draft of own writing, 30/10/03)

(A.436) *A.* And the average income for **a players** in the top half of the Premiership now runs to over EUR50,000 weekly. *B.* And the average income for **a player** in the top half of the Premiership now runs to over EUR50,000 weekly. *C.* And the average income for **players** in the top half of the Premiership now runs to over EUR50,000 weekly.
(Evening Herald, p.37,12/11/03)

(A.437) *A.* The growth in volume servers points to growing trend in modular computing. *B.* The growth in volume servers points to **a** growing trend in modular computing.
(IT's Monday News electronic newsletter, 1/12/03)

(A.438) *A.* A dotted rule is **a** similar to an ordinary rule except that a dot marks out a position in its body and the body can contain a single token instead of categories. *B.* A dotted rule is similar to an ordinary rule except that a dot marks out a position in its body and the body can contain a single token instead of categories.
(User Manual for the JavaChart 0.1.3 system, p.4, downloaded May 2003)

(A.439) *A.* The first thing Xwerx discusses with a client is not the look of the site but **what** the goals the client wants the website to achieve. *B.* The first thing Xwerx discusses with a client is not the look of the site but the goals the client wants the website to achieve. *C.* The first thing Xwerx discusses with a client is not the look of the site but what **are** the goals the client wants the website to achieve.
(Irish Independent, 9/10/03, Business supplement, e-Thursday page)

(A.440) *A.* And don't fiddle about the page layout or the margin size. *B.* And don't fiddle about **with** the page layout or the margin size.
(http://www.sfb441.uni-tuebingen.de/ling.evidence/instr.html, 18/11/03)

(A.441) *A.* Just wondering **how** if the papers due in on Friday can be any longer than the submitted papers. *B.* Just wondering if the papers due in on Friday can be any longer than the submitted papers.
(personal email, 3/9/03)

(A.442) *A.* It's a black day **for** if you want retail therapy spent in relative peace and quiet as today's the traditional hardy annual excursion when everyone who possibly can from Down the Country descends on Dublin for their humongous Christmas shop. *B.* It's a black day for **you** if you want retail therapy spent in relative peace and quiet as today's the traditional hardy annual excursion

when everyone who possibly can from Down the Country descends on Dublin for their humongous Christmas shop. *C.* It's a black day if you want retail therapy spent in relative peace and quiet as today's the traditional hardy annual excursion when everyone who possibly can from Down the Country descends on Dublin for their humongous Christmas shop.
(Evening Herald, p.22, 8/12/03)

(A.443) *A.* There are other aspects of robustness which **even** have not been treated at all, including the flexible adaptation to external time constraints or internal resource limitations.*B.* There are other aspects of robustness which have not **even** been treated at all, including the flexible adaptation to external time constraints or internal resource limitations.
(Menzel 1995)

(A.444) *A.* Neosupervital is strictly **is** his baby though. *B.* Neosupervital is strictly his baby though.
(Foggy Notions magazine, Iss. 1, May 2003)

(A.445) *A.* To see when **when** training sessions for other subjects will take place, take a look at our website. B. To see when training sessions for other subjects will take place, take a look at our website.
(electronic newsletter, 11/11/03)

(A.446) *A.* **It's** purpose is to create an array of strings that contains the 8 pronouns. *B.* **Its** purpose is to create an array of strings that contains the 8 pronouns.
(student assignment, Feb 2003)

(A.447) *A.* According to Byrne the stated goal of bringing an international point of presence to the region is still being worked on but **that** the company hopes to announce a number of significant initiatives between now and Christmas. *B.* According to Byrne the stated goal of bringing an international point of presence to the region is still being worked on but the company hopes to announce a number of significant initiatives between now and Christmas.
(IT's Monday News electronic newsletter, 8/9/03)

(A.448) *A.* **They** will be case studies and demonstrations from 11:45 to 1:15 in LCR and SCR. *B.* **There** will be case studies and demonstrations from 11:45 to 1:15 in LCR and SCR.
(personal email, 12/5/03)

(A.449) *A.* The major technical innovation is the use of a "maximum-entropy-inspired" model for conditioning and smoothing that let us successfully **to** test and combine many different conditioning events. *B.* The major technical innovation is the use of a "maximum-entropy-inspired" model for conditioning and smoothing that let us successfully test and combine many different conditioning events.
(Charniak 2000).

(A.450) *A.* There were **there** hundreds of well-wishers ensconced in the airport arrivals lounge, while Ray was mobbed back home and virtually everywhere he set foot. *B.* There were hundreds of well-wishers ensconced in the airport arrivals lounge, while Ray was mobbed back home and virtually everywhere he set foot.
(http://bigbrother.channel4.com/bigbrother/news/newsstory_t.jsp?id=5380, 6/8/03)

(A.451) *A.* A trained group of annotators individually parse (i.e. assign meaningful structure) **to** the sentences in a corpus using a grammar as their reference. *B.* A trained group of annotators individually parse (i.e. assign meaningful structure **to**) the sentences in a corpus using a grammar as their reference.
(draft of own writing, 31/8/03)

(A.452) *A.* points the grammar file that shall be compiled *B.* points **to** the grammar file that shall be compiled *C.* points **out** the grammar file that shall be compiled
(User Manual for the JavaChart 0.1.3 system, p.7, downloaded May 2003)

(A.453) *A.* This year the House **if** full of attractive young males and females. *B.* This year the House **is** full of attractive young males and females.
(http://www.channel4.com/bigbrother/news/newsstory_t.jsp?id=2144, 10th June 2003)

(A.454) *A.* The Interactive Poster/Demo Sessions run in parallel **of** the main session. *B.* The Interactive Poster/Demo Sessions run in parallel **to** the main session.
(ACL electronic newsletter, 10/10/03)

(A.455) *A.* Just how much sex **in** there **is** a skeleton? *B.* Just how much sex **is** there **in** a skeleton?
("The Female Eunuch" by Germaine Greer, p.35)

(A.456) *A.* Conditions set for Heatwole's release included a ban on him entering any airport or **board** any airplane. *B.* Conditions set for Heatwole's release included a ban on him entering any airport or **boarding** any airplane.
(Evening Herald, p.17, 21/10/03)

(A.457) *A.* The scare was almost over before it **had began**. *B.* The scare was almost over before it **had begun**. *C.* The scare was almost over before it **began**.
(http://bigbrother.channel4.com/bigbrother/news/newsstory_t.jsp?id=5185, 17/7/03)

(A.458) *A.* The method SyntaxTree **take** a parameter 's1' which is the root of sentence. *B.* The method SyntaxTree **takes** a parameter 's1' which is the root of **the** sentence. *C.* The method SyntaxTree **takes** a parameter 's1' which is the root of **a** sentence.
(student assignment, 2/4/03)

(A.459) *A.* Europe is bucking the 'tough and uncertain' global outsourcing trend with increases in the number and size of deals, according the latest market statistics. *B.* Europe is bucking the 'tough and uncertain' global outsourcing trend with increases in the number and size of deals, according **to** the latest market statistics.
(IT's Monday News email, 13/10/03)

(A.460) *A.* As I went outside to join the others, it struck me that Sirius Recruitment must have known about the layoffs in order to deliver the monthly cheques **pay** several days early. *B.* As I went outside to join the others, it struck me that Sirius Recruitment must have known about the layoffs in order to deliver the monthly **pay** cheques several days early.
("An Evening of Long Goodbyes" by Paul Murray, p.372)

(A.461) *A.* You feel a little vulnerable and **its** up to family members to give support. *B.* You feel a little vulnerable and **it's** up to family members to give support.
(Evening Herald, p.28, 14/10/03)

(A.462) *A.* When I heard **was** the crashing of glass, I knew there was something coming in again. *B.* When I heard **what** was the crashing of glass, I knew there was something coming in again. *C.* When I heard the crashing of glass, I knew there was something coming in again.
(Evening Herald, p.8, 21/10/03)

(A.463) *A.* ClearPath is being positioned **at** the ideal hardware platform for the Business Blueprinting strategy that Unisys unveiled last month. *B.* ClearPath is being positioned **as** the ideal hardware platform for the Business Blueprinting strategy that Unisys unveiled last month.
(IT's Monday News electronic newsletter, 4/8/03)

(A.464) *A.* Thank **your** for your time. *B.* Thank **you** for your time.
(personal email, 18/9/03)

(A.465) *A.* Even if form of the word is in the lexicon, it does not follow that the lexicon will contain the relevant sense. *B.* Even if **the** form of the word is in the lexicon, it does not follow that the lexicon will contain the relevant sense.
(Gazdar 2003, p.20)

(A.466) *A.* Quality information is available as well for the word as for **every** information about a word. *B.* Quality information is available as well for the word as for **all** information about a word.*C.* Quality information is available as well for the word as for every **piece of** information about a word.
(Quasthoff 1998)

(A.467) *A.* The lucky man who hits the right note with the judges will be off to record an album with **others** winners from all over Europe sometime between October and December. *B.* The lucky man who hits the right note with the judges will be off to record an album with **other** winners from all over Europe sometime between October and December.
(Evening Herald, p.17, 23/9/03)

(A.468) *A.* This version **is** has a precision/recall of about 89.8% on the standard Penn treebank test set, about 0.3% lower than what can be achieved with significantly more search. *B.* This version has a precision/recall of about 89.8% on the standard Penn treebank test set, about 0.3% lower than what can be achieved with significantly more search.
(README file for Charniak's parser, downloaded fromftp://ftp.cs.brown.edu/pub/nlparser/ in March 2003)

(A.469) *A.* Validating parsers allow **to control** the input and the output of XML conversion routines. *B.* Validating parsers allow the input and the output of XML conversion routines **to be controlled.** *C.* Validating parsers allow **us** to control the input and the output of XML conversion routines.
(Mengel and Lezius 2000)

(A.470) *A.* IBM spent USD 5 billion on R&D worldwide in 2002 and as one of only seven CAS sites across IBM's 165 worldwide offices, the Dublin Software Lab will use the new status to build stronger relationships with **Ireland** academic institutions, supporting local Research and Development projects. *B.* IBM spent USD 5 billion on R&D worldwide in 2002 and as one of only seven CAS sites across IBM's 165 worldwide offices, the Dublin Software Lab will use the new status to

build stronger relationships with **Ireland's** academic institutions, supporting local Research and Development projects. *C.* IBM spent USD 5 billion on R&D worldwide in 2002 and as one of only seven CAS sites across IBM's 165 worldwide offices, the Dublin Software Lab will use the new status to build stronger relationships with **Irish** academic institutions, supporting local Research and Development projects.
(IT's Monday News electronic newsletter, 8/9/03)

(A.471) *A.* HP America is **ending** staff over to work on the programmes rather than basing them with a US university and already has six US researchers working full time in Ireland. *B.* HP America is **sending** staff over to work on the programmes rather than basing them with a US university and already has six US researchers working full time in Ireland.
(IT's Monday News electronic newsletter, 21st July 2001)

(A.472) *A.* Another traditional definition consists **with** replacing in Definition 1 the set of edges with a transition function. *B.* Another traditional definition consists **of** replacing in Definition 1 the set of edges with a transition function.
(Roche and Schabes 1997, p.3)

(A.473) *A.* I fail to see what is strange in the first, given that **fox** is not strange and, as to glasnost, it is simply not English. *B.* I fail to see what is strange in the first, given that **fax** is not strange and, as to glasnost, it is simply not English.
(James 1998, p.75)

(A.474) *A.* They can no longer indulge themselves by driving a "peace" process with the **the** bomb and bullet in the background. *B.* They can no longer indulge themselves by driving a "peace" process with the bomb and bullet in the background.
(Evening Herald, p.14, 1/12/03)

(A.475) *A.* The collection of a parallel corpus of grammatical sentences in conjunction with the **conjunction with the** collection of a corpus of ungrammatical ones provides a ready-made evaluation procedure. *B.* The collection of a parallel corpus of grammatical sentences in conjunction with the collection of a corpus of ungrammatical ones provides a ready-made evaluation procedure.
(draft of own writing, 12/9/03)

(A.476) *A.* Beef **of** Salmon **of** the menu at Punchestown *B.* Beef **or** Salmon **off** the menu at Punchestown
(ireland.com, 30/4/03)

(A.477) *A.* So **lets** stop bashing the Byrnes and pick on someone who deserves it for a change. *B.* So **let's** stop bashing the Byrnes and pick on someone who deserves it for a change.
(Irish Daily Mirror, p.27, 3/12/03)

(A.478) *A.* These figures suggest that the parser copes quite well with the problem of ungrammaticality but they will make more sense when compared with figures for other **parser**, which is the next step in this research. *B.* These figures suggest that the parser copes quite well with the problem of ungrammaticality but they will make more sense when compared with figures for other **parsers**, which is the next step in this research.
(draft of own writing, 6/11/03)

(A.479) *A.* The **apartments** is suitable for 1-2 professional persons. *B.* The **apartment** is suitable for 1-2 professional persons.
(estate agent's electronic newsletter, 10/10/03)

(A.480) *A.* A utility class StringTokenizer provides methods divide a string into tokens. *B.* A utility class StringTokenizer provides methods **to** divide a string into tokens. *C.* A utility class StringTokenizer provides methods **which** divide a string into tokens.
(student assignment, Feb 2003)

(A.481) *A.* Product licence revenue is still rising, we have introduced a subscription model that allows companies to rent our software and we have turned our cost centres into **a** profit centres. *B.* Product licence revenue is still rising, we have introduced a subscription model that allows companies to rent our software and we have turned our cost centres into profit centres.
(IT's Monday News electronic newsletter, 29/9/03)

(A.482) *A.* The **had** disk password prevents unauthorized access to the hard disk drives and is checked internally each time the system is booted. *B.* The **hard** disk password prevents unauthorized access to the hard disk drives and is checked internally each time the system is booted.
(Laptop user guide, p.67, 20/11/03)

(A.483) *A.* I hope, if they have their photos, **then** will leave us in peace, and let us simply enjoy the day. *B.* I hope, if they have their photos, **they** will leave us in peace, and let us simply enjoy the day. *C.* I hope, if they have their photos, **then they** will leave us in peace, and let us simply enjoy the day.
(Evening Herald, p.3, 12/11/03)

(A.484) *A.* Michael managed to get a job as London bus driver for 10 years, and in his mid-30s he became a truck driver. *B.* Michael managed to get a job as **a** London bus driver for 10 years, and in his mid-30s he became a truck driver.
(Evening Herald, p.4, 28/11/03)

(A.485) *A.* Award-winning Irish comedian Tommy Tiernan **choose** last night's premiere of Irish movie Spin the Bottle to show **of** the new love in his life, Yvonne McMahon, a record company executive from Dublin. *B.* Award-winning Irish comedian Tommy Tiernan **chose** last night's premiere of Irish movie Spin the Bottle to show **off** the new love in his life, Yvonne McMahon, a record company executive from Dublin.
(Evening Herald, p.23, 27/11/03)

(A.486) *A.* Ossidian had originally put feelers out into Australian **market** through the Australian office of Enterprise Ireland. *B.* Ossidian had originally put feelers out into **the** Australian market through the Australian office of Enterprise Ireland. *C.* Ossidian had originally put feelers out into Australian **markets** through the Australian office of Enterprise Ireland.
(IT's Monday News electronic newsletter, 9th June)

(A.487) *A.* The pragmatic components are supported by knowledge sources that hold for each participant about his knowledge of the world, the partner and the course of the dialogue dependent **of** the task. *B.* The pragmatic components are supported by knowledge sources that hold for each participant about his knowledge of the world, the partner and the course of the dialogue dependent **on** the task.
(Gehrke 1983, p.185)

(A.488) *A.* I'm not sure what I'm up tomorrow but if I'm in town I'll call over that way. *B.* I'm not sure what I'm up **to** tomorrow but if I'm in town I'll call over that way.
(personal email, 9/1/04)

(A.489) *A.* It is able to **chose** a pronoun which matches with the verb. *B.* It is able to **choose** a pronoun which matches with the verb.
(student assignment, Feb 2003)

(A.490) *A.* **Its** really georgous here. *B.* **It's** really georgous here.
(personal email, 15/1/04)

(A.491) *A.* The tortoise duly carries the scorpion to safety, but just as they on the bank, the scorpion digs its poison into the tortoise. *B.* The tortoise duly carries the scorpion to safety, but just as they **are** on the bank, the scorpion digs its poison into the tortoise.
(Irish Times letters page, 19/6/03)

(A.492) *A.* Parsing, if understood as a disambiguation **procedures**, is put down to the principle of parsimony.*B.* Parsing, if understood as a disambiguation **procedure**, is put down to the principle of parsimony.
(Ballim *et al.* 1999, p.11)

(A.493) *A.* The nearest to disaster the soccer version of our saints and scholars came when two international players shared a prostitute on one trip abroad and then refused to pay her. *B.* The nearest to disaster the soccer version of our saints and scholars came **was** when two international players shared a prostitute on one trip abroad and then refused to pay her.
(Evening Herald, p.4, 9/10/03)

(A.494) *A.* Ruud van Nistelrooy is fearing a much tougher challenge from Aston Villa **in Sunday's in** FA Cup third round clash than he encountered the last time his Manchester United team-mates took on David O'Leary's men. *B.* Ruud van Nistelrooy is fearing a much tougher challenge from Aston Villa **in Sunday's FA** Cup third round clash than he encountered the last time his Manchester United team-mates took on David O'Leary's men. *C.* Ruud van Nistelrooy is fearing a much tougher challenge from Aston Villa **on Sunday in the** FA Cup third round clash than he encountered the last time his Manchester United team-mates took on David O'Leary's men.
(Evening Herald, p.60, 2/1/04)

(A.495) *A.* Collins then extends his model to a history based model so that the probability of **a** generating a modifier can be conditioned on any of the structure previously generated. *B.* Collins then extends his model to a history based model so that the probability of generating a modifier can be conditioned on any of the structure previously generated.
(draft of paper on probabilistic parsing by Deirdre Hogan, May 2003)

(A.496) *A.* The error position is indicated by the edges and nodes **build** upon the words. *B.* The error position is indicated by the edges and nodes **built** upon the words.
(Becker *et al.* 1999)

(A.497) *A.* This is the case especially for higher frequency words because they should possess **these** information.*B.* This is the case especially for higher frequency words because they should possess **this** information. *C.* This is the case especially for higher frequency words because they should possess these **pieces of** information.
(Quasthoff 1998)

(A.498) *A.* The six were gathered on the sofa looking nervous when Big Brother made the announcement to **the** shortly after 13.30. *B.* The six were gathered on the sofa looking nervous when Big Brother made the announcement to **the house** shortly after 13.30. *C.* The six were gathered on the sofa looking nervous when Big Brother made the announcement to **them** shortly after 13.30.
(http://bigbrother.channel4.com/bigbrother/news/newsstory_t.jsp?id=5142, 15/7/03)

(A.499) *A.* From 2001 to 2002 I worked for **for** Reuters in London as part of their Research and Standards Group where, among other things, I was involved in the production and dissemination of the Reuters Corpus. *B.* From 2001 to 2002 I worked for Reuters in London as part of their Research and Standards Group where, among other things, I was involved in the production and dissemination of the Reuters Corpus.
(http://www.dcs.shef.ac.uk/ marks/, April 2003)

(A.500) *A.* The closure **in** computed breadth-first. *B.* The closure **is** computed breadth-first.
(ALE User Guide, Version 3.2.1. December 2001, p.56)

(A.501) *A.* For those of you who know and for those of you who don't, I **will** leaving Ireland to head off to the US to work for a year (or more!) at the start of March. *B.* For those of you who know and for those of you who don't, I **will be** leaving Ireland to head off to the US to work for a year (or more!) at the start of March. *C.* For those of you who know and for those of you who don't, I **am** leaving Ireland to head off to the US to work for a year (or more!) at the start of March.
(personal email, 15/1/04)

(A.502) *A.* The 52 year-old married **father-of-two** adult children consented to police searches of his homes in London and Sri Lanka during the investigation. *B.* The 52 year-old married **father of two** adult children consented to police searches of his homes in London and Sri Lanka during the investigation.
(Evening Herald, p.39, 10/10/03)

(A.503) *A.* In particular biometrics can be used **use** to prove a person's right to access personal data that may be held by an organisation and also prevent identity theft. *B.* In particular biometrics can be used to prove a person's right to access personal data that may be held by an organisation and also prevent identity theft.
(IT's Monday News electronic newsletter, 13/10/03)

(A.504) *A.* Among those standing in front a solitary bin truck was Mary Lee, sister of jailed TD and bin charges opponent Joe Higgins. *B.* Among those standing in front **of** a solitary bin truck was Mary Lee, sister of jailed TD and bin charges opponent Joe Higgins.
(Evening Herald, 29/9/03, p.11)

(A.505) *A.* Please tell **you** friends. *B.* Please tell **your** friends.
(personal email, 25/8/03)

(A.506) *A.* Will round town from 3. *B.* Will **be** round town from 3.
(text message, 22/12/03)

(A.507) *A.* While there have been point to point failovers demonstrated before, according to Dr Helmut Beck of Fujitsu Siemens, was that the failover was a world first in an asynchronous networking environment. *B.* While there have been point to point failovers demonstrated before, according to Dr Helmut Beck of Fujitsu Siemens, **the difference** was that the failover was a world first in an asynchronous networking environment.
(IT's Monday News electronic newsletter, 13/10/03)

(A.508) *A.* Galway light-welterweight Francis Barrett repeated last year's points win over David Kirk, winning every round of a six-rounder at **Londons** Alexandra Palace. *B.* Galway light-welterweight Francis Barrett repeated last year's points win over David Kirk, winning every round of a six-rounder at **London's** Alexandra Palace.
(Evening Herald, p.53, 11/10/03)

(A.509) *A.* Limiting **one attention's** to one particular type of text means that only the error types typically associated with this text type are likely to be found. *B.* Limiting **one's attention** to one particular type of text means that only the error types typically associated with this text type are likely to be found.
(draft of own writing, 30/10/03)

(A.510) *A.* However, this fad about Atkin's diet sounds almost medieval. *B.* However, this fad about **the** Atkin's diet sounds almost medieval.
(Evening Herald, p.26, 21/10/03)

(A.511) *A.* Belfast based software company Latens has developed an access control system for digital cable and two way broadband operators, which could **renders** smartcards redundant. *B.* Belfast based software company Latens has developed an access control system for digital cable and two way broadband operators, which could **render** smartcards redundant.
(IT's Monday News electronic newsletter, 14th July 2003)

(A.512) *A.* **Its** time to unleash the power within. *B.* **It's** time to unleash the power within.
(Evening Herald, p.26, 11/10/03)

(A.513) *A.* We had a bit of bad weather which we **where** not expecting. *B.* We had a bit of bad weather which we **were** not expecting.
(personal email, 18/5/03)

(A.514) *A.* I also think it might be useful to elect **the our** Departmental rep at the end of the month, certainly before the 22 October departmental meeting. *B.* I also think it might be useful to elect **our** Departmental rep at the end of the month, certainly before the 22 October departmental meeting. *C.* I also think it might be useful to elect **the** Departmental rep at the end of the month, certainly before the 22 October departmental meeting.
(personal email, 2/9/03)

(A.515) *A.* The evidence was that a gunman positioned on the roof had fired three shots **shots** through the bathroom window, breaking it. *B.* The evidence was that a gunman positioned on the roof had fired three shots through the bathroom window, breaking it.
(Evening Herald, p.6, 14/10/03)

(A.516) *A.* A pause followed where I realised I was supposed to **responded.** *B.* A pause followed where I realised I was supposed to **have** responded. *C.* A pause followed where I realised I was supposed to **respond.**
(It's Friday magazine, p.23, Evening Herald, 26/9/03)

(A.517) *A.* Wait until tomorrow before making changes of any kind **if where** home matters are concerned. *B.* Wait until tomorrow before making changes of any kind **where** home matters are concerned. *C.* Wait until tomorrow before making changes of any kind **if** home matters are concerned.
(Evening Herald, p.26, 11/10/03)

(A.518) *A.* Is she always on the go or **act** as if she is driven by a motor? B. Is she always on the go or **acting** as if she is driven by a motor? C. Is she always on the go or **does she** act as if she is driven by a motor?
(Evening Herald, p.25, 21/10/03)

(A.519) *A.* Perhaps the most obvious similarity is the reduced role of transformations in **the these** theories. *B.* Perhaps the most obvious similarity is the reduced role of transformations in **these** theories *C.* Perhaps the most obvious similarity is the reduced role of transformations in **the** theories
(Wasow 1985, p.200)

(A.520) *A.* The new IT infrastructure will also supply the University with powerful back-up and recovery capabilities and very high availability and will enable it to reduce number of e-mail servers from 60 to six with estimated cost savings of more than 50 per cent. *B.* The new IT infrastructure will also supply the University with powerful back-up and recovery capabilities and very high availability and will enable it to reduce **the** number of e-mail servers from 60 to six with estimated cost savings of more than 50 per cent.
(IT's Monday electronic newsletter, 30th June 2003)

(A.521) *A.* The class of 2003 have turned their back on business and science in favour of "old reliable" professions such medicine, law and teaching, the CAO offers confirm. *B.* The class of 2003 have turned their back on business and science in favour of "old reliable" professions such **as** medicine, law and teaching, the CAO offers confirm.
(Evening Herald, Going to College supplement, p.2, 19/8/03)

(A.522) *A.* WE **WE** HAVE THEM ALL Prescribed Online! *B.* WE HAVE THEM ALL Prescribed Online!
(header of spam email, 13/10/03)

(A.523) *A.* The midfield is **were** Dublin could control the contest. *B.* The midfield is **where** Dublin could control the contest.
(Evening Herald, p.55, 27/9/03)

(A.524) *A.* US must **sent** inspectors back. *B.* US must **send** inspectors back.
(Ceefax News Headline, 22/4/03)

(A.525) *A.* Andy's bird turned up **to**, but it was all grandy grand. *B.* Andy's bird turned up **too**, but it was all grandy grand.
(personal email, 27/10/02)

(A.526) *A.* Last March in the neighbouring Serpentine Avenue developer Sean Dunne also received an exemption for 14 apartments in this posh Ballsbridge near AIB's head office. *B.* Last March in the neighbouring Serpentine Avenue developer Sean Dunne also received an exemption for 14 apartments in this posh Ballsbridge **development** near AIB's head office.
(Irish Independent Property, p.2., 11th July 2003)

(A.527) *A.* They are responsible for purchasing ad break slots for **client's** who want their brands associated with a given programme. *B.* They are responsible for purchasing ad break slots for **clients** who want their brands associated with a given programme.
(Evening Herald, p.4, 27/9/03)

(A.528) *A.* Although the logical meaning of a unification is that the two values should be the same, procedurally there a processing direction. *B.* Although the logical meaning of a unification is that the two values should be the same, procedurally there **is** a processing direction.
(Fouvry 2003b, p.227)

(A.529) *A.* If the predictions about these word sequences made by the analysis **matches** the grammaticality judgements collected by the linguist, the linguist can conclude that his analysis is correct. *B.* If the predictions about these word sequences made by the analysis **match** the grammaticality judgements collected by the linguist, the linguist can conclude that his analysis is correct.
(draft of own writing, 25/8/03)

(A.530) *A.* Let cooperate to achieve the result that is right for the student. *B.* Let **us** cooperate to achieve the result that is right for the student.
(personal email, 14/4/03)

(A.531) *A.* The point is this: when dealing with uncertainty, the probabilities of the model must reflect, as **good** as possible, the missing knowledge, which is, to a large extent, extra-linguistic in nature. *B.* The point is this: when dealing with uncertainty, the probabilities of the model must reflect, as **well** as possible, the missing knowledge, which is, to a large extent, extra-linguistic in nature.
(Sima'an 2003, p.9)

(A.532) *A.* They would need **to** good experience of database technology including getting different systems to talk to each other, and XML and they should have done something similar before. *B.* They would need to **have** good experience of database technology including getting different systems to talk to each other, and XML and they should have done something similar before. *C.* They would need good experience of database technology including getting different systems to talk to each other, and XML and they should have done something similar before.
(personal email, 11/4/03)

(A.533) *A.* What does a single line **yellow** mean? *B.* What does a single **yellow** line mean?
(Irish School of Motoring, Rules of the Road Question Sheet)

(A.534) *A.* The method nextInt gets a random number that upper bound of 17. *B.* The method nextInt gets a random number that **has an** upper bound of 17.
(student assignment, Feb 2003)

(A.535) *A.* Retailers selling gifts are expected to be the big winner this year but expenditure on most other categories **are** expected to be lower. *B.* Retailers selling gifts are expected to be the big winner this year but expenditure on most other categories **is** expected to be lower.
(Evening Herald, p.6, 8/12/03)

(A.536) *A.* This week, a trio of fortysomething women get some intensive coaching on how to **to** combat, fashion-wise, the hazards of those advancing years. *B.* This week, a trio of fortysomething women get some intensive coaching on how to combat, fashion-wise, the hazards of those advancing years.
(Evening Herald, p.37,12/11/03)

(A.537) *A.* This class also contains a static method called IndexOfSubject, which indicates index of the pronoun in an array. *B.* This class also contains a static method called IndexOfSubject, which indicates **the** index of the pronoun in an array.
(student assignment, Feb 2003)

(A.538) *A.* Mr Latham suggested Huntley appreciated from the papers that witnesses had claimed to have seen **he** girls well after 6.45pm. *B.* Mr Latham suggested Huntley appreciated from the papers that witnesses had claimed to have seen **the** girls well after 6.45pm.
(Evening Herald, p.3, 2/12/03)

(A.539) *A.* Our aim in this paper is to show that combining both learning styles produces a grammar that assigns more plausible parses than is **then** case for grammars learnt using either learning style in isolation. *B.* Our aim in this paper is to show that combining both learning styles produces a grammar that assigns more plausible parses than is **the** case for grammars learnt using either learning style in isolation.
(Osbourne and Bridge 1994)

(A.540) *A.* Tom also is taking us to **an** footie match which shall be interesting. *B.* Tom also is taking us to **a** footie match which shall be interesting. C. Tom also is taking us to an **australian** footie match which shall be interesting.
(personal email, 22/11/02)

(A.541) *A.* It needs to be updated as it's a bit messy but it has links to previous **years** talks at the bottom. *B.* It needs to be updated as it's a bit messy but it has links to previous **years'** talks at the bottom.
(personal email, 23/10/03)

(A.542) *A.* It's not that we haven't heard love songs this overt before, but not since Van Morrison has an Irish singer opened his heart and put it over the length a record in such an honest manner. *B.* It's not that we haven't heard love songs this overt before, but not since Van Morrison has an Irish singer opened his heart and put it over the length **of** a record in such an honest manner.
(Dublin Event Guide, p. 19, 13th-26th August)

(A.543) *A.* The preposition which was required by another lexical item in the sentence. *B.* The preposition which was **omitted was** required by another lexical item in the sentence. *C.* The preposition omitted was required by another lexical item in the sentence.
(draft of own writing, 23/10/03)

(A.544) *A.* Morris used 53 players last year, only 30 are permitted on an official squad, while **medicals** bills were also high. B. Morris used 53 players last year, only 30 are permitted on an official squad, while **medical** bills were also high.
(Evening Herald, p.54, 11/10/03)

(A.545) *A.* Well, back during the summer when we hadn't addressed shifting from the 8pm schedule, I agreed to give a research talk Centre for Language and Communication Studies taught postgraduates at 6 today. *B.* Well, back during the summer when we hadn't addressed shifting from the 8pm schedule, I agreed to give a research talk **to** Centre for Language and Communication Studies taught postgraduates at 6 today.
(personal email, 13/11/02)

(A.546) *A.* Despite the widespread media interest in the camp it was a lot smaller **that** many of the public realised. *B.* Despite the widespread media interest in the camp it was a lot smaller **than** many of the public realised.
(Irish Times, p.3. 5th February, 2003)

(A.547) *A.* The company also said it is cooperating with law enforcement authorities and broadband providers to stop illegal online trading of the games, which normally sell for **between** over STG 25. *B.* The company also said it is cooperating with law enforcement authorities and broadband providers to stop illegal online trading of the games, which normally sell for over STG 25.
(IT's Monday News electronic newsletter, 17/11/03)

(A.548) *A.* She speaks of it in the past tense and says it was a problem which occurred when she was faced with things **were** out of her control. *B.* She speaks of it in the past tense and says it was a problem which occurred when she was faced with things **which** were out of her control. *C.* She speaks of it in the past tense and says it was a problem which occurred when she was faced with things out of her control.
(Irish Times Magazine, p.14, 16/8/03)

(A.549) *A.* The linguistic grammars are devices that delimit the exact borders of the **the** language as a set of utterances. B. The linguistic grammars are devices that delimit the exact borders of the language as a set of utterances.
(Sima'an 2003, p.4)

(A.550) *A.* In An Irishman's Diary, he denigrates those who take a different view of the Iraq war to his own as simplistic peace-lovers, blithering idiots, cretins and, **wit** regard to one honest lady, having epic stupidity!B. In An Irishman's Diary, he denigrates those who take a different view of the Iraq war to his own as simplistic peace-lovers, blithering idiots, cretins and, **with** regard to one honest lady, having epic stupidity!
(Letters to the Editor, Irish Times, Saturday 5th April, 2003)

(A.551) *A.* Since humans are capable **to deal** with acceptable ill-formed text without any deep competence of the underlying structure of the language, it seems that proposed theory and systems are not able to perform an approximate matching between input and pre-defined structures. *B.* Since humans are capable **of dealing** with acceptable ill-formed text without any deep competence of the underlying structure of the language, it seems that proposed theory and systems are not able to perform an approximate matching between input and pre-defined structures.
(Ballim *et al.* 1999, p.3)

(A.552) *A.* The accommodation in brief **comprises of** two double bedrooms, main ensuite, fully equipped kitchen. *B.* The accommodation in brief **comprises** two double bedrooms, main ensuite, fully equipped kitchen. *C.* The accommodation in brief **consists of** two double bedrooms, main ensuite, fully equipped kitchen.
(estate agent's electronic newsletter, 9/10/03)

(A.553) *A.* Bespoke software for this kind of problem is just too expensive and **need** dedicated IT people who are expensive, and businesses in every sector are grappling with delivering product faster, cheaper and better. *B.* Bespoke software for this kind of problem is just too expensive and **needs** dedicated IT people who are expensive, and businesses in every sector are grappling with delivering product faster, cheaper and better.
(IT's Monday News electronic newsletter, 20/10/03)

(A.554) *A.* Campbell Scott, product director with O2 said that, while the text market is being driven by younger consumers, SMS is increasingly **been** used as a business tool for the same reasons. *B.* Campbell Scott, product director with O2 said that, while the text market is being driven by younger consumers, SMS is increasingly **being** used as a business tool for the same reasons.
(IT's Monday news electronic newsletters, 25th August 2003)

(A.555) *A.* A closer look at the situations in which NLP-systems perform **suboptimal** reveals that problems arise in the following areas. *B.* A closer look at the situations in which NLP-systems perform **suboptimally** reveals that problems arise in the following areas. *C.* A closer look at the situations in which NLP-systems perform **in a** suboptimal **way** reveals that problems arise in the following areas.
(Ballim *et al.* 1999, p.4)

(A.556) *A.* But the pen and paper I find for her are not as good as **the those** my son has taken from his bag and a huge row erupts. *B.* But the pen and paper I find for her are not as good as **those** my son has taken from his bag and a huge row erupts. *C.* But the pen and paper I find for her are not as good as **the pen and paper** my son has taken from his bag and a huge row erupts.
(Evening Herald, p.25, 14/10/03)

(A.557) *A.* However, most of the strategies are capable of using contextual restrictions on what incorrect lexical item might be, and therefore are most effective when the constraints on the unknown word are strongest. *B.* However, most of the strategies are capable of using contextual restrictions on what **the** incorrect lexical item might be, and therefore are most effective when the constraints on the unknown word are strongest.
(Carbonell and Hayes 1983)

(A.558) *A.* This, in turn, requires an examination the use of evidence and structure of arguments characteristic of a given research program. *B.* This, in turn, requires an examination **of** the use of evidence and structure of arguments characteristic of a given research program.
(Perry 1979b, p.3)

(A.559) *A.* Everybody says that and **its** not true. *B.* Everybody says that and **it's** not true.
(IT's Monday electronic newsletter, 23rd June 2003)

(A.560) *A.* Rule-based taggers generally involve a large database of hand-written disambiguation **rule** which specify , for example, that an ambiguous word is a noun rather than a verb if it follows a determiner. *B.* Rule-based taggers generally involve a large database of hand-written disambiguation **rules** which specify , for example, that an ambiguous word is a noun rather than a verb if it follows a determiner.
(Jurafsky and Martin 2000, p.300)

(A.561) *A.* either a word following the same rules **at** a category name, or a reference to a rule category occurrence *B.* either a word following the same rules **as** a category name, or a reference to a rule category occurrence
(User Manual for the JavaChart 0.1.3 system, p.9, downloaded May 2003)

(A.562) *A.* We should I think be able to make that **attractive concept** even if J2EE programming isn't all that interesting. *B.* We should I think be able to make that **concept attractive** even if J2EE programming isn't all that interesting. *C.* We should I think be able to make that **an attractive**

concept even if J2EE programming isn't all that interesting.
(personal email, 8/12/03)

(A.563) *A.* The company has simply hired the capability it needed in the form of 12 additional staff for **is** engineering department and is now only looking for system integration partners on an ad hoc basis. *B.* The company has simply hired the capability it needed in the form of 12 additional staff for **its** engineering department and is now only looking for system integration partners on an ad hoc basis.
(IT's Monday News electronic newsletter, 13/12/03)

(A.564) *A.* Although the US parent's legal travails did not directly **effect** the Irish operations, Walsh said his focus was on ensuring service levels did not drop during the reorganisation. *B.* Although the US parent's legal travails did not directly **affect** the Irish operations, Walsh said his focus was on ensuring service levels did not drop during the reorganisation.
(IT's Monday News electronic newsletter, 20/10/03)

(A.565) *A.* In his presence, Brent **would** often **became** stunned by resentment, occasionally trickling out as blithering half-insults, prefaced by a muted whine as he searched desperately for the killer punchline. *B.* In his presence, Brent **would** often **become** stunned by resentment, occasionally trickling out as blithering half-insults, prefaced by a muted whine as he searched desperately for the killer punchline. C. In his presence, Brent often **became** stunned by resentment, occasionally trickling out as blithering half-insults, prefaced by a muted whine as he searched desperately for the killer punchline.
(Irish Times Weekend Review, p.18, 3/1/04)

(A.566) *A.* It all points to the same thing - businesses use technology to reduce costs or gain greater market share at the expense of others and their jobs. *B.* It all points to the same thing - businesses use technology to reduce costs or gain greater market share at the expense of **others'** and their jobs.
(personal email, 8/12/03)

(A.567) *A.* Large-size corpora, are used to evaluate and compare each other the systems which have already been developed. *B.* Large-size corpora, are used to evaluate and compare **to** each other the systems which have already been developed.
(www.elra.info, Language Resources section, accessed 16/10/03)

(A.568) *A.* Like so many before him, Eamon Zayed had struggled to find his feet in the league following his release from a big English club, but the Dubliner really impressed in the 2003 season, scoring 19 goals, and he could even have **score** more had he not missed the last four games of the season as he was on international duty with the Irish U20s at the World Cup finals in the UAE, where he played in all four games. *B.* Like so many before him, Eamon Zayed had struggled to find his feet in the league following his release from a big English club, but the Dubliner really impressed in the 2003 season, scoring 19 goals, and he could even have **scored** more had he not missed the last four games of the season as he was on international duty with the Irish U20s at the World Cup finals in the UAE, where he played in all four games.
(Evening Herald, p.59, 2/1/04)

(A.569) *A.* The longest **journey begin** with one step. *B.* The longest journey **begins** with one step. *C.* The longest **journeys** begin with one step.
(electronic newsletter about petition to EU on behalf of Irish language, 8/1/04)

(A.570) *A.* Last week Sun announced the creation of the Java Tools Community (JTC) but critics have already doomed the initiative to failure when it became clear that **the** two **of the** leading development tools providers in the Java space - IBM and Borland, which have 70 per cent of the market between them - have decided not to support it. *B.* Last week Sun announced the creation of the Java Tools Community (JTC) but critics have already doomed the initiative to failure when it became clear that two **of the** leading development tools providers in the Java space - IBM and Borland, which have 70 per cent of the market between them - have decided not to support it. *C.* Last week Sun announced the creation of the Java Tools Community (JTC) but critics have already doomed the initiative to failure when it became clear that **the** two leading development tools providers in the Java space - IBM and Borland, which have 70 per cent of the market between them - have decided not to support it.
(IT's Monday News electronic newsletter, 19/1/04)

(A.571) *A.* Everyone was looking at me to see what I would do, but **the he** just reached in for his teeth, picked them up, licked them, put them back in his mouth and slapped me around the head, telling me to eat up. *B.* Everyone was looking at me to see what I would do, but **he** just reached in for his teeth, picked them up, licked them, put them back in his mouth and slapped me around the head, telling me to eat up. *C.* Everyone was looking at me to see what I would do, but **the man** just reached in for his teeth, picked them up, licked them, put them back in his mouth and

slapped me around the head, telling me to eat up.
(Evening Herald, p.3, 28/11/03)

(A.572) *A.* Closing date for submissions **in** Friday 27th June. *B.* Closing date for submissions **is** Friday 27th June.
(http://www.cs.tcd.ie/aiai/aics/, June 2003)

(A.573) *A.* So if any of you are in or around town **lets** meet up some time for a drink or lunch now that I'm on Mount Street. *B.* So if any of you are in or around town **let's** meet up some time for a drink or lunch now that I'm on Mount Street.
(personal email, 8/1/04)

(A.574) *A.* In November, NetApp made a swift entry into the ILM market with its purchase of Spinnaker Networks, in a deal aimed at boosting NetApp's high-end storage products line and **get** its 'storage grid' architecture to market more quickly. *B.* In November, NetApp made a swift entry into the ILM market with its purchase of Spinnaker Networks, in a deal aimed at boosting NetApp's high-end storage products line and **getting** its 'storage grid' architecture to market more quickly.
(IT's Monday News electronic newsletter, 8/12/03)

(A.575) *A.* Generally speaking, it's not a bad idea to identify a common neurosis and publish something **geared** precisely **targeted** towards the sufferers if you want to make a publishing fortune. *B.* Generally speaking, it's not a bad idea to identify a common neurosis and publish something **geared** precisely towards the sufferers if you want to make a publishing fortune. *C.* Generally speaking, it's not a bad idea to identify a common neurosis and publish something precisely **targeted** towards the sufferers if you want to make a publishing fortune.
(Evening Herald, 14/10/03, p.22)

(A.576) *A.* The increased financial clout will help the struggling Pivotal stabilise and improve its chances of finding a strong sales partner to **breath** life into its moribund Irish operation. *B.* The increased financial clout will help the struggling Pivotal stabilise and improve its chances of finding a strong sales partner to **breathe** life into its moribund Irish operation.
(IT's Monday News electronic newsletter, 13/12/03)

(A.577) *A.* **Its** just like a game! *B.* **It's** just like a game!
(personal email, 6/10/04)

(A.578) *A.* In this paper we present the design of a digital resource which will be used as a repository of information **of** linguistic errors.*B.* In this paper we present the design of a digital resource which will be used as a repository of information **on** linguistic errors.
(Aduriz *et al.* 2002)

(A.579) *A.* Ok **ill** try to get the five o clock then. *B.* Ok **i'll** try to get the five o clock then.
(personal text message, 7/1/04)

(A.580) *A.* Not **everyones** cup of tea, but some people might just like to know that IT offers a perhaps unique possibility in terms of professional employment of fitting work into **ones** life and not the other way around. *B.* Not **everyone's** cup of tea, but some people might just like to know that IT offers a perhaps unique possibility in terms of professional employment of fitting work into **one's** life and not the other way around.
(personal email, 8/12/03)

(A.581) *A.* Yet it captures **quite plausibility** a certain predominant form of life today.*B.* Yet it captures quite **plausibly** a certain predominant form of life today.*C.* Yet it captures **with** plausibility a certain predominant form of life today.
(The Sociological Imagination: Hilary Tovey and Brian Torode: Course Outline for Part Two, Jan 2003)

(A.582) *A.* But a source **as** Lillie's Bordello said today: "Irish soccer stars are a lot gentler than the British stars to handle."B. But a source **at** Lillie's Bordello said today: "Irish soccer stars are a lot gentler than the British stars to handle."
(Evening Herald, p.2, 9/10/03)

(A.583) *A.* This corpus contains the text of many articles taken from the Wall Street Journal without modification, except for the separation of punctuation marks from the words **the** adjoin.*B.* This corpus contains the text of many articles taken from the Wall Street Journal without modification, except for the separation of punctuation marks from the words **they** adjoin.
(Charniak 1997)

(A.584) *A.* Don't let anyone tell you**.** You are being overambitious. *B.* Don't let anyone tell you you are being overambitious. *C.* Don't let anyone tell you **that** you are being overambitious.
(Evening Herald, p.28, 14/10/03)

(A.585) *A.* In the following paragraphs, I show that the order ideal still gives us too many elements, but that at the same time it is not possible reduce it further without losing some crucial properties. B. In the following paragraphs, I show that the order ideal still gives us too many elements, but that at the same time it is not possible **to** reduce it further without losing some crucial properties.
(Fouvry 2003b, p.127)

(A.586) *A.* If not, i was wondering if you and james wanted **top** call up for a drink or two or three before the gig ?? *B.* if not, i was wondering if you and james wanted **to** call up for a drink or two or three before the gig ??
(personal email, Feb 2003)

(A.587) *A.* It is locked in place and if **its** moved it is extremely difficult for me to move it back in place and to get at the other shelves. *B.* It is locked in place and if **it's** moved it is extremely difficult for me to move it back in place and to get at the other shelves.
(personal email, 26/11/03)

(A.588) *A.* **There** are going to **hit** a lot of problems along the way, people running small enough companies, working their asses off for next to nothing. *B.* **They** are going to hit a lot of problems along the way, people running small enough companies, working their asses off for next to nothing. *C.* There are going to **be** a lot of problems along the way, people running small enough companies, working their asses off for next to nothing.
(Evening Herald, p. 13, 18/8/03)

(A.589) *A.* Joe immediately makes plans to **to** rob Underworld before Mike can resume the stock check and expose the fraud. *B.* Joe immediately makes plans to rob Underworld before Mike can resume the stock check and expose the fraud.
(Woman's Way, 3/6/03, p.48)

(A.590) *A.* Of course, our infrastructure needs radical improvement but when motorists can benefit from a 20pc decrease in premiums simply because everyone is simply driving more carefully in an effort not pick up points, it proves that road manners can always be improved. *B.* Of course, our infrastructure needs radical improvement but when motorists can benefit from a 20pc decrease in premiums simply because everyone is simply driving more carefully in an effort not **to** pick up points, it proves that road manners can always be improved.
(Evening Herald, p.14, 29/10/03)

(A.591) *A.* **Its** a software company down the road in pearse street so I'm enjoying myself down here for the moment. *B.* **It's** a software company down the road in pearse street so I'm enjoying myself down here for the moment.
(personal email, 27/5/03)

(A.592) *A.* Two fire units, an ambulance and an emergency unit **was** sent to the scene. *B.* Two fire units, an ambulance and an emergency unit **were** sent to the scene.
(Evening Herald, 2/12/03, p.8)

(A.593) *A.* They can be extracted either from the previous analysis at lower linguistic level or directly from the text. *B.* They can be extracted either from the previous analysis at **a** lower linguistic level or directly from the text.
(Ballim *et al.* 1999, p.13)

(A.594) *A.* If you ordered a Powerpoint Table Top Fridge and Larder Fridge Powerpoint from DID Electrical in October please let me asap. *B.* If you ordered a Powerpoint Table Top Fridge and Larder Fridge Powerpoint from DID Electrical in October please let me **know** asap.
(personal email, 19/11/03)

(A.595) *A.* Phrase-structure syntax allows **describing** the syntactic structure of utterances in terms of the phrases which constitute them, using a hierarchical and recursive set of concepts. *B.* Phrase-structure syntax allows the syntactic structure of utterances **to be described** in terms of the phrases which constitute them, using a hierarchical and recursive set of concepts. *C.* Phrase-structure syntax allows **the** describing **of the** syntactic structure of utterances in terms of the phrases which constitute them, using a hierarchical and recursive set of concepts.
(Musillo and Sima'an 2002)

(A.596) *A.* A new credit control policy has also been implemented at the IEDR **.** which requires domains to be paid for before activation and renewals will have to be paid within 30 days of the date of invoice or they will face suspension followed by deactivation. *B.* A new credit control policy has also been implemented at the IEDR which requires domains to be paid for before activation and renewals will have to be paid within 30 days of the date of invoice or they will face suspension followed by deactivation.
(IT's Monday News electronic newsletter, 2nd June 2003)

(A.597) *A.* A Java program consists of **of** several class declarations. *B.* A Java program consists of several class declarations.
(lecture notes, 25/11/03)

(A.598) *A.* The walk was beautiful you walk right along the edge of the beaches and there are amazing views. *B.* The walk was beautiful**.** You walk right along the edge of the beaches and there are amazing views.
(personal email, 19/4/03)

(A.599) *A.* holds a feature structure where skipped token number i occurs as value of attribute i *B.* holds a feature structure where skipped token number i occurs as **the** value of attribute i
(User Manual for the JavaChart 0.1.3 system, p.11, downloaded May 2003)

(A.600) *A.* The crocs **where** very obviously sedated. *B.* The crocs **were** very obviously sedated.
(personal email, 6/10/04)

(A.601) *A.* It is important because a parser needs to identify a sentencial structure which **reflect's** the sentence's meaning, and in order to so, the parser must be able to recognize that an error has occured and know how this error can be corrected. *B.* It is important because a parser needs to identify a sentencial structure which **reflects** the sentence's meaning, and in order to **do** so, the parser must be able to recognize that an error has occured and know how this error can be corrected.
(draft of own writing, 31/8/03)

(A.602) *A.* We can not be held responsible for loose buttons or loose trimmings but in the case of same **been** misplaced we will replace with the closest match. *B.* We can not be held responsible for loose buttons or loose trimmings but in the case of same **being** misplaced we will replace with the closest match.
(Sign in Express Alterations Tailoring Services, Abbey Street, noticed 1st August 2003)

(A.603) *A.* Because counselling in general is not regulated officially here, you can find yourself spending excessive amounts, whether you're depressed, professionally stalled or in a crisis pregnancy from people with no better qualification **that** the ability to recognize vulnerability when they see it. *B.* Because counselling in general is not regulated officially here, you can find yourself spending excessive amounts, whether you're depressed, professionally stalled or in a crisis pregnancy from people with no better qualification **than** the ability to recognize vulnerability when they see it.
(Evening Herald, p.17, 21/8/03)

(A.604) *A.* Occasionally, an **an** untelligible ungrammatical sentence was encountered which couldn't be corrected. *B.* Occasionally, an untelligible ungrammatical sentence was encountered which couldn't be corrected.
(draft of own writing 25/11/03)

(A.605) *A.* In last **weeks** issue of IT's Monday we misquoted the name of the consumer desktop Linux software as Fedora 4.1, when instead it should have read as Fedora Core 1. *B.* In last **week's** issue of IT's Monday we misquoted the name of the consumer desktop Linux software as Fedora 4.1, when instead it should have read as Fedora Core 1.
(IT's Monday News electronic newsletter, 17/11/03)

(A.606) *A.* As normal working hours began for most of the nation this week, the Dail chamber remained **a** silent as a morgue. *B.* As normal working hours began for most of the nation this week, the Dail chamber remained **as** silent as a morgue. *C.* As normal working hours began for most of the nation this week, the Dail chamber remained silent as a morgue.
(Evening Herald, p.4, 29/10/03)

(A.607) *A.* Very few us now stand in a queue to withdraw money from our bank account. *B.* Very few **of** us now stand in a queue to withdraw money from our bank account.
(Evening Herald, p.18, 2/1/04)

(A.608) *A.* The gorgeous model has lined a yet-to-be-announced date in Whelan's of Wexford Street. *B.* The gorgeous model has lined **up** a yet-to-be-announced date in Whelan's of Wexford Street.
(Evening Herald, p.17, 14/10/03)

(A.609) *A.* The information that such grammars lack **is described** a method for obtaining such information is described. *B.* The information that such grammars lack **is described and** a method for obtaining such information is described. *C.* The information that such grammars lack **and** a method for obtaining such information is described.
(draft of own writing, 27/11/03)

(A.610) *A.* His one previous conviction for stealing a car. *B.* His one previous conviction **was** for stealing a car.
(Evening Herald, p.18, 14/10/03)

(A.611) *A.* The Budapest local organisers did a superb job of catering to the **participants** needs and making us feel welcome. *B.* The Budapest local organisers did a superb job of catering to the **participants'** needs and making us feel welcome.
(ACL electronic newsletter, 17/10/03)

(A.612) *A.* All **you** credit to you! *B.* All credit to you!
(http://bigbrother.channel4.com/bigbrother/news/newsstory_t.jsp?id=5372, 5th August 2003)

(A.613) *A.* Thankfully, little Luke made a full recovery but it was six months before could go home. *B.* Thankfully, little Luke made a full recovery but it was six months before **he** could go home. *C.* Thankfully, little Luke made a full recovery but it was six months before **the baby** could go home.
(Evening Herald, p.27, 25/11/03)

(A.614) *A.* We are currently **up dating** our records of our lettings database, if you wish to remain on the list please reply or send a message with update to be kept on that list, and remove as the subject if you wish to be removed from the list. *B.* We are currently **updating** our records of our lettings database, if you wish to remain on the list please reply or send a message with update to be kept on that list, and remove as the subject if you wish to be removed from the list.
(estate agent's electronic newsletter, 10/10/03)

(A.615) *A.* Please note that **a** provisional undergrad class **list are** accessible on the Student Information System. *B.* Please note that a provisional undergrad class list **is** accessible on the Student Information System. *C.* Please note that provisional undergrad class **lists** are accessible on the Student Information System.
(personal email, 16/10/03)

(A.616) *A.* Managers would be wise to get to grips with outsourcing and be prepared to be positive as it seems to becoming an inevitable part of business life. *B.* Managers would be wise to get to grips with outsourcing and be prepared to be positive as it seems to **be** becoming an inevitable part of business life.
(IT's Monday News electronic newsletter, 10/11/03)

(A.617) *A.* The entry is well-formed iff **each feature path** that **occur** in some equation in E has the principal attribute c. *B.* The entry is well-formed iff each feature path that **occurs** in some equation in E has the principal attribute c. *C.* The entry is well-formed iff **all feature paths** that occur in some equation in E have the principal attribute c.
(User Manual for the JavaChart 0.1.3 system, p.3, downloaded May 2003)

(A.618) *A.* Gardai are today hunting a chief suspect for the latest west Dublin murder which claimed the life of a notorious joyrider - who himself got off on a charge of **a** killing a taxi driver. *B.* Gardai are today hunting a chief suspect for the latest west Dublin murder which claimed the life of a notorious joyrider - who himself got off on a charge of killing a taxi driver. *C.* Gardai are today hunting a chief suspect for the latest west Dublin murder which claimed the life of a notorious joyrider - who himself got off on a charge of a killing **of** a taxi driver.
(Evening Herald, p.5, 12/11/03)

(A.619) *A.* Around 25 people are to **loose** their jobs over the next few weeks from the Dublin end of telecommunications giant, Cable and Witness. *B.* Around 25 people are to **lose** their jobs over the next few weeks from the Dublin end of telecommunications giant, Cable and Witness.
(Evening Herald, p.18, 21/10/03)

(A.620) *A.* In fact, when Homage suddenly became popular within **the** Dublin fashion, he responded by printing new issues in smaller runs and placing them where "that certain crew" were unlikely to find them. *B.* In fact, when Homage suddenly became popular within the Dublin fashion **scene**, he responded by printing new issues in smaller runs and placing them where "that certain crew" were unlikely to find them. *C.* In fact, when Homage suddenly became popular within Dublin fashion, he responded by printing new issues in smaller runs and placing them where "that certain crew" were unlikely to find them.
(Irish Times Weekend Magazine, p.17, 24th May 2003)

(A.621) *A.* This method has been **fruitful** exploited in Nas as a method for completing partial parses of ill-formed sentences for which the parser cannot build a unique structure. *B.* This method has been **fruitfully** exploited in Nas as a method for completing partial parses of ill-formed sentences for which the parser cannot build a unique structure. *C.* This method has been exploited **in a** fruitful **way** in Nas as a method for completing partial parses of ill-formed sentences for which the parser cannot build a unique structure.
(Ballim *et al.* 1999, p.13)

(A.622) *A.* Who would have thought **the** Sir Alex Ferguson and Mr Roy Keane would suddenly become the good boys in a confrontation of this sort? *B.* Who would have thought **that** Sir Alex Ferguson

and Mr Roy Keane would suddenly become the good boys in a confrontation of this sort?
(Evening Herald, p.14, 23/9/03)

(A.623) *A.* But the marriage ban ensured that mum, and many women like her never received a **descent** wage. B. But the marriage ban ensured that mum, and many women like her never received a **decent** wage.
(Evening Herald, p.41, 26/9/03)

(A.624) *A.* Joe Brown, country manager for Xerox Ireland, claims the company has seen a revenue increase of 42 per cent for the first half of this year alone and **its** gaining credibility in the public and corporate sectors. *B.* Joe Brown, country manager for Xerox Ireland, claims the company has seen a revenue increase of 42 per cent for the first half of this year alone and **it's** gaining credibility in the public and corporate sectors. *C.* Joe Brown, country manager for Xerox Ireland, claims the company has seen a revenue increase of 42 per cent for the first half of this year alone and is gaining credibility in the public and corporate sectors.
(IT's Monday News electronic newsletter, 14th July 2003)

(A.625) *A.* **Issuers** of negotiating time and privacy for study with children and family members, sharing household responsibilities and maintaining commitment to important obligations will be covered. *B.* **Issues** of negotiating time and privacy for study with children and family members, sharing household responsibilities and maintaining commitment to important obligations will be covered.
(electronic newsletter, 14/10/03)

(A.626) *A.* Reuters **in** now in danger of being kicked off the FTSE index. *B.* Reuters **is** now in danger of being kicked off the FTSE index.
(IT's Monday News electronic newsletter, 28/4/03)

(A.627) *A.* **He's be** quite open to what you'd to do i'd say. *B.* **He'd be** quite open to what you'd **want** to do i'd say.
(personal email, 13/1/04)

(A.628) *A.* This will allow to **make** the robust behaviour almost transparent to the grammar writer. *B.* This will allow the robust behaviour to **be made** almost transparent to the grammar writer. *C.* This will allow **us** to make the robust behaviour almost transparent to the grammar writer.
(Fouvry 2003b, p.134)

(A.629) *A.* inactive edges that **starts** at position 0 and **ends** at position length-of-sentence plus 1 *B.* inactive edges that **start** at position 0 and **end** at position length-of-sentence plus 1
(User Manual for the JavaChart 0.1.3 system, p.8, downloaded May 2003)

(A.630) *A.* With the release of SAS 9.1, which was first mooted earlier this summer and will become generally available towards the end of the year, it will have a much more complete and integrated platform at a technical level, which will **from** the basis of its higher level horizontal and vertical solutions. *B.* With the release of SAS 9.1, which was first mooted earlier this summer and will become generally available towards the end of the year, it will have a much more complete and integrated platform at a technical level, which will **form** the basis of its higher level horizontal and vertical solutions.
(IT's Monday News electronic newsletter, 18/8/03)

(A.631) *A.* While the Public Service Broker will help smaller agencies and **department** wield the buying power of Government and give them access to cheaper technologies, Graham believes the Government should go further and establish a cross-department management system. *B.* While the Public Service Broker will help smaller agencies and **departments** wield the buying power of Government and give them access to cheaper technologies, Graham believes the Government should go further and establish a cross-department management system.
(IT's Monday electronic newsletter, 23rd June 2003)

(A.632) *A.* Heat the oil in a shallow flameproof casserole **a over** medium-high heat. *B.* Heat the oil in a shallow flameproof casserole **over a** medium-high heat.
(Marks & Spencers One Pot Recipe Book, p.52)

(A.633) *A.* Wait until tomorrow before making changes of any kind **if where** home matters are concerned. *B.* Wait until tomorrow before making changes of any kind **where** home matters are concerned. *C.* Wait until tomorrow before making changes of any kind **if** home matters are concerned.
(Evening Herald, 26/9/03, p.26)

(A.634) *A.* Despite being one the most bankable stars, George Clooney has always remained grounded to reality. *B.* Despite being one **of** the most bankable stars, George Clooney has always remained grounded to reality.
(Evening Herald, p.13, 14/10/03)

(A.635) *A.* While **a the** spokesperson did say that Reuters intended to fill Sherry's position in Ireland, it is understood that it has been **a** three weeks since he left Reuters Ireland and so far there has been no announcement as to who his successor might be. *B.* While **a** spokesperson did say that Reuters intended to fill Sherry's position in Ireland, it is understood that it has been three weeks since he left Reuters Ireland and so far there has been no announcement as to who his successor might be. *C.* While **the** spokesperson did say that Reuters intended to fill Sherry's position in Ireland, it is understood that it has been three weeks since he left Reuters Ireland and so far there has been no announcement as to who his successor might be.
(IT's Monday News electronic newsletter, 28/4/03)

(A.636) *A.* Dougan says there is proof of how powerful a united Irish team could be: back in 1973 he was one of the main organisers of a charity game in Dublin, when an all-Ireland side played Brazil, and **giving** the then World Cup holders a run for their money, the Brazilians lucky to get a 4-3 win. *B.* Dougan says there is proof of how powerful a united Irish team could be: back in 1973 he was one of the main organisers of a charity game in Dublin, when an all-Ireland side played Brazil, and **gave** the then World Cup holders a run for their money, the Brazilians lucky to get a 4-3 win. *C.* Dougan says there is proof of how powerful a united Irish team could be: back in 1973 he was one of the main organisers of a charity game in Dublin, when an all-Ireland side played Brazil, and giving the then World Cup holders a run for their money, the Brazilians **were** lucky to get a 4-3 win.
(Evening Herald, p.60, 11/10/03)

(A.637) *A.* A linguistic theory provides an analysis for **the** certain word sequences and should reject as ungrammatical others. *B.* A linguistic theory provides an analysis for certain word sequences and rejects as ungrammatical others.
(draft of own writing, 27/8/03)

(A.638) *A.* The problem with this approach to parsing is that it overly influenced by the notion of grammaticality advocated by the generative linguists. *B.* The problem with this approach to parsing is that it **was** overly influenced by the notion of grammaticality advocated by the generative linguists.
(draft of own writing, 28th July 2003)

(A.639) *A.* There seems to be a lot of **emphasise** on planning rather than trusting in ce sera sera. *B.* There seems to be a lot of **emphasis** on planning rather than trusting in ce sera sera.
(Weddings Irish Style, Vol.1 No.2 Winter 2003, p.30)

(A.640) *A.* The quality of your life is in direct proportion **of** your commitment to excellence. *B.* The quality of your life is in direct proportion **to** your commitment to excellence.
(Evening Herald, p.36, 8/10/03)

(A.641) *A.* Our grandparents believed they would be back in year or so, but that wasn't the case. *B.* Our grandparents believed they would be back in **a** year or so, but that wasn't the case.
(Evening Herald, p.3, 28/11/03)

(A.642) *A.* Ellen **its** not good for pregnant women but can you just spread the word for me. *B.* Ellen **it's** not good for pregnant women but can you just spread the word for me.
(personal email, 29/12/04)

(A.643) *A.* Miriam would kill me **is** I didn't support her charity and you would not want to mess with a mother in-law like her. *B.* Miriam would kill me **if** I didn't support her charity and you would not want to mess with a mother in-law like her.
(Evening Herald, Monday 1/12/03, p.22)

(A.644) *A.* While it may sound unwieldy, Hands pointed out that **that** big bang theory is hard to get right and is very costly even when it is. *B.* While it may sound unwieldy, Hands pointed out that **the** big bang theory is hard to get right and is very costly even when it is.
(IT's Monday News electronic newsletter, 29/9/03)

(A.645) *A.* This class contains attributes that similar to Verb Class. *B.* This class contains attributes that **are** similar to Verb Class.
(student assignment, Feb 2003)

(A.646) *A.* Levelt, Gent, Haans and Meijers **reports that** this is precisely what people do when asked to judge the grammaticality of an isolated sentence. *B.* Levelt, Gent, Haans and Meijers **report** that this is precisely what people do when asked to judge the grammaticality of an isolated sentence. *C.* **According to Levelt**, Gent, Haans and Meijers this is precisely what people do when asked to judge the grammaticality of an isolated sentence.
(draft of own writing, 18/8/03)

(A.647) *A.* We start by considering several hard factors that **is** driving the development of computers in general and distributed systems specifically. *B.* We start by considering several hard factors that **are** driving the development of computers in general and distributed systems specifically.
(personal email, 12/12/03)

(A.648) *A.* Ms. Kelly-Ann Moore, who had known Mr. Carney and who met him by chance on night of Tuesday, January 22, in Dame Street. **B.** Ms. Kelly-Ann Moore, who had known Mr. Carney and who met him by chance on **the** night of Tuesday, January 22, in Dame Street.
(Irish Times, p.6. 8/2/03)

(A.649) *A.* Of course I didn't **wanted** to delete the track. *B.* Of course I didn't **want** to delete the track.
(Sleevenotes, "Senor My Friend", Mumbling Deaf Ro, Feb 2003)

(A.650) *A.* Neither girl pressed the returnee on what he had meant, neither did they discuss the matter further before wiping off the day's make-up and drifting into the **lad** of nod. *B.* Neither girl pressed the returnee on what he had meant, neither did they discuss the matter further before wiping off the day's make-up and drifting into the **land** of nod.
(http://bigbrother.channel4.com/bigbrother/news/newsstory_t.jsp?id=5136, 15/7/03)

(A.651) *A.* When it comes to my personal happiness over my partner's, it **would depends** on the type of relationship I'm involved in. *B.* When it comes to my personal happiness over my partner's, it **would depend** on the type of relationship I'm involved in. *C.* When it comes to my personal happiness over my partner's, it **depends** on the type of relationship I'm involved in.
(Evening Herald, p.17, 11/10/03)

(A.652) *A.* In the rest of the paper we will try to summarize what has been done so far and try to understand which are the most promising approaches to **reach** a valuable robust NLP system. *B.* In the rest of the paper we will try to summarize what has been done so far and try to understand which are the most promising approaches to **reaching** a valuable robust NLP system.
(Ballim *et al.* 1999, p.3)

(A.653) *A.* **Their** is the case of Ronald Reagan. *B.* **There** is the case of Ronald Reagan.
(Evening Herald, p.13, 21/10/03)

(A.654) *A.* The grammars or knowledge bases arising from linguistic theory, **is** generally based on the evidence provided by linguistic intuition. *B.* The grammars or knowledge bases arising from linguistic theory, **are** generally based on the evidence provided by linguistic intuition.
(draft of my own writing, 6/8/03)

(A.655) *A.* Most of the company's recent wins in this **are** lately have been from the Government. *B.* Most of the company's recent wins in this **area** lately have been from the Government.
(IT's Monday electronic newsletter, 23rd June 2003)

(A.656) *A.* After Jon's outburst during yesterday's bell-ringing practice, he could be forgiven for showing an **interested** in the other housemates' private conversations. *B.* After Jon's outburst during yesterday's bell-ringing practice, he could be forgiven for showing an **interest** in the other housemates' private conversations. *C.* After Jon's outburst during yesterday's bell-ringing practice, he could be forgiven for showing an interested **look** in the other housemates' private conversations.
(http://www.channel4.com/bigbrother/news/newsstory_t.jsp?id=1733, 9th June 2003)

(A.657) *A.* These errors specific **for** written text are due to the format in which the language is produced. *B.* These errors specific **to** written text are due to the format in which the language is produced.
(Fouvry 2003b, p.8)

(A.658) *A.* How much information does the parser return, and what is **it's** quality? *B.* How much information does the parser return, and what is **its** quality?
(Musillo and Sima'an 2002)

(A.659) *A.* Right now a romance that will blow your hat off **in** coming your way. *B.* Right now a romance that will blow your hat off **is** coming your way.
(Evening Herald, p.32, 9/10/03)

(A.660) *A.* Himself is the head, **and** which gives rise to an inconsistency in its SPR, imposed by the rule. *B.* Himself is the head, which gives rise to an inconsistency in its SPR, imposed by the rule.
(Fouvry 2003b, p.189)

(A.661) *A.* However, despite such dismal figures, Gartner has given the company a thumbs-up, claiming that decisions not sell its management consulting group, A T Kearney and its software development group, Product Life Cycle Management Solutions and a definitive agreement with Fiserv, which plans to acquire EDS's Credit Union Industry Group, **shows** EDS's direction is taking shape. *B.* However, despite such dismal figures, Gartner has given the company a thumbs-up, claiming that decisions not **to** sell its management consulting group, A T Kearney and its software development group, Product Life Cycle Management Solutions and a definitive agreement with Fiserv, which

plans to acquire EDS's Credit Union Industry Group, **show** EDS's direction is taking shape.
(IT's Monday news electronic newsletter, 28th July 2003)

(A.662) *A.* **It's** fields are the root node, the left subtree and the right subtree. *B.* **Its** fields are the root node, the left subtree and the right subtree.
(student assignment, 2/4/03)

(A.663) *A.* The end result was that **be** became a member of the 'famous five' who were to win eight All-Ireland SFC medals. *B.* The end result was that **he** became a member of the 'famous five' who were to win eight All-Ireland SFC medals.
(Evening Herald, p.69, 8/10/03)

(A.664) *A.* But no matter how much like a media-studies module the programme becomes, it is still hard not to shake the sense that this is little more than an oxygen mask of publicity for Beckham and **her a** floundering career. *B.* But no matter how much like a media-studies module the programme becomes, it is still hard not to shake the sense that this is little more than an oxygen mask of publicity for Beckham and **her** floundering career. *C.* But no matter how much like a media-studies module the programme becomes, it is still hard not to shake the sense that this is little more than an oxygen mask of publicity for Beckham and **a** floundering career.
(Sunday Times Culture Magazine, p.67, 27/4/03)

(A.665) *A.* With traditional formalisms this sentence cannot **analysed** because the unification of 2 and the synsem value of the subject fails. *B.* With traditional formalisms this sentence cannot **be** analysed because the unification of 2 and the synsem value of the subject fails. *C.* With traditional formalisms **we** cannot **analyse** this sentence because the unification of 2 and the synsem value of the subject fails.
(Fouvry 2003b, p.104)

(A.666) *A.* I **as** wondering if we could go to the cinema tomorrow instead. *B.* I **was** wondering if we could go to the cinema tomorrow instead.
(personal email, 17/3/03)

(A.667) *A.* Dublin-based software company Telcotec has taken the brave decision to step out **form** under the wing of major mobile companies and is going it alone with Content Guardian, a filtering programme that blocks undesirable content from mobile phones. B. Dublin-based software company Telcotec has taken the brave decision to step out **from** under the wing of major mobile companies and is going it alone with Content Guardian, a filtering programme that blocks undesirable content from mobile phones.
(IT's Monday News electronic newsletter, 19th May 2003)

(A.668) *A.* If you're interested, please email me and I'll forward **his** you contact details. *B.* If you're interested, please email me and I'll forward you **his** contact details.
(personal email, 27/3/03)

(A.669) *A.* The Business section editor of the Irish Times, Barry O'Keeffe and his former partner Paula Hatton **has** appeared in the Swords District Court when both were charged with breaching a Protection Order issued by the Court on September 10 last. *B.* The Business section editor of the Irish Times, Barry O'Keeffe and his former partner Paula Hatton **have** appeared in the Swords District Court when both were charged with breaching a Protection Order issued by the Court on September 10 last.
(Evening Herald, p.4, 27/11/03)

(A.670) *A.* But this situation **is has** changed, for reasons we will attend to later in the paper. *B.* But this situation **has** changed, for reasons we will attend to later in the paper. *C.* But this situation **is** changed, for reasons we will attend to later in the paper.
(Gazdar 2003, p.3)

(A.671) *A.* What you say is **just** not words, but the feelings they create. *B.* What you say is not **just** words, but the feelings they create.
(Evening Herald, p.26, 23/8/03)

(A.672) *A.* Eventually catching up with her, the Liver lad grabbed Nush in **an** headlock and rubbed his tomato juice streaked hands into her face. *B.* Eventually catching up with her, the Liver lad grabbed Nush in **a** headlock and rubbed his tomato juice streaked hands into her face.
(http://www.channel4.com/bigbrother/news/newsstory_t.jsp?id=4768, 26/6/03)

(A.673) *A.* The range of options is the **employers** decision. *B.* The range of options is the **employers'** decision.
(IT's Monday News electronic newsletter, 1/9/03)

(A.674) *A.* While SSL provides **come** security at present, it is not transport neutral and only provides point-to-point security, not the end-to-end security required for asynchronous communications in a supply chain. *B.* While SSL provides **some** security at present, it is not transport neutral and only provides point-to-point security, not the end-to-end security required for asynchronous communications in a supply chain.
(IT's Monday News electronic newsletter, 18/8/03)

(A.675) *A.* Example 28 seemed to cause **may** Ss trouble.*B.* Example 28 seemed to cause **many** Ss trouble.
(Langendoen *et al.* 1977, p.209)

(A.676) *A.* When there is **there** more than one correction for an ill-formed sentence with each correction expressing the same meaning, all the possible corrections are added to the parallel corpus. *B.* When there is more than one correction for an ill-formed sentence with each correction expressing the same meaning, all the possible corrections are added to the parallel corpus.
(draft of own writing, 30/10/03)

(A.677) *A.* Various researchers have developed methods that attempt **at** transforming the Penn WSJ format into the different output formats of their parsers. *B.* Various researchers have developed methods that attempt *to* transform the Penn WSJ format into the different output formats of their parsers. C. Various researchers have developed methods that **make an** attempt at transforming the Penn WSJ format into the different output formats of their parsers.
(Musillo and Sima'an 2002)

(A.678) *A.* According to her, this is "not acceptable, as everyone knows a sentence is **continue**, not started, with 'and"'. B. According to her, this is "not acceptable, as everyone knows a sentence is **continued**, not started, with 'and"'.
(transcription error in my writing, original sentence is from the Evening Herald, 25/8/03)

(A.679) *A.* The second **sentence** contains grammatical versions of the sentences in the first corpus. *B.* The second corpus contains grammatical versions of the sentences in the first corpus.
(draft of own writing, 21/11/03)

(A.680) *A.* It is important to observe how the syntactic module passes information to the semantic module through two **mechanism**. *B.* It is important to observe how the syntactic module passes information to the semantic module through two **mechanisms**.
(Sima'an 2003, p.11-12)

(A.681) *A.* At a time when there seems to a neverending supply of horrifying statistics about our pernicious love affair with alcohol this survey is truly worrying. *B.* At a time when there seems to **be** a neverending supply of horrifying statistics about our pernicious love affair with alcohol this survey is truly worrying.
(Evening Herald, p.14, 25/9/03)

(A.682) *A.* Just to let you know that we are **plying** in Gleeson's this friday. *B.* Just to let you know that we are **playing** in Gleeson's this friday.
(Turn electronic newsletter, 22/10/03)

(A.683) *A.* The Department of Foreign Affairs is investigating reports that an Irish woman may have killed in the suicide bombings at three Westerners' compounds in the Saudi Arabian capital. *B.* The Department of Foreign Affairs is investigating reports that an Irish woman may have **been** killed in the suicide bombings at three Westerners' compounds in the Saudi Arabian capital.
(http://www.ireland.com/newspaper/breaking/2003/0514/index.htm, 14/05/03)

(A.684) *A.* Well, it looks like james won't be able to go so he's a no , so i mulled it over and found that i'd still quite like to go and **hat** basically my main problem is money , so then i decided if that's my only problem, i may as well go. *B.* Well, it looks like james won't be able to go so he's a no , so i mulled it over and found that i'd still quite like to go and **that** basically my main problem is money , so then i decided if that's my only problem, i may as well go.
(personal email, 17/4/03)

(A.685) *A.* That was the signal for Wood to limp off, his job well done, to be replaced by Shane Byrne, while before the end Ronan O'Gara, Guy Easterby, Donncha O'Callaghan and John Kelly also got a run and Hickie **grabbed got** another try, which O'Gara goaled before Maftei scored a Romanian try, which Andrei goaled. *B.* That was the signal for Wood to limp off, his job well done, to be replaced by Shane Byrne, while before the end Ronan O'Gara, Guy Easterby, Donncha O'Callaghan and John Kelly also got a run and Hickie **grabbed** another try, which O'Gara goaled before Maftei scored a Romanian try, which Andrei goaled. *C.* That was the signal for Wood to limp off, his job well done, to be replaced by Shane Byrne, while before the end Ronan O'Gara, Guy Easterby, Donncha O'Callaghan and John Kelly also got a run and Hickie **got** another try, which O'Gara goaled before Maftei scored a Romanian try, which Andrei goaled.
(Evening Herald, p.64, 11/10/03)

(A.686) *A.* In all but one of these cases, a determiner needed to be before a singular noun. *B.* In all but one of these cases, a determiner needed to be **added** before a singular noun.
(draft of own writing, 23/10/03)

(A.687) *A.* If **all instantiation are** rejected, then the input string W is deemed ungrammatical. *B.* If all **instantiations** are rejected, then the input string W is deemed ungrammatical. *C.* If **each** instantiation **is** rejected, then the input string W is deemed ungrammatical.
(Osbourne and Bridge 1994)

(A.688) *A.* His next insult **as** to call me a Republican. *B.* His next insult **was** to call me a Republican.
(transcription error in my own writing, original sentence from Matthews (1993, p.38))

(A.689) *A.* After numerous attempts **of** trying to track down the following people, I still have had no luck. *B.* After numerous attempts **at** trying to track down the following people, I still have had no luck.
(personal email, 14/10/03)

(A.690) *A.* We trekked **though** rice fields , camped overnight by a huge waterfall, ate brilliant Thai food cooked by our guide and had great craic singing songs around the campfire. *B.* We trekked **through** rice fields , camped overnight by a huge waterfall, ate brilliant Thai food cooked by our guide and had great craic singing songs around the campfire.
(personal email, 13/10/03)

(A.691) *A.* It might be interesting to investigate to what extent missing word error rules **provides** a treatment for the grammatical phenomenon of ellipsis. *B.* It might be interesting to investigate to what extent missing word error rules **provide** a treatment for the grammatical phenomenon of ellipsis.
(draft of own writing, 5/9/03)

(A.692) *A.* Presiding magistrate Christine Simon Rosenthal who has already twice adjourned the case since May, granted requests by the O'Keeffe family, **the** and families of **21-year old** postal worker Patrice Dennis and 18-year-old Patrick Gache to press ahead with the trial and swore in a jury. *B.* Presiding magistrate Christine Simon Rosenthal who has already twice adjourned the case since May, granted requests by the O'Keeffe family, and the families of **21-year-old** postal worker Patrice Dennis and 18-year-old Patrick Gache to press ahead with the trial and swore in a jury.
(Evening Herald, p.4, 14/10/03)

(A.693) *A.* But it is widely accepted as axiomatic that a parser will operate by **resting** an input in some systematic manner against one after another of the structures provided by a grammar. *B.* But it is widely accepted as axiomatic that a parser will operate by **testing** an input in some systematic manner against one after another of the structures provided by a grammar.
(Sampson 1987)

(A.694) *A.* Some of the above scoring parameters score an edge according what sort of parses it could contribute to. *B.* Some of the above scoring parameters score an edge according **to** what sort of parses it could contribute to.
(Mellish 1989)

(A.695) *A.* The fourth quarter for Lucent turned out to be its first **in profitable one** since March 2000. *B.* The fourth quarter for Lucent turned out to be its first **profitable one** since March 2000. *C.* The fourth quarter for Lucent turned out to be its first **one in profit** since March 2000.
(IT's Monday News electronic newsletter, 27/10/03)

(A.696) *A.* The postgrad who is bikeless would like to where he can get a cheap, second-hand bike? *B.* The postgrad who is bikeless would like to **know** where he can get a cheap, second-hand bike?
(personal email, 7/7/03)

(A.697) *A.* A phone call will reveal more if you wish, call me. *B.* A phone call will reveal more**.** If you wish, call me.
(Evening Herald, p.36, 8/10/03)

(A.698) *A.* The competition is seeking games ideas that can be developed for mobile phones with the winner receiving a 5,000 advance on future earnings from the published game and earning an internship **at** a games developer. *B.* The competition is seeking games ideas that can be developed for mobile phones with the winner receiving a 5,000 advance on future earnings from the published game and earning an internship **as** a games developer.
(IT's Monday News electronic newsletter, 26th May 2003)

(A.699) *A.* Expect a much more positive state of affairs from tomorrow onwards**.** If you've been out of luck recently. *B.* Expect a much more positive state of affairs from tomorrow onwards if you've been out of luck recently.
(Evening Herald, p.26, 26/9/03)

(A.700) *A.* Detectives were today investigating the worrying incident when **when** northside criminals tried to attack the driver and passenger of a four wheel drive. *B.* Detectives were today investigating the worrying incident when northside criminals tried to attack the driver and passenger of a four wheel drive.
(Evening Herald, p. 16, 13/8/03)

(A.701) *A.* **According to** Gleitman et al 1972 call **call** grammaticality judgements 'very derived and rather artificial psycholinguistic phenomena which develop late in language acquisition'. *B.* Gleitman et al 1972 call grammaticality judgements 'very derived and rather artificial psycholinguistic phenomena which develop late in language acquisition'. *C.* According to Gleitman et al 1972 grammaticality judgements **are** 'very derived and rather artifical psycholinguistic phenomena which develop late in language acquisition'.
(my written notes from 15/8/03)

(A.702) *A.* The term expansion definition makes use of a relation that takes the LHS and RHS of the as input and which instantiates three lists. *B.* The term expansion definition makes use of a relation that takes the LHS and RHS of the **rule** as input and which instantiates three lists.
(draft of "Enabling Left Recursive DCGs" by Carl Vogel, March 2003)

(A.703) *A.* Furthermore, constraints can be annotated by their degree of vulnerability, hence allowing to include aspects of error anticipation into the relaxation framework.*B.* Furthermore, constraints can be annotated by their degree of vulnerability, hence allowing **us** to include aspects of error anticipation into the relaxation framework. *C.* Furthermore, constraints can be annotated by their degree of vulnerability, hence allowing aspects of error anticipation **to be included** into the relaxation framework.
(Menzel 1995)

(A.704) *A.* However much **efforts** has been spent in computing sentence meaning in a compositional way.*B.* However much **effort** has been spent in computing sentence meaning in a compositional way.
(Ballim *et al.* 1999, p.12)

(A.705) *A.* Because for the rest **in** English the orthography has rather little interaction with the syntax, the value is duplicated in a syntactic feature specific to prepositions, called PFORM. *B.* Because for the rest **of** English the orthography has rather little interaction with the syntax, the value is duplicated in a syntactic feature specific to prepositions, called PFORM.
(Fouvry 2003b, p.89)

(A.706) *A.* Logical Apps is claiming **it** most recent product is a cure for the problems Oracle customers are experiencing with Oracle financial and enterprise resource planning applications. *B.* Logical Apps is claiming **its** most recent product is a cure for the problems Oracle customers are experiencing with Oracle financial and enterprise resource planning applications.
(IT's Monday News eletronic newsletter, 12th May 2003)

(A.707) *A.* The ovular nodes on the right end of each edge represent activation **level**. *B.* The ovular nodes on the right end of each edge represent activation **levels**. *C.* The ovular nodes on the right end of each edge represent **an** activation level.
(Kim 1994, p.314)

(A.708) *A.* But found that **found** Dublin's most exclusive social event was no better than a boozy brawl. *B.* But found that Dublin's most exclusive social event was no better than a boozy brawl.
(The Slate, p.4, Issue No.25, June 2003, 16th June 2003)

(A.709) *A.* We found the **skip** of normal rules is very problematic. *B.* We found the **skipping** of normal rules is very problematic.
(Wang 2001)

(A.710) *A.* Even though e_list, ne_list and synsem_list are all three **are** sister types, the relation between e_list and ne_list seems to be of a different nature than between synsemlist and ne_list. *B.* Even though e_list, ne_list and synsem_list are all three sister types, the relation between e_list and ne_list seems to be of a different nature than between synsemlist and ne_list.
(Fouvry 2003b, p.176)

(A.711) *A.* vivacious young Canadian girls fresh a previous summer in HKB. *B.* vivacious young Canadian girls fresh **from** a previous summer in HK
(personal email, 2/10/02)

(A.712) *A.* The accommodation **comprises of** 1 bedroom, 1 reception, fully equipped kitchen. *B.* The accommodation **comprises** 1 bedroom, 1 reception, fully equipped kitchen. *C.* The accommodation **consists of** 1 bedroom, 1 reception, fully equipped kitchen.
(estate agent's electronic newsletter, 9/10/03)

(A.713) *A.* **He** average contribution was generally higher for these accounts, because of higher minimum contributions and restrictions that meant that they appealed to more affluent customers. *B.* **The** average contribution was generally higher for these accounts, because of higher minimum contributions and restrictions that meant that they appealed to more affluent customers.
(Your Money supplement, p.3, Irish Independent, 11/9/03)

(A.714) *A.* Typically the ambiguity is taken **in** account at a higher linguistic level. *B.* Typically the ambiguity is taken **into** account at a higher linguistic level.
(Ballim *et al.* 1999, p.6)

(A.715) *A.* Windows 95 users will see a greater benefit by moving to Linux **that** Windows 2000 or XP users. *B.* Windows 95 users will see a greater benefit by moving to Linux **than** Windows 2000 or XP users.
(IT's Monday News electronic newsletter, 22/9/03)

(A.716) *A.* Using Picky's grammar, these test sentences generate, on average, over 100 parses per sentence, **with** some sentences **generated** over 1,000 parses. *B.* Using Picky's grammar, these test sentences generate, on average, over 100 parses per sentence, with some sentences **generating** over 1,000 parses. *C.* Using Picky's grammar, these test sentences generate, on average, over 100 parses per sentence, **while** some sentences generated over 1,000 parses.
(Magerman and Weir 1992, p.44)

(A.717) *A.* One is that the results of part-of-speech taggers are poor when considered at sentence level rather **that** word level. *B.* One is that the results of part-of-speech taggers are poor when considered at sentence level rather **than** word level.
(Chanod and Tapanainen 1996, p.20)

(A.718) *A.* Probably true that bands with a buzz around **then** do get booked but then that's probably because they have a buzz as a result of being good and so the demo they submit is worth accepting. *B.* Probably true that bands with a buzz around **them** do get booked but then that's probably because they have a buzz as a result of being good and so the demo they submit is worth accepting.
(http://www.inthecity.co.uk/itc-board/itc-unsigned-mb.shtml, 19/8/03)

(A.719) *A.* Bush attempts **breath** life into Middle East peace plan. *B.* Bush attempts **to breathe** life into Middle East peace plan. *C.* Bush attempts breath **of** life into Middle East peace plan.
(http://www.ireland.com, 4th June 2003)

(A.720) *A.* In particular Dublin is resented every cent spent on it and it was only when this giant golden egg was threatening to seize up entirely did it get the Luas and the port tunnel and several other inadequate decades-in-the-waiting improvements. *B.* In particular Dublin is resented **for** every cent spent on it and it was only when this giant golden egg was threatening to seize up entirely did it get the Luas and the port tunnel and several other inadequate decades-in-the-waiting improvements.
(Evening Herald, p.13, 10/12/03)

(A.721) *A.* It's common-sense when you consider that the rate **your** getting is effectively increasing all the time and we are now over two years in. *B.* It's common-sense when you consider that the rate **you're** getting is effectively increasing all the time and we are now over two years in.
(Your Money supplement, p.1, Irish Independent, 11/9/03)

(A.722) *A.* The performer who in his early days was more **like** to spit on audiences than lecture them from a stool is giving the second lecture in a series kicked off by philosopher John Moriarty in the Shelbourne last week. *B.* The performer who in his early days was more **likely** to spit on audiences than lecture them from a stool is giving the second lecture in a series kicked off by philosopher John Moriarty in the Shelbourne last week.
(Evening Herald, p. 21, 13/8/03)

(A.723) *A.* Look, I'm trying to write in the **the** voice of a three-foot-tall alien puppet. *B.* Look, I'm trying to write in the voice of a three-foot-tall alien puppet.
(Evening Herald, p.29, 12/11/03)

(A.724) *A.* The decision to classify a sentence as ungrammatical will depend on whether or not the corpus compiler thinks that an error has occured, and as such is dependent on the linguistic intuition of the corpus **compilation.** *B.* The decision to classify a sentence as ungrammatical will depend on whether or not the corpus compiler thinks that an error has occured, and as such is dependent on the linguistic intuition of the corpus **compiler.**
(draft of own writing, 31/8/03)

(A.725) *A.* Eriksson's side will play France and Croatia in the 65,000 capacity new Stadium of Light in Lisbon either side of short trip north to Coimbra to face Switzerland. *B.* Eriksson's side will play France and Croatia in the 65,000 capacity new Stadium of Light in Lisbon either side of **a** short

trip north to Coimbra to face Switzerland. C. Eriksson's side will play France and Croatia in the 65,000 capacity new Stadium of Light in Lisbon either side of **the** short trip north to Coimbra to face Switzerland.
(Evening Herald, Monday 1/12/03, p.88)

(A.726) *A.* **Lets** try to arrange that. *B.* **Let's** try to arrange that.
(personal email, January 2004)

(A.727) *A.* If this is so, it returns **an** the root node value, an opening bracket, the value of the root node of the right subtree and finally a closing bracket. *B.* If this is so, it returns the root node value, an opening bracket, the value of the root node of the right subtree and finally a closing bracket.
(student assignment, 2/4/03)

(A.728) *A.* If **the** that was the unadorned cake, then the icing was slathered on in 2000 when he won an award from Hillary Clinton for his business achievements. *B.* If that was the unadorned cake, then the icing was slathered on in 2000 when he won an award from Hillary Clinton for his business achievements.
(Evening Herald, p.3, 19/8/03)

(A.729) *A.* **It's** annual results have provided scary reading for shareholders of late. *B.* **Its** annual results have provided scary reading for shareholders of late.
(IT's Monday News eletronic newsletter, 7th April 2003)

(A.730) *A.* Please remember that **the** it is not safe to send confidential data around the network. *B.* Please remember that it is not safe to send confidential data around the network. *C.* Please remember that the **network** is not safe to send confidential data **around**.
(personal email, 14/4/03)

(A.731) *A.* But if you don't want to **got** the cinema, do you want to meet for coffee? *B.* But if you don't want to **go to** the cinema, do you want to meet for coffee?
(personal email, 23/9/03)

(A.732) *A.* In this case, the utterance made **be** A doesn't lead to an understanding of B's response. *B.* In this case, the utterance made **by** A doesn't lead to an understanding of B's response.
(draft of own writing, 18/8/03)

(A.733) *A.* The parser is a bottom-up probabilistic chart parser which finds the parse tree with the best score by best-first search algorithm. *B.* The parser is a bottom-up probabilistic chart parser which finds the parse tree with the best score by **a** best-first search algorithm. *C.* The parser is a bottom-up probabilistic chart parser which finds the parse tree with the best score by **using a** best-first search algorithm.
(http://nlp.cs.nyu.edu/app/, March 2003)

(A.734) *A.* Scott and Nush spent a good deal time chatting away to each other, but definitely as friends. *B.* Scott and Nush spent a good deal **of** time chatting away to each other, but definitely as friends.
(http://bigbrother.channel4.com/bigbrother/news/newsstory_t.jsp?id=5366, 30th July 2003)

(A.735) *A.* Statistical methods are not able to extract useful probabilities from **modesty** sized corpora. *B.* Statistical methods are not able to extract useful probabilities from **modestly** sized corpora.
(Ballim *et al.* 1999, p.10)

(A.736) *A.* A mini-lexicon entry contains a suffix category, possibly a **pointers** to a subsequent suffix category, and data about the suffix as a feature structure. *B.* A mini-lexicon entry contains a suffix category, possibly a **few** pointers to a subsequent suffix category, and data about the suffix as a feature structure. *C.* A mini-lexicon entry contains a suffix category, possibly **pointers** to a subsequent suffix category, and data about the suffix as a feature structure.
(User Manual for the JavaChart 0.1.3 system, p.13, downloaded May 2003)

(A.737) *A.* A basic proposal along the same lines, but in the context HPSG and also for the purposes of Computer-Aided Language Learning was proposed in Heift. *B.* A basic proposal along the same lines, but in the context **of** HPSG and also for the purposes of Computer-Aided Language Learning was proposed in Heift.
(Fouvry 2003b, p.206)

(A.738) *A.* Have **on** idea where **johns** place is though. *B.* Have **no** idea where **john's** place is though.
(personal text message, 22/11/03)

(A.739) *A.* It's all pretty flexible but the talks can be divided into the following **group**. *B.* It's all pretty flexible but the talks can be divided into the following **groups**.
(personal email, 23/10/03)

(A.740) *A.* In my book **its** quality that counts. *B.* In my book **it's** quality that counts.
(Evening Herald, 14/10/03, p.22)

(A.741) *A.* In the end, the album was professionally mastered and many, though not all, of clumsier decisions corrected. *B.* In the end, the album was professionally mastered and many, though not all, of **the** clumsier decisions corrected.
(Sleevenotes, "Senor My Friend", Mumbling Deaf Ro, Feb 2003)

(A.742) *A.* They reassured them that **it was they** came back because they truly wanted to. *B.* They reassured them that **they** came back because they truly wanted to. *C.* They reassured them that **it was** because they truly wanted to.
(personal email, 15/10/02)

(A.743) *A.* This means **to allow** structure-sharing and links from terminal nodes. *B.* This means structure-sharing and links from terminal nodes **are allowed**.
(Mengel and Lezius 2000)

(A.744) *A.* If there is no parameter list, it might get more **that** one random number which is exception. *B.* If there is no parameter list, it might get more **than** one random number which is **an** exception.
(student assignment, Feb 2003)

(A.745) *A.* I **awaits** your response. *B.* I **await** your response.
(header to a spam mail, 21/2/03)

(A.746) *A.* Mummy the **others boys** won't talk to me. *B.* Mummy the **other** boys won't talk to me. *C.* Mummy the **others** won't talk to me.
(Sunday Independent Living, p.2, 10/8/03)

(A.747) *A.* In a frank statement today, Deputy Wright admitted that subsequent tests showed **the** he exceeded the breath alcohol limit. *B.* In a frank statement today, Deputy Wright admitted that subsequent tests showed **that** he exceeded the breath alcohol limit. *C.* In a frank statement today, Deputy Wright admitted that subsequent tests showed he exceeded the breath alcohol limit.
(Evening Herald, p.1, 25/9/03)

(A.748) *A.* In practice, we actually subdivide the lexicon further **in** lexicon and mini lexicon. *B.* In practice, we actually subdivide the lexicon further **into** lexicon and mini lexicon.
(User Manual for the JavaChart 0.1.3 system, p.3, downloaded May 2003)

(A.749) *A.* The old couch we **has** now has **it's** own web page recording all its press coverage. *B.* The old couch we **had** now has **its** own web page recording all its press coverage.
(personal email, 16/9/03)

(A.750) *A.* **Its** got a killer product in BCM and its new strategy, Engaged Solutions, has already got the large corporates sniffing around. *B.* **It's** got a killer product in BCM and its new strategy, Engaged Solutions, has already got the large corporates sniffing around.
(IT's Monday News eletronic newsletter, 7/4/03)

(A.751) *A.* Instead of imposing a normative description on the input data it takes them as starting point and tries to find a plausible interpretation for them.*B.* Instead of imposing a normative description on the input data it takes them as **a** starting point and tries to find a plausible interpretation for them. *C.* Instead of imposing a normative description on the input data it takes them as **the** starting point and tries to find a plausible interpretation for them.
(Menzel 1995)

(A.752) *A.* Yeah Wednesday is **find** with me. *B.* Yeah Wednesday is **fine** with me.
(text message, 9/12/03)

(A.753) *A.* This fault has **not** been rectified and future mailshots will have this functionality. *B.* This fault has **now** been rectified and future mailshots will have this functionality.
(mail from estage agents, 10/1/04)

(A.754) *A.* Part of this meaning is the structure or grammar or episememe which is the aspect of the sentence which must be **judgement.** *B.* Part of this meaning is the structure or grammar or episememe which is the aspect of the sentence which must be **judged.**
(a draft of own writing, 15/8/03)

(A.755) *A.* For more details **of** any of the above chaplaincy services and events, visit our website. *B.* For more details **on** any of the above chaplaincy services and events, visit our website.
(electronic newsletter, 30/4/03)

(A.756) *A.* The Government is keen extend fibre to more than one hundred more towns on a shared risk basis. *B.* The Government is keen **to** extend fibre to more than one hundred more towns on a shared risk basis.
(IT's Monday News electronic newsletter, 8/12/03)

(A.757) *A.* I am at a loss to know what exactly all the ballyhoo **surround** the recent wedding of one Georgina Ahern to her guitar-strumming boyfriend was meant to convey to us readers. *B.* I am at a loss to know what exactly all the ballyhoo **surrounding** the recent wedding of one Georgina Ahern to her guitar-strumming boyfriend was meant to convey to us readers. *C.* I am at a loss to know what exactly all the ballyhoo **that surrounded** the recent wedding of one Georgina Ahern to her guitar-strumming boyfriend was meant to convey to us readers.
(Evening Herald, p.15, 21/8/03)

(A.758) *A.* Despite the difficulties faced by the company following a sustained period of reduced expenditure on telecoms infrastructure, Nortel in Galway has managed to position itself close to the heart of the **organisations** long term technology strategy while generating significant local business for the technology it has helped develop. *B.* Despite the difficulties faced by the company following a sustained period of reduced expenditure on telecoms infrastructure, Nortel in Galway has managed to position itself close to the heart of the **organisation's** long term technology strategy while generating significant local business for the technology it has helped develop.
(IT's Monday News electronic newsletter, 1/12/03)

(A.759) *A.* Being so gifted can find fulfillment in many employments. *B.* Being so gifted **they** can find fulfillment in many employments. *C.* Being so gifted **Scorpians** can find fulfillment in many employments.
(http://www.crystalinks.com/astrology3.html, Feb 2003)

(A.760) *A.* While Mortell was reluctant to give an exact figure for revenue he did say that combined new contracts in the last **mine** months had brought in an additional EUR 1.5 million. *B.* While Mortell was reluctant to give an exact figure for revenue he did say that combined new contracts in the last **nine** months had brought in an additional EUR 1.5 million.
(IT's Monday News electronic newsletter, 13/12/03)

(A.761) *A.* We use a variety of null elements to show **show** how non-adjacent constituents are related. *B.* We use a variety of null elements to show how non-adjacent constituents are related.
(Marcus *et al.* 1994)

(A.762) *A.* Spoken language is not only characterized by hesitations, breaks and restarts but is additionally **narked** by different kinds of syntactic and semantic inconsistencies. *B.* Spoken language is not only characterized by hesitations, breaks and restarts but is additionally **marked** by different kinds of syntactic and semantic inconsistencies.
(Menzel 1996)

(A.763) *A.* After all, the catch statement contains code to potential expectations. *B.* After all, the catch statement contains code to **catch** potential expectations.
(student assignment, Feb 2003)

(A.764) *A.* **Digger** a deeper grave for himself, the exchange housemate said: "I don't mean it in a bad way, but you eat a lot." *B.* **Digging** a deeper grave for himself, the exchange housemate said: "I don't mean it in a bad way, but you eat a lot."
(http://www.channel4.com/bigbrother/news/newsstory_t.jsp?id=4079, 24/6/03)

(A.765) *A.* If a sentence is encountered where it isn't obvious that an error has actually occurred, the sentence is **no** included as evidence. *B.* If a sentence is encountered where it isn't obvious that an error has actually occurred, the sentence is **not** included as evidence.
(draft of own writing, 11/12/03)

(A.766) *A.* They can be compiled **compiled** into finite state transducers. *B.* They can be compiled into finite state transducers.
(overhead slides at talk by Lauri Kartunnen at Dublin Computational Linguistics Research Seminar (DCLRS) 10/10/03)

(A.767) *A.* Moving from the domain of linguistic theory to the more practical domain of natural language processing such strings are less useful because they have usually been made up by linguists in order to test their theory and they do not **necessary** reflect the kind of ungrammaticality encountered in real language use. *B.* Moving from the domain of linguistic theory to the more practical domain of natural language processing such strings are less useful because they have usually been made up by linguists in order to test their theory and they do not **necessarily** reflect the kind of ungrammaticality encountered in real language use.
(draft of own writing, 31/8/03)

(A.768) *A.* Try to make up your mind about pressing decisions, your charm will **amazing**, so don't waste your time, go for it. *B.* Try to make up your mind about pressing decisions, your charm will **be** amazing, so don't waste your time, go for it. *C.* Try to make up your mind about pressing decisions, your charm will **amaze**, so don't waste your time, go for it.
(Evening Herald, p.36, 8/10/03)

(A.769) *A.* David O' Leary probably has more strikers to **chose** from **that** any other manager in the Premiership. *B.* David O' Leary probably has more strikers to **choose** from **than** any other manager in the Premiership.
(Evening Herald, p. 71, 13/8/03)

(A.770) *A.* It remains to be shown that the unification of two feature **structure** is itself a feature structure. *B.* It remains to be shown that the unification of two feature **structures** is itself a feature structure.
(Fouvry 2003a, p.109)

(A.771) *A.* In example given, the distal context set up by the words barge and towpath indicates that it is the second sense which is relevant to the interpretation of the final sentence. *B.* In **the** example given, the distal context set up by the words barge and towpath indicates that it is the second sense which is relevant to the interpretation of the final sentence.
(Gazdar 2003, p.7-8)

(A.772) *A.* Eric Lagier, director business development in Oracle's Wireless & Voice Division says there are three areas of business where the benefits are greatest. *B.* Eric Lagier, director **of** business development in Oracle's Wireless & Voice Division says there are three areas of business where the benefits are greatest.
(IT's Monday News electronic newsletter, 8/9/03)

(A.773) *A.* They will NOT be playing despite the rumours and there more bands added next week. *B.* They will NOT be playing despite the rumours and there **will be** more bands added next week. *C.* They will NOT be playing despite the rumours and there **are to be** more bands added next week.
(personal email, 17/4/03)

(A.774) *A.* **Our** do we need to use context-free rules or even context-sensitive rules? *B.* **Or** do we need to use context-free rules or even context-sensitive rules?
(Jurafsky and Martin 2000, p.481)

(A.775) *A.* This seemingly innocent analogy between the derivation of a grammatical syntactic structure and the construction of a proof in a metamathematical framework became one of the characteristic features TGG-style syntactic theory.*B.* This seemingly innocent analogy between the derivation of a grammatical syntactic structure and the construction of a proof in a metamathematical framework became one of the characteristic features **of** TGG-style syntactic theory.
(Tomalin 2002)

(A.776) *A.* Detectives are aware of the first name of the individual who sold the car and today questioned the driver about events leading up to the fatal crash, and how **contacted** the second-hand dealer was contacted. *B.* Detectives are aware of the first name of the individual who sold the car and today questioned the driver about events leading up to the fatal crash, and how the second-hand dealer was contacted.
(Evening Herald, p.4, 21/10/03)

(A.777) *A.* The **apartments** is located to the rear of the complex, and has secure communal. *B.* The **apartment** is located to the rear of the complex, and has secure communal **parking**.
(estate agent's electronic newsletter, 10/10/03)

(A.778) *A.* This method contains a for loop which initializes a variable of type **into** to zero. *B.* This method contains a for loop which initializes a variable of type **int** to zero.
(student assignment, Feb 2003)

(A.779) *A.* Dolores steals on Fair City to try pay back her credit card debts. *B.* Dolores steals on Fair City to try **to** pay back her credit card debts. *C.* Dolores steals on Fair City to try **and** pay back her credit card debts.
(Evening Herald, p.5, 25/11/03)

(A.780) *A.* Change the condition in each of the above 3 loops **to** that they use <= rather than <. *B.* Change the condition in each of the above 3 loops **so** that they use <= rather than <.
(lecture notes, November 2003)

(A.781) *A.* PalmSource, the provider of the operating system for handheld devices and smart phones that is the midst of being spun off as a separate company, said it formed a strategic alliance with PalmGear, a resource for freeware, shareware and commercial software for the Palm OS platform. *B.* PalmSource, the provider of the operating system for handheld devices and smart phones that is **in** the midst of being spun off as a separate company, said it formed a strategic alliance with PalmGear, a resource for freeware, shareware and commercial software for the Palm OS platform.
(IT's Monday News electronic newsletter, 8/9/03)

(A.782) *A*. But it was definitely worth it we got up quite far which meant we could walk through a big crevice of ice. *B*. But it was definitely worth it**.** We got up quite far which meant we could walk through a big crevice of ice.
(personal email, 19/4/03)

(A.783) *A*. If the members of the set constituting the domain represent contingently real objects, then we say the theory **in** empirical. **B.** If the members of the set constituting the domain represent contingently real objects, then we say the theory **is** empirical.
(Perry 1979b, p.4)

(A.784) *A*. Using people to parse sentences in a corpus is so much better than using a computer because people are able use their understanding of language to produce one correct parse of a sentence, whereas computers are unable to distinguish between possible parses. *B*. Using people to parse sentences in a corpus is so much better than using a computer because people are able **to** use their understanding of language to produce one correct parse of a sentence, whereas computers are unable to distinguish between possible parses.
(draft of own writing, 27/8/03)

(A.785) *A*. King estimates the BPM application licensing arena to worth around STG 550 million currently, while the Aberdeen Group expects application licensing spending to grow to USD1.56 billion in 2005. *B*. King estimates the BPM application licensing arena to **be** worth around STG 550 million currently, while the Aberdeen Group expects application licensing spending to grow to USD1.56 billion in 2005.
(IT's Monday News electronic newsletter, 1/12/03)

(A.786) *A*. Oracle's new E-Business suite **it** targeted at SMEs and sees Oracle going after the same business as its biggest rival, SAP. *B*. Oracle's new E-Business suite **is** targeted at SMEs and sees Oracle going after the same business as its biggest rival, SAP.
(IT's Monday News electronic newsletter, 8/9/03)

(A.787) *A*. **Agent architecture have** been taken as a good approximating model of human mind where simple concurrent and communicating cognitive processes are able to "implement"the human thought process.*B*. Agent architecture **has** been taken as a good approximating model of **the** human mind where simple concurrent and communicating cognitive processes are able to "implement" the human thought process. *C*. **Agent architectures** have been taken as a good approximating model of human **minds** where simple concurrent and communicating cognitive processes are able to "implement" the human thought process.
(Ballim *et al.* 1999, p.11)

(A.788) *A*. Cable firm NTL plans to cut off 15,000 customers in the next six months as part of a campaign to crackdown on subscribers who fail to pay **to** time. *B*. Cable firm NTL plans to cut off 15,000 customers in the next six months as part of a campaign to crackdown on subscribers who fail to pay **on** time.
(Evening Herald, p. 18, 13/8/03)

(A.789) *A*. I was questioned by mary and louise what happened and I said nothing. *B*. I was questioned by mary and louise **about** what happened and I said nothing.
(personal email, 2/11/02)

(A.790) *A*. Thank you to those who **send** me pictures. *B*. Thank you to those who **sent** me pictures.
(personal email, 11/9/03)

(A.791) *A*. It is also obvious that many further rules remain be written. *B*. It is also obvious that many further rules remain **to** be written.
(Roche and Schabes 1997, p.303)

(A.792) *A*. The survey also found that in families where both parents work, parents will spend up to EUR 188 a **a** week per child on food, clothing and childcare. *B*. The survey also found that in families where both parents work, parents will spend up to EUR 188 a week per child on food, clothing and childcare.
(Evening Herald, p.46,28/11/03)

(A.793) *A*. Each subject judged one **lexicalizations** and 24 fillers.*B*. Each subject judged one **lexicalization** and 24 fillers.
(Keller and Asudeh 2000)

(A.794) *A*. Money will be a major cause of concern in the days ahead so **holds** on to what you have for dear life. *B*. Money will be a major cause of concern in the days ahead so **hold** on to what you have for dear life.
(Evening Herald, 10/10/03, p.26)

(A.795) *A.* Finally admitting defeat, Nush then flattened his hair down into **it's** original style. *B.* Finally admitting defeat, Nush then flattened his hair down into **its** original style.
(http://www.channel4.co.uk/bigbrother/news/newsstory_t.jsp?id=1236, 26/5/03)

(A.796) *A.* **Both interfaces takes an** parse sentence as their main input, but have other parameters as well. *B.* Both interfaces **take a** parse sentence as their main input, but have other parameters as well. *C.* **Each interface** takes **a** parse sentence as its main input, but has other parameters as well.
(User Manual for the JavaChart 0.1.3 system, p.5, downloaded May 2003)

(A.797) *A.* From all of the above considerations, the following roadmap for the has been derived **derived** for this presentation.*B.* From all of the above considerations, the following roadmap for the **project** has been derived for this presentation.
(Ballim *et al.* 1999, p.14)

(A.798) *A.* These basic phrases are produced by a collection finite-state rules compiled into a transducer. *B.* These basic phrases are produced by a collection **of** finite-state rules compiled into a transducer.
(Jurafsky and Martin 2000, p.388)

(A.799) *A.* An answer to this question can be given only **taking in** account which application the parser is defined for. *B.* An answer to this question can be given only **by** taking **into** account which application the parser is defined for. *C.* An answer to this question can be given only **if one takes into** account which application the parser is defined for.
(Ballim *et al.* 1999, p.7)

(A.800) *A.* But there has certainly been enough parallel publication that this history will **error** on the side of succinctness in giving only a characteristic early mention of each algorithm. *B.* But there has certainly been enough parallel publication that this history will **err** on the side of succinctness in giving only a characteristic early mention of each algorithm.
(Jurafsky and Martin 2000, p.392)

(A.801) *A.* I know **its** necessary and all that but Jesus it doesn't make good reading. *B.* I know **it's** necessary and all that but Jesus it doesn't make good reading.
(http://www.thumped.com/bbs/showthread.php?+=9918, 11/11/03)

(A.802) *A.* We just find that it is easier to find common ground with Irish academics **that** American. *B.* We just find that it is easier to find common ground with Irish academics **than** American.
(IT's Monday News electronic newsletter, 21st July 2001)

(A.803) *A.* However, while it **it** clear that a parser need not be concerned with where a sentence lies on some spectrum of grammaticality, it is also clear that such a parser should be aware that grammatical errors do occur. *B.* However, while it **is** clear that a parser need not be concerned with where a sentence lies on some spectrum of grammaticality, it is also clear that such a parser should be aware that grammatical errors do occur.
(draft of own writing, 31/8/03)

(A.804) *A.* It was the **fist** signs of the facade beginning to fracture. *B.* It was the **first** signs of the facade beginning to fracture.
(Evening Herald, p.15, 23/9/03)

(A.805) *A.* In many cases without total collapse, it is hard to judge the performance since only recently convergence **to** towards common criteria, measures and methods for performance evaluation of NLP-systems has started.*B.* In many cases without total collapse, it is hard to judge the performance since only recently convergence towards common criteria, measures and methods for performance evaluation of NLP-systems has started.
(Ballim *et al.* 1999, p.4)

(A.806) *A.* In fact we must assume the validity this induction principle as we did the validity of the original induction principle of that section. *B.* In fact we must assume the validity **of** this induction principle as we did the validity of the original induction principle of that section.
(Hopcroft *et al.* 2001, p.24)

(A.807) *A.* However it should be better defined what we mean **for** illegitimate utterance. *B.* However it should be better defined what we mean **by** illegitimate utterance.
(Ballim *et al.* 1999, p.7)

(A.808) *A.* He described the County Board attitude as "penny pinching" and says he was "pushed into a corner without support" and felt he had no option to resign. *B.* He described the County Board attitude as "penny pinching" and says he was "pushed into a corner without support" and felt he had no option **but** to resign.
(Evening Herald, p.54, 11/10/03)

(A.809) *A.* The independent **auditors** report comes as both the US Attorney General's office and the Securities & Exchange Commission are conducting a joint investigation into whether Computer Associates manipulated its revenue numbers in an effort to pump up the company's stock price. *B.* The independent **auditor's** report comes as both the US Attorney General's office and the Securities & Exchange Commission are conducting a joint investigation into whether Computer Associates manipulated its revenue numbers in an effort to pump up the company's stock price.
(IT's Monday News electronic newsletter, 13/10/03)

(A.810) *A.* Under my superhero guise of Mumblin Deaf Ro I **having** mostly been moving flat. *B.* Under my superhero guise of Mumblin Deaf Ro I **have** mostly been moving flat.
(personal email, 16/5/03)

(A.811) *A.* I loved the Chameleon, but I know it doesn't suit everyone so do you have **another suggestions**? *B.* I loved the Chameleon, but I know it doesn't suit everyone so do you have **any other** suggestions? *C.* I loved the Chameleon, but I know it doesn't suit everyone so do you have another **suggestion**?
(personal email, 28/5/03)

(A.812) *A.* I would like to hear the dissenting voice re: **Americas** professional dissenter du jour. *B.* I would like to hear the dissenting voice re: **America's** professional dissenter du jour.
(http://www.thumped.com/bbs/showthread.php?+=9918, 11/11/03)

(A.813) *A.* **he** property has been entirely refurbished and redecorated, there is a communal bike shed and rear area to the back and is independently alarmed with colour CCTV intercom access to the building. *B.* **The** property has been entirely refurbished and redecorated, there is a communal bike shed and rear area to the back and is independently alarmed with colour CCTV intercom access to the building.
(estate agent's electronic newsletter, 10/10/03)

(A.814) *A.* Generation should be able take into account stylistic issues.*B.* Generation should be able **to** take into account stylistic issues.
(Ballim *et al.* 1999, p.7)

(A.815) *A.* Since the likes of Lenny Bruce and Sam Kinnison, stand-up comedians have tended to see themselves as gun slingers and with that image often comes the belief that to be great comedian you have to drink hard, take a lot of drugs and leave a beautiful corpse. *B.* Since the likes of Lenny Bruce and Sam Kinnison, stand-up comedians have tended to see themselves as gun slingers and with that image often comes the belief that to be **a** great comedian you have to drink hard, take a lot of drugs and leave a beautiful corpse.
(Evening Herald, p.13, 27/9/03)

(A.816) *A.* Event organisers effectively laid siege to an enduring tradition Hollywood frivolity, it seems, has been a casualty of the war in Iraq. *B.* Event organisers effectively laid siege to an enduring tradition**.** Hollywood frivolity, it seems, has been a casualty of the war in Iraq.
(Night and Day supplement, Irish Independent, 28/3/03)

(A.817) *A.* Finite-state automata techniques are used in a wide range of domains, including switching theory, pattern matching, pattern recognition, speech processing, handwriting recognition, optical character recognition, encryption **algorithm**, data compression, indexing and operating system analysis. *B.* Finite-state automata techniques are used in a wide range of domains, including switching theory, pattern matching, pattern recognition, speech processing, handwriting recognition, optical character recognition, encryption **algorithms**, data compression, indexing and operating system analysis.
(Roche and Schabes 1997, p.1)

(A.818) *A.* If we accept that the function of language is to communicate, it is hard not to conclude that sentences which **have** produced in order to communicate are somehow closer to the reality of language than those whose sole reason for existence is to have their grammaticality judged. *B.* If we accept that the function of language is to communicate, it is hard not to conclude that sentences which have **been** produced in order to communicate are somehow closer to the reality of language than those whose sole reason for existence is to have their grammaticality judged. *C.* If we accept that the function of language is to communicate, it is hard not to conclude that sentences which **were** produced in order to communicate are somehow closer to the reality of language than those whose sole reason for existence is to have their grammaticality judged.
(draft of own writing, 27/8/03)

(A.819) *A.* The attraction is that it is a passive system, it sits on the side of the network and it fails, which none have to date, there is no negative effect on the network. *B.* The attraction is that it is a passive system, it sits on the side of the network and **if** it fails, which none have to date, there is

no negative effect on the network.
(IT's Monday News electronic newsletter, 4/8/03)

(A.820) *A*. Over the history of generative grammar, much has made of its heavy reliance on introspective judgments and their nonequivalence to production and comprehension. *B*. Over the history of generative grammar, much has **been** made of its heavy reliance on introspective judgments and their nonequivalence to production and comprehension.
(Schütze 1996, p.48)

(A.821) *A*. Well I think we can guess **whose** laughing now. *B*. Well I think we can guess **who's** laughing now.
(Evening Herald, p.15, 25/9/03)

(A.822) *A*. Each department is trying to add the IT to help them **to** their bit which makes the penalty points system as a whole harder to implement, harder to manage and amounts to less efficient use of expensive IT systems. *B*. Each department is trying to add the IT to help them **do** their bit which makes the penalty points system as a whole harder to implement, harder to manage and amounts to less efficient use of expensive IT systems. *C*. Each department is trying to add the IT to help them **to do** their bit which makes the penalty points system as a whole harder to implement, harder to manage and amounts to less efficient use of expensive IT systems.
(IT's Monday electronic newsletter, 23rd June 2003)

(A.823) *A*. With this notation in hand, the following operations are defined for two sets of strings L1 and **and** L2. *B*. With this notation in hand, the following operations are defined for two sets of strings L1 and L2.
(Roche and Schabes 1997, p.3).

(A.824) *A*. Ask **they** user whether they wish to continue. *B*. Ask **the** user whether they wish to continue.
(CS1LL4 Lecture Notes, February 2003)

(A.825) *A*. The use of one **just** error rule at a time limits the search space. *B*. The use of **just** one error rule at a time limits the search space.
(draft of own writing, 16/9/03)

(A.826) *A*. For instance, if the cont:C specification had been left out of either the input **our** output category description, then the output category of the rule would have a completely unconstrained content value.*B*. For instance, if the cont:C specification had been left out of either the input **or** output category description, then the output category of the rule would have a completely unconstrained content value.
(ALE User Guide, Version 3.2.1. December 2001, p.49)

(A.827) *A*. I don't know but it's **its** crawling up your leg. *B*. I don't know but it's crawling up your leg.
(http://www.channel4.com/bigbrother/news/newsstory_t.jsp?id=4242, 25/6/03)

(A.828) *A*. I am psychologist and student of linguistics and oriental languages. *B*. I am **a** psychologist and student of linguistics and oriental languages.
a letter quoted in James (1998, p.268)

(A.829) *A*. Microsoft marketing tends to the at the enterprise level but the bulk of the market is SMEs particularly in Ireland. *B*. Microsoft marketing tends to **be** at the enterprise level but the bulk of the market is SMEs particularly in Ireland.
(IT's Monday news electronic newsletters, 25th August 2003)

(A.830) *A*. If you are a current ACL member for 2003 and did not receive a PIN for voting, please send email immediately to carberry@cis.udel.edu with the subject heading MISSING PIN, and in the body of the message give your full name and indicate when you paid your dues for 2003. *B*. If you are a current ACL member for 2003 and did not receive a PIN for voting, please send **an** email immediately to carberry@cis.udel.edu with the subject heading MISSING PIN, and in the body of the message give your full name and indicate when you paid your dues for 2003.
(ACL electronic newsletter, 10/10/03)

(A.831) *A*. I **planning** on organising some gigs on friday nights, where the first 101 people in will get free cd's featuring my band **an** my band alone. B. I **am** planning on organising some gigs on friday nights, where the first 101 people in will get free cd's featuring my band **and** my band alone. C. I **plan** on organising some gigs on friday nights, where the first 101 people in will get free cd's featuring my band **and** my band alone.
(http://thumpedbbs.com/bbs/showthread.php?s=1bbe2c085f846a35bff110a1e3e93b51&threadid=4981, February 2003)

(A.832) *A*. If you're interested send a mail to Eileen Murphy at eileen@fakeemailaddress.com and let her **no** where you heard about the job. *B*. If you're interested send a mail to Eileen Murphy at eileen@fakeemailaddress.com and let her **know** where you heard about the job.
(personal email, 16/1/04)

(A.833) *A.* If the probability of the most likely parse is very small, we must **take** sigma very small, which means that N gets very large. *B.* If the probability of the most likely parse is very small, we must **make** sigma very small, which means that N gets very large.
(Bod 1998, p.48)

(A.834) *A.* **It's** purpose is to use the methods in the syntax tree class to create a particular syntax tree. *B.* **Its** purpose is to use the methods in the syntax tree class to create a particular syntax tree.
(student assignment, 2/4/03)

(A.835) *A.* It was obvious from seeing the couple, pictured below, at the Westlife party that they are **a great together**. *B.* It was obvious from seeing the couple, pictured below, at the Westlife party that they are **great together**. *C.* It was obvious from seeing the couple, pictured below, at the Westlife party that they are **a great couple**.
(Irish Daily Mirror, p.27, 3/12/03)

(A.836) *A.* To extend the partial parses with the complete parse A, the parser **exams** the chart for incomplete constituent, and constructs new constituents with various dot positions as long as all the symbols between A and the new dot position are optional. *B.* To extend the partial parses with the complete parse A, the parser **examines** the chart for **an** incomplete constituent, and constructs new constituents with various dot positions as long as all the symbols between A and the new dot position are optional.
(Wang 2001)

(A.837) *A.* This makes **is** possible for IT companies to create low-cost, fixed-function devices, which can deliver a high end multimedia workstation experience, while simultaneously avoiding the usual demand for upgrades as application resource requirements increase. *B.* This makes **it** possible for IT companies to create low-cost, fixed-function devices, which can deliver a high end multimedia workstation experience, while simultaneously avoiding the usual demand for upgrades as application resource requirements increase.
(IT's Monday News electronic newsletter, 27/10/03)

(A.838) *A.* Sunderland's win over Mansfield Town in the Carling Cup has broken that awful losing sequence but in the **sort** and medium term, it looks like woe all the way for Mick McCarthy. *B.* Sunderland's win over Mansfield Town in the Carling Cup has broken that awful losing sequence but in the **short** and medium term, it looks like woe all the way for Mick McCarthy.
(Evening Herald, p.69, 15/8/03)

(A.839) *A.* There all these Rio Ferdinands out there now. *B.* There **are** all these Rio Ferdinands out there now.
(Evening Herald, p.5, 8/10/03)

(A.840) *A.* Mercury promises a satisfying venture your favourite activities. *B.* Mercury promises a satisfying venture **involving** your favourite activities.
(Evening Herald, p.26, 11/10/03)

(A.841) *A.* In this section we argue that both suggestions seem not **be** workable in practice. *B.* In this section we argue that both suggestions seem not **to** be workable in practice. *C.* In this section we argue that both suggestions seem not workable in practice.
(Musillo and Sima'an 2002)

(A.842) *A.* Where Charniak describes his model in terms of the probability of a constituent (made up of the probability of the **constituents** POS tag, its head word and its unlexicalised expansion) I describe **the his** probability in terms of the probability of a lexicalised rule. *B.* Where Charniak describes his model in terms of the probability of a constituent (made up of the probability of the **constituent's** POS tag, its head word and its unlexicalised expansion) I describe **his** probability in terms of the probability of a lexicalised rule. *C.* Where Charniak describes his model in terms of the probability of a constituent (made up of the probability of the **constituent's** POS tag, its head word and its unlexicalised expansion) I describe **the** probability in terms of the probability of a lexicalised rule.
(draft of paper on probabilistic parsing by Deirdre Hogan, May 2003)

(A.843) **A.** Apartments seized by CAB **goes** under the hammer. *B.* Apartments seized by CAB **go** under the hammer.
(Ireland.com, breaking news, 6th Feb 2003, 14:46)

(A.844) *A.* It is not going to be possible to take everything off the **flour** and I have indicated this to the Buildings Office. *B.* It is not going to be possible to take everything off the **floor** and I have indicated this to the Buildings Office.
(personal email, 14/8/03)

(A.845) *A.* Exclusivity ties you into a company, regardless of how well it performs when pushing **you** product. *B.* Exclusivity ties you into a company, regardless of how well it performs when pushing **your** product.
(IT's Monday electronic newsletter, 30th June 2003)

(A.846) *A.* This method takes one parameter to **be** make sure that only one random number **has** chosen in IndexOfVerb. *B.* This method takes one parameter to make sure that only one random number **is** chosen in IndexOfVerb.
(student assignment, Feb 2003)

(A.847) *A.* Brasilia was built in less time than the Irish are taking to construct a single light urban railway with two **just** lines. *B.* Brasilia was built in less time than the Irish are taking to construct a single light urban railway with **just** two lines.
(Irish Times, p.15, 17/9/03)

(A.848) *A.* I have had an e-mail from Eamon O'Briain, Director of Innovation Services reminding staff that bookings **he** deleted whenever and as soon as it is known that the scheduled event is not going to take place. *B.* I have had an e-mail from Eamon O'Briain, Director of Innovation Services reminding staff that bookings **be** deleted whenever and as soon as it is known that the scheduled event is not going to take place.
(personal email, 7/1/04)

(A.849) *A.* JavaChart **system** includes a **of** compiler that takes a feature-based grammar, a sequence of lexicons and a mini-lexicon. *B.* **The** JavaChart system includes a compiler that takes a feature-based grammar, a sequence of lexicons and a mini-lexicon. *C.* JavaChart includes a compiler that takes a feature-based grammar, a sequence of lexicons and a mini-lexicon.
(User Manual for the JavaChart 0.1.3 system, p.5, downloaded May 2003)

(A.850) *A.* It is also worth **to remark** that robustness is a necessary prerequisite in any system designed for real-life applications. *B.* It is also worth **remarking** that robustness is a necessary prerequisite in any system designed for real-life applications.*C.* It is also **worthwhile** to remark that robustness is a necessary prerequisite in any system designed for real-life applications.
(Ballim *et al.* 1999, p.14)

(A.851) *A.* The plan is that we meet **is** Ritchie's at around 11. *B.* The plan is that we meet **at** Ritchie's at around 11. *C.* The plan is that we meet **in** Ritchie's at around 11.
(personal email, 13/2/04)

(A.852) *A.* The project is being backed with STG 1.5 million of public funding **from by** Invest NI. *B.* The project is being backed with STG 1.5 million of public funding **from** Invest NI. *C.* The project is being backed with STG 1.5 million of public funding **by** Invest NI.
(IT's Monday News electronic newsletter, 8/12/03)

(A.853) *A.* Real IRA leader Liam Campbell smirks as he leaves the Four Courts in Dublin where he is appealing his conviction for membership of the **the** renegade terror group. *B.* Real IRA leader Liam Campbell smirks as he leaves the Four Courts in Dublin where he is appealing his conviction for membership of the renegade terror group.
(Evening Herald, 29/10/03, p.17)

(A.854) *A.* Gardai are said to told Mr. Campbell's family that a prosecution is not likely to take place because of the age of the suspect. *B.* Gardai are said to **have** told Mr. Campbell's family that a prosecution is not likely to take place because of the age of the suspect.
(Evening Herald, p.6, 15/8/03)

(A.855) *A.* David Sommeil played a short pass to Shaun Wright-Phillips who tried his luck with a shot that went in **of** TNS goalkeeper Gerard Doherty. *B.* David Sommeil played a short pass to Shaun Wright-Phillips who tried his luck with a shot that went in **off** TNS goalkeeper Gerard Doherty.
(Irish Times, p.21, 15/8/03)

(A.856) *A.* Wait until tomorrow before making changes of any kind **if where** home matters are concerned. *B.* Wait until tomorrow before making changes of any kind **where** home matters are concerned. *C.* Wait until tomorrow before making changes of any kind **if** home matters are concerned.
(Evening Herald, p. 30,14/8/03)

(A.857) *A.* Word sequences which have been deemed **ungrammaticality** by native speakers provide negative evidence for **a linguistic theories**. *B.* Word sequences which have been deemed **ungrammatical** by native speakers provide negative evidence for a linguistic **theory**. *C.* Word sequences which have been deemed **ungrammatical** by native speakers provide negative evidence for linguistic **theories**.
(draft of own writing, 31/8/03)

(A.858) *A.* This enables them to claim that if well-formedness outweighs faithfulness, the null parse, instead of a **party** underparsed structure, may come out as the optimal candidate. *B.* This enables them to claim that if well-formedness outweighs faithfulness, the null parse, instead of a **partly** underparsed structure, may come out as the optimal candidate.
(Boersma *et al.* 2000, p.23)

(A.859) *A.* The first of these visiting **scientist** will begin their contracts on the first of January 2004. *B.* The first of these visiting **scientists** will begin their contracts on the first of January 2004.
(IT's Monday News electronic newsletter, 8/9/03)

(A.860) *A.* Never one for the muck, Sissy spent the same **mount** of time working her hair. *B.* Never one for the muck, Sissy spent the same **amount** of time working her hair.
(http://www.channel4.com/bigbrother/news/newsstory.jsp?id=1526, June 2003)

(A.861) *A.* Btw., I'll be away from Wednesday this week, and if you're away next week, it looks like the week of the 13th **of** the 17th. *B.* Btw., I'll be away from Wednesday this week, and if you're away next week, it looks like the week of the 13th **to** the 17th.
(personal email, January 2004)

(A.862) *A.* Don't let anyone tell you**.** You are over-ambitious. *B.* Don't let anyone tell you you are over-ambitious. *C.* Don't let anyone tell you **that** you are over-ambitious.
(Evening Herald, p.26, 26/9/03)

(A.863) *A.* Whereas at the end of the **of** 80s the average number of words in an NLP-system was sometimes reported to be less than 50, currently even the vocabularies of complex NLP-applications like spoken-language-machine-translation systems count by the thousands.(**?**) Whereas at the end of the 80s the average number of words in an NLP-system was sometimes reported to be less than 50, currently even the vocabularies of complex NLP-applications like spoken-language-machine-translation systems count by the thousands.
(Ballim *et al.* 1999, p.3)

(A.864) *A.* The US military said 320 Iraqi foot soldiers were killed in the fighting for control the airport, just 12 miles from the city centre. *B.* The US military said 320 Iraqi foot soldiers were killed in the fighting for control **of** the airport, just 12 miles from the city centre.
(http://www.ireland.com/newspaper/breaking/2003/0404/index.htm)

(A.865) *A.* Jones and Lennon **gives** examples where native speakers couldn't find agreement. *B.* Jones and Lennon **give** examples where native speakers couldn't find agreement.
(draft of own writing, 11/10/03)

(A.866) *A.* **You** credit card has been charged for 234.65. *B.* **Your** credit card has been charged for 234.65.
(header of spam mail, 23/9/03)

(A.867) *A.* They use that system and **this** have to declare ungrammatical a vast amount of what people in fact say, which is equally absurd. *B.* They use that system and **thus** have to declare ungrammatical a vast amount of what people in fact say, which is equally absurd.
(transcription error, original sentence in Gethin (1990, p.144))

(A.868) *A.* The college defended its reputation last night, pointing out that it has the largest number of students at both undergraduate and postgraduate level, and has greatest number of students with higher CAO points results. *B.* The college defended its reputation last night, pointing out that it has the largest number of students at both undergraduate and postgraduate level, and has **the** greatest number of students with higher CAO points results.
(Evening Herald, Going to College supplement, p.2, 19/8/03)

(A.869) *A.* In 2001, this global market was valued at USD 9 billion, the Phase IV element of which **comprises of** approximately USD 4 billion. *B.* In 2001, this global market was valued at USD 9 billion, the Phase IV element of which **comprises** approximately USD 4 billion. *C.* In 2001, this global market was valued at USD 9 billion, the Phase IV element of which **consists of** approximately USD 4 billion.
(IT's Monday News electronic newsletter, 13/10/03)

(A.870) *A.* Widespread frost is expected blanket the country. *B.* Widespread frost is expected **to** blanket the country.
(Evening Herald, p.5, 29/10/03)

(A.871) *A.* Write an application class called PointProgram in which two point objects are created and the distance between the two points **are** calculated. *B.* Write an application class called PointProgram in which two point objects are created and the distance between the two points **is** calculated.
(lecture notes, November 2002)

(A.872) *A.* There are essentially four problems that cannot be solved simply by extending standard parsing techniques in **the** such a direction.*B.* There are essentially four problems that cannot be solved simply by extending standard parsing techniques in such a direction.
(Ballim *et al.* 1999, p.10)

(A.873) *A.* According to AJ McKee, public relations officer for the Linux Users Group, not only are users finding it impossible to see an adequate amount of code to prove exactly what the product is, there also seems to be confusion **at** to whether or not the SCO Linux license is available in Europe. *B.* According to AJ McKee, public relations officer for the Linux Users Group, not only are users finding it impossible to see an adequate amount of code to prove exactly what the product is, there also seems to be confusion **as** to whether or not the SCO Linux license is available in Europe.
(IT's Monday News electronic newsletter, 27/10/03)

(A.874) *A.* Head-driven Phrase Structure Grammar is **an** theoretical framework for expressing generalizations about linguistic phenomena in a way that integrates information supplied from syntax, semantics and pragmatics. *B.* Head-driven Phrase Structure Grammar is **a** theoretical framework for expressing generalizations about linguistic phenomena in a way that integrates information supplied from syntax, semantics and pragmatics.
(http://www.cs.tcd.ie/Carl.Vogel, April 2003)

(A.875) *A.* Based on this analysis **a** error rules were derived from conventional grammar rules. *B.* Based on this analysis error rules were derived from conventional grammar rules. *C.* Based on this analysis a **set of** error rules was derived from conventional grammar rules.
(draft of own writing, 16/9/03)

(A.876) *A.* While everyone applauds the work done **of** by the rehabilitation teams in our prisons, there is simply no excuse for the casual brutality that is becoming an increasing feature of our society. *B.* While everyone applauds the work done by the rehabilitation teams in our prisons, there is simply no excuse for the casual brutality that is becoming an increasing feature of our society.
(Evening Herald, p.14, 27/11/03)

(A.877) *A.* It was an apt chat seeing as though the twelve of them are living in very challenging circumstances for the chance to the not inconsiderable sum of 70,000. *B.* It was an apt chat seeing as though the twelve of them are living in very challenging circumstances for the chance to **win** the not inconsiderable sum of 70,000.
(http://www.channel4.co.uk/bigbrother/news/newsstory_t.jsp?id=1259,27/5/03)

(A.878) *A.* The approach is **be** an alternative to that of Abramson and Dahl. *B.* The approach is an alternative to that of Abramson and Dahl.
(draft of "Enabling Left Recursive DCGs" by Carl Vogel, March 2003)

(A.879) *A.* Humans are able to combine different **level** of understanding in order achieve an acceptable or even partial understanding of the input text. *B.* Humans are able to combine different **levels** of understanding in order **to** achieve an acceptable or even partial understanding of the input text.
(Ballim *et al.* 1999, p.3)

(A.880) *A.* The courses in French, German and Irish aim to give students sufficient competence in their language of focus to be able to operate in that language in **in** their future careers. *B.* The courses in French, German and Irish aim to give students sufficient competence in their language of focus to be able to operate in that language in their future careers.
(http://www.cs.tcd.ie/courses/csll/CSLLcourse.html, 14/4/03)

(A.881) *A.* Thus, the type of this attribute will the type of a node, which is the type we are currently defining. *B.* Thus, the type of this attribute will **be** the type of a node, which is the type we are currently defining.
(lecture notes, March 2003)

(A.882) *A.* CDC has committed to increasing Pivotal's marketing budget by 200 per cent and its technical support staff by 40 per cent - an area that Gartner believes Pivotal had cut back **to** far on. *B.* CDC has committed to increasing Pivotal's marketing budget by 200 per cent and its technical support staff by 40 per cent - an area that Gartner believes Pivotal had cut back **too** far on.
(IT's Monday News electronic newsletter, 13/12/03)

(A.883) *A.* Write a program that generates sentences in German language. *B.* Write a program that generates sentences in **the** German language.
(student assignment, Feb 2003)

(A.884) *A.* This means that if the node which starts at the top of the syntax tree **and** has no subtrees then it evaluates to null. *B.* This means that if the node which starts at the top of the syntax tree has no subtrees then it evaluates to null.
(student assignment, 2/4/03)

(A.885) *A.* Nush tried to warn him that could be more shocks to come. *B.* Nush tried to warn him that **there** could be more shocks to come.
(http://bigbrother.channel4.com/bigbrother/news/newsstory_t.jsp?id=5125, 14th July 2003)

(A.886) *A.* One method being used is to put the grenade with **it's** pin out in a glass of water and let it freeze rock-solid in the frigid Kosovo winter temperatures which hit minus seven degrees centigrade this week. *B.* One method being used is to put the grenade with **its** pin out in a glass of water and let it freeze rock-solid in the frigid Kosovo winter temperatures which hit minus seven degrees centigrade this week.
(Evening Herald, p.13, 10/12/03)

(A.887) *A.* I got my essay done no problem though I still have to fix it up and **to** the references and all. *B.* I got my essay done no problem though I still have to fix it up and **do** the references and all.
(personal email, 6/11/03)

(A.888) *A.* At Dublin City University, the **pints** on its Business Studies course remain the same as last year at 430, although the college said many students have far higher points, while its Accounting and Finance, Financial and Actuarial Mathematics and European Business were also up. *B.* At Dublin City University, the **points** on its Business Studies course remain the same as last year at 430, although the college said many students have far higher points, while its Accounting and Finance, Financial and Actuarial Mathematics and European Business were also up.
(Evening Herald, Going to College supplement, p.3, 19/8/03)

(A.889) *A.* Our guide attempted to teach us a **childrens** song in Thai about an Elephant which was very appropriate as we took an elephant ride the next day. *B.* Our guide attempted to teach us a **children's** song in Thai about an Elephant which was very appropriate as we took an elephant ride the next day.
(personal email, 13/10/03)

(A.890) *A.* The connective is replaced by a connective which **describe** the error. *B.* The connective is replaced by a connective which **describes** the error.
(draft of own writing, 16/9/03)

(A.891) *A.* A large body of literature exists which criticizes the reliability of such **such** data. *B.* A large body of literature exists which criticizes the reliability of such data.
(draft of own writing, 26/8/03)

(A.892) *A.* Inter-subject variation, for example, is bound to arise if informants have differing ideas about what aspect of a sentence they are supposed to be giving their opinion on, or if the ability **of** perform the judgement task varies from individual to individual. *B.* Inter-subject variation, for example, is bound to arise if informants have differing ideas about what aspect of a sentence they are supposed to be giving their opinion on, or if the ability **to** perform the judgement task varies from individual to individual.
(draft of own writing, 2/12/03)

(A.893) *A.* Work should enjoyed not endured. *B.* Work should **be** enjoyed not endured.
(Evening Herald, p.28, 11/8/03)

(A.894) *A.* On the surface, her relationship with seems idyllic. *B.* On the surface, her relationship with **her partner** seems idyllic.
(Now magazine, 22/10/03, p.7)

(A.895) *A.* The next day we headed up to a lovely town just north of Queenstown which again had the most amazing scenery around the **towns** lake. *B.* The next day we headed up to a lovely town just north of Queenstown which again had the most amazing scenery around the **town's** lake.
(personal email, 19/4/03)

(A.896) *A.* Unpack the gzipped tar-ball and you will find the files at top level of the unpacked file tree. *B.* Unpack the gzipped tar-ball and you will find the files at **the** top level of the unpacked file tree.
(User Manual for the JavaChart 0.1.3 system, p.6, downloaded May 2003)

(A.897) *A.* Therefore the application logic can be implemented according to the semantic class structure, in parallel with the development of linguistic context free **grammar**. *B.* Therefore the application logic can be implemented according to the semantic class structure, in parallel with the development of **a** linguistic context free grammar. *C.* Therefore the application logic can be implemented according to the semantic class structure, in parallel with the development of linguistic context free **grammars**.
(Wang 2001)

(A.898) *A.* I wanted to go out drinking as did Andy and it turned out Sarah did **to** even though she didn't know it. *B.* I wanted to go out drinking as did Andy and it turned out Sarah did **too** even though she didn't know it.
(personal email, 13/11/02)

(A.899) *A.* The feature structure F called the rule feature structure **and** holds the actual values for the equations of the dotted rule of the chart edge. *B.* The feature structure F **is** called the rule feature structure and holds the actual values for the equations of the dotted rule of the chart edge. *C.* The feature structure F called the rule feature structure holds the actual values for the equations of the dotted rule of the chart edge.
(User Manual for the JavaChart 0.1.3 system, p.4, downloaded May 2003)

(A.900) *A.* Director Peter Smith who been heavily involved with the development of the Seasiders' stadium since joining the board has resigned. *B.* Director Peter Smith who **has** been heavily involved with the development of the Seasiders' stadium since joining the board has resigned. *C.* Director Peter Smith who was heavily involved with the development of the Seasiders' stadium since joining the board has resigned.
(Teletext, p.413, 00:52, 10/9/03)

(A.901) *A.* I may have **mislead** some people with regard to the name of the new socks5 server. *B.* I may have **misled** some people with regard to the name of the new socks5 server.
(personal email, 17/4/03)

(A.902) *A.* There are four to five times the number of opportunities in healthcare **that** there are in call centres. *B.* There are four to five times the number of opportunities in healthcare **than** there are in call centres.
(IT's Monday News electronic newsletter, 13/12/03)

(A.903) *A.* Is it the **actually** payment you are worried about, or the job itself? *B.* Is it the **actual** payment you are worried about, or the job itself?
(personal email, 10/10/02)

(A.904) *A.* I still see my girlfriend but she can't get away very often, so **its** usually just in a hotel for a couple of hours. *B.* I still see my girlfriend but she can't get away very often, so **it's** usually just in a hotel for a couple of hours.
(Evening Herald, p.26, 14/10/03)

(A.905) *A.* The phrasing in this sentence from the word 'as' onwards is unusual enough to suggest that what the **the** writer of the sentence had in mind was 'as they are needed immediately' but as this same sense is captured by 'as they are needed immediately', the decision was taken not to classify this sentence as erroneous. *B.* The phrasing in this sentence from the word 'as' onwards is unusual enough to suggest that what the writer of the sentence had in mind was 'as they are needed immediately' but as this same sense is captured by 'as they are needed immediately', the decision was taken not to classify this sentence as erroneous.
(draft of own writing, 22/10/03)

(A.906) *A.* Safety **features has** also been improved with an 80pc increase in rigidity. *B.* Safety **features have** also been improved with an 80pc increase in rigidity. *C.* **Safety** has also been improved with an 80pc increase in rigidity.
(Evening Herald, p.42, 8/10/03)

(A.907) *A.* The if-statement condition needs to check whether the index selected greater than or equal to six. *B.* The if-statement condition needs to check whether the index selected **is** greater than or equal to six.
(student assignment, Feb 2003)

(A.908) *A.* Several competing enterprise portal vendors **recently** came together **recently** to form an alliance to drive portal demand based on - wait for it - open source. *B.* Several competing enterprise portal vendors **recently** came together to form an alliance to drive portal demand based on - wait for it - open source. *C.* Several competing enterprise portal vendors came together **recently** to form an alliance to drive portal demand based on - wait for it - open source.
(IT's Monday News electronic newsletter, 17/11/03)

(A.909) *A.* Chomsky was of the opinion that the above sentence has no meaning and he used **that** fact that it is grammatical to argue that meaning and grammar are two autonomous concepts. *B.* Chomsky was of the opinion that the above sentence has no meaning and he used **the** fact that it is grammatical to argue that meaning and grammar are two autonomous concepts.
(draft of own writing, 18/8/03)

(A.910) *A.* This **makes** that the information loss may well be able to effectively fulfil its function as a sign post in the search space. *B.* This **means** that the information loss may well be able to effectively fulfil its function as a sign post in the search space.
(Fouvry 2003a, p.111)

(A.911) *A.* Lan's recent success **is** this area is all the more remarkable when recent studies suggest that Irish businesses are not keen on outsourcing. *B.* Lan's recent success **in** this area is all the more

remarkable when recent studies suggest that Irish businesses are not keen on outsourcing.
(IT's Monday News electronic newsletter, 17/2/03)

(A.912) *A.* The robustness to ungrammaticality and noise can be attributed to its ability **of skipping** minimum unparsable segments in the input. *B.* The robustness to ungrammaticality and noise can be attributed to its ability **to skip** minimum unparsable segments in the input.
(Wang 2001)

(A.913) *A.* This is because it seems to me to be the only theoretical linguistics that **are** concerned with something real, interesting or important. *B.* This is because it seems to me to be the only theoretical linguistics that **is** concerned with something real, interesting or important.
(http://www.media.uwe.ac.uk/masoud/editor/ais/ais-1/ais-1-6.htm, July 2002)

(A.914) *A.* He was dressed up in the 70s gear and had a big afro wig on, and so at one stage he lost his heart medalion, and who found it only me, so I got to be really cheesy and go up to him, and tell how he had lost his **hear**. *B.* He was dressed up in the 70s gear and had a big afro wig on, and so at one stage he lost his heart medalion, and who found it only me, so I got to be really cheesy and go up to him, and tell how he had lost his **heart.**
(personal email, 27/10/02)

(A.915) *A.* I wanted to chat about the research seminar idea too if anyone **if** interested. *B.* I wanted to chat about the research seminar idea too if anyone **is** interested.
(personal email, 21/10/03)

(A.916) *A.* We just watched and could nothing. *B.* We just watched and could **do** nothing.
(Evening Herald, p.4, 28/11/03)

(A.917) *A.* In this section we examined the basic notions of robust text analysis, as well as empirical evidence related to robustness **are reviewed**.*B.* In this section we examined the basic notions of robust text analysis, as well as **reviewing** empirical evidence related to robustness.
(Ballim *et al.* 1999, p.14)

(A.918) *A.* Learn how **you to** organise your life at university so that you can manage your study, recreation and social life. *B.* *B.* Learn how **to** organise your life at university so that you can manage your study, recreation and social life. *C.* Learn how **you can** organise your life at university so that you can manage your study, recreation and social life.
(electronic newsletter, 15/1/04)

(A.919) *A.* Write and tell **be** all your news. *B.* Write and tell **me** all your news.
(personal email, 19/4/03)

(A.920) *A.* Managed accounts have **provide** successful elsewhere in Europe. *B.* Managed accounts have **proved** successful elsewhere in Europe.
(IT's Monday News electronic newsletter, 1/12/03)

(A.921) *A.* Visit one of the many monuments the city has to offer, such as The Saint-Etienne Cathedral with **it's** French Gothic architecture. *B.* Visit one of the many monuments the city has to offer, such as The Saint-Etienne Cathedral with **its** French Gothic architecture.
(aerlingus.com electronic newsletter, 14/8/03)

(A.922) *A.* Accordingly, parsing can be viewed as low-level type of text analysis.*B.* Accordingly, parsing can be viewed as **a** low-level type of text analysis.
(Ballim *et al.* 1999, p.6)

(A.923) *A.* VERB_ARRAY **has** initialised at the beginning of the program. *B.* VERB_ARRAY **was** initialised at the beginning of the program.
(student assignment, Feb 2003)

Appendix B

Questionnaire

CORRECTING LANGUAGE ERRORS - A QUESTIONNAIRE

PART ONE
There are six passages of text. Each passage contains an error (the general area containing the error is indicated in italics). Please read the entire passage and indicate in the space provided how you would correct the error.
You can either write out the corrected sentence in full or just make a note of the change or changes you would make to correct the error.
It is possible that the text contains more than one error. Only correct an error that is shown in italics.
If you think of another correction which fits in with the surrounding text, please write it down also.
If you cannot correct the error, please indicate the reason why you cannot. For example, one reason why you might not want to correct the error is if you don't think it is an error. Another reason might be that you do not think the text makes any sense.

EXAMPLE:
Eight couples will be picked out from a series of auditions across the country to compete for a roof over their heads. All of the entrants must not own a house, be married and have to be over 18 years old but the good news is if they win the house *it can sold for* hard cash. The three bedroom pile is part of a smart new development called Lakepark Drive just outside of Mullingar.

```
add be --> it can be sold for
```

1. Manchester United manager Alex Ferguson has just concluded a new pay deal which will give him a total weekly income of nearly EUR 100,000. Roy Keane, of course, earns even more per week. And the average income *for a players* in the top half of the Premiership now runs to over EUR50,000 weekly.

2. In addition, signings for the first-half of the year are so far USD 6.4 billion, compared with USD 13.4 billion in the comparable 2002 period. However, despite such dismal figures, Gartner has given the company a thumbs-up, claiming that *decisions not sell its* management consulting group, A T Kearney and its software development group, Product Life Cycle Management (PLM) Solutions and a definitive agreement with Fiserv, which plans to acquire EDS's Credit Union Industry Group, shows EDS's direction is taking shape.

3. Making the case for biometrics, which uses a person's unique physical attributes such as a fingerprint or face scan to establish their identity, was Martin Walsh, head of Legal and Regulatory Affairs with Daon and also chair of the European Biometric Forum. Walsh said that while there are privacy concerns around the introduction of biometrics, when used as part of a comprehensive identity management system it can actually enhance the ethical user's privacy. In particular biometrics can *be used use to* prove a person's right to access personal data that may be held by an organisation and also prevent identity theft.

4. The most recent attack happened two weeks ago when raiders used a steel bar to break a large window in the back of his house while Fr. Coughlan was home alone. When I *heard was the crashing* of glass, I knew there was something coming in again.

5. Give your soul a chance to catch up with your body. The quality of your life is in direct *proportion of your commitment* to excellence. One of your strengths is the ability to assess situations. The emphasis this week is on all matters to do with creativity, self-expression and your general level of vitality.

6. Beckham and company want to show their support for Ferdinand, but realise they would compromise their England careers if they carry through with their threat. And for now the FA are playing *down the players action.* Director of communications, Paul Barber said: "They obviously wanted to express their opinions and views on what has happened. We were very happy to listen. They also listened to our explanation of what has happened."

PART TWO
Below are six sentences. Each sentence contains some kind of error (in italics). Indicate in the space provided how you would correct this sentence.
You can either write out the corrected sentence in full or just make a note of the change or changes you would make to the sentence.
If you think that there is another way to correct the sentence, please write it down also.
Only correct errors that are shown in italics.
If you cannot correct the sentence, please indicate the reason why you cannot. For example, one reason why you might not want to correct the sentence is if you don't think there is anything wrong with it. Another reason might be that you just don't know what the sentence means.

1. All are very welcome - *please tell you friends.*

2. HP America *is ending staff over* to work on the programmes rather than basing them with a US university and already has six US researchers working full time in Ireland.

3. We *might found out what* was going on every six months, during which time researchers could have followed a path we didn't want them to take, wasting both our time and our money.

4. We *order then directly from* the web.

5. As with the previous *elections, ensure the involvement* of the community, the new FoLLI president will be elected directly by the members.

6. We just find that it is easier to find common ground with *Irish academics that American.*

CORRECTING LANGUAGE ERRORS - A QUESTIONNAIRE

PART ONE
There are six passages of text. Each passage contains an error (the general area containing the error is indicated in italics). Please read the entire passage and indicate in the space provided how you would correct the error.
You can either write out the corrected sentence in full or just make a note of the change or changes you would make to correct the error.
It is possible that the text contains more than one error. Only correct an error that is shown in italics.
If you think of another correction which fits in with the surrounding text, please write it down also.
If you cannot correct the error, please indicate the reason why you cannot. For example, one reason why you might not want to correct the error is if you don't think it is an error. Another reason might be that you do not think the text makes any sense.

EXAMPLE:
Eight couples will be picked out from a series of auditions across the country to compete for a roof over their heads. All of the entrants must not own a house, be married and have to be over 18 years old but the good news is if they win the house *it can sold for* hard cash. The three bedroom pile is part of a smart new development called Lakepark Drive just outside of Mullingar.

```
add be --> it can be sold for
```

1. The quiz is to support the establishment of a library in a school in the north of Cameroon, West Africa. Why not form a team? Lots of prizes, including one for the top scoring research group! All are very welcome - *please tell you friends.*

2. Dunne estimated that HP spent between EUR 10 to EUR 20 million last year with Irish universities and that figure is set to grow. The company is currently talking to the Science Foundation about setting up two large R&D programmes in the areas of nano science and sensor research. Eighty per cent of the funds earmarked for these projects will be diverted to universities involved in the projects. HP America *is ending staff over* to work on the programmes rather than basing them with a US university and already has six US researchers working full time in Ireland.

3. One of the clauses, which we have a big problem with, is that Stanford owns all the intellectual property arising from the project, despite the fact that we have paid for it! We also found there was little communication - we didn't know what was happening within Stanford. We *might found out what* was going on every six months, during which time researchers could have followed a path we didn't want them to take, wasting both our time and our money. Irish universities, on the other hand, are happy to negotiate terms in regards to IP and they understand we need to get some thing out of the relationship too.

4. Here in the office, we came up with the idea of personalised bottles of wine for John with the name of his son on the label. We can *order then directly from* the web. The idea is to get a wine that ages well (like a St.Emilion or Medoc), so that they can open them for the boy's 18th (or 21st) birthday

5. The primary task of the FoLLI President is to initiate efforts aimed at securing the long term interests of the Logic, Language & Information community. For example, during his presidency, Hans Uszkoreit has strongly contributed to raising awareness of the association both within and outside the field, by supporting links with other associations and strengthening EU relations. As with the previous *elections, ensure the involvement* of the community, the new FoLLI president will be elected directly by the members. The election process consists of a nomination and an election phase.

6. According to Dunne, academic arrogance in the US is increasingly putting private companies off working in partnership with academics, and they are looking to overseas operations to provide the data needed for R&D projects. "We just find that it is easier to find common ground with *Irish academics that American*," said Dunne. "The relationship is a lot more flexible."

PART TWO

Below are six sentences. Each sentence contains some kind of error (in italics). Indicate in the space provided how you would correct this sentence.

You can either write out the corrected sentence in full or just make a note of the change or changes you would make to the sentence.

If you think that there is another way to correct the sentence, please write it down also.

Only correct errors that are shown in italics.

If you cannot correct the sentence, please indicate the reason why you cannot. For example, one reason why you might not want to correct the sentence is if you don't think there is anything wrong with it. Another reason might be that you just don't know what the sentence means.

1. And the average income *for a players in* the top half of the Premiership now runs to over EUR50,000 weekly.

2. However, despite such dismal figures, Gartner has given the company a thumbs-up, claiming that *decisions not sell its* management consulting group, A T Kearney and its software development group, Product Life Cycle Management (PLM) Solutions and a definitive agreement with Fiserv, which plans to acquire EDS's Credit Union Industry Group, show EDS's direction is taking shape.

3. In particular biometrics can *be used use to* prove a person's right to access personal data that may be held by an organisation and also prevent identity theft.

4. When I *heard was the crashing* of glass, I knew there was something coming in again.

5. The quality of your life is in direct *proportion of your commitment* to excellence.

6. And for now the FA are playing *down the players action.*

Appendix C

Two-Stage Error Anticipation Parser Code & Test Data

This appendix contains a listing of the Prolog code which implements the parser described in Chapter 5. The original code for the chart parser which is used for the first phase was developed by my supervisor Dr. Carl Vogel. I used it as a basis for the second phase. The sentences which were used to evaluate this parser are given after the code.

```
:- op(1200,xfx,missingop).        % Error rule connective
:- op(1200,xfx,spellop).          % Error rule connective
:- op(1200,xfx,agreeop).          % Error rule connective
:- op(1200,xfx,vformop).          % Error rule connective
:- op(1200,xfx,extraop).          % Error rule connective
:- op(1200,xfx,lexop).            % Error rule connective
:- op(1200,xfx,possop).           % Error rule connective
:- op(1200,xfx,prepop).           % Error rule connective
:- op(1200,xfx,indefop).          % Error rule connective
:- op(1200,xfx,infprepop).        % Error rule connective
:- op(1200,xfx,auxop).            % Error rule connective
:- op(1200,xfx,adjop).            % Error rule connective
:- op(1200,xfx,conjop).           % Error rule connective


:-use_module(library(charsio)).
:-dynamic to_process/1,edge/6,error_edge/6,edge_num/1,node_num/1,num_name/2,
cycle_count/1,parse_count/1,success/0,forest/2,forest_error/2,packed/1,
packed_error/1,propose/1.
:-unknown(_,fail).
not(X) :- \+ X.


% Grammar rules are interpreted and their right hand sides are converted
% into lists.
term_expansion((Lhs -->  Rhs), rule(Lhs,List)) :-
  convert(Rhs,List).
term_expansion((Lhs missingop Rhs), missing_error_rule(Lhs,List)):-
   convert(Rhs,List).
term_expansion((Lhs spellop Rhs), spell_error_rule(Lhs,List)):-
   convert(Rhs,List).
term_expansion((Lhs agreeop Rhs), agree_error_rule(Lhs,List)):-
   convert(Rhs,List).
term_expansion((Lhs vformop Rhs), vform_error_rule(Lhs,List)):-
   convert(Rhs,List).
term_expansion((Lhs extraop Rhs), extra_error_rule(Lhs,List)):-
   convert(Rhs,List).
term_expansion((Lhs lexop Rhs), lex_error_rule(Lhs,List)):-
   convert(Rhs,List).
term_expansion((Lhs possop Rhs), poss_error_rule(Lhs,List)):-
   convert(Rhs,List).
term_expansion((Lhs prepop Rhs), prep_error_rule(Lhs,List)):-
   convert(Rhs,List).
term_expansion((Lhs indefop Rhs), indef_error_rule(Lhs,List)):-
   convert(Rhs,List).
term_expansion((Lhs infprepop Rhs), infprep_error_rule(Lhs,List)):-
   convert(Rhs,List).
term_expansion((Lhs auxop Rhs), aux_error_rule(Lhs,List)):-
   convert(Rhs,List).
term_expansion((Lhs adjop Rhs), adj_error_rule(Lhs,List)):-
   convert(Rhs,List).
term_expansion((Lhs conjop Rhs), conj_error_rule(Lhs,List)):-
   convert(Rhs,List).

convert((Element,Rest),[Element|List]) :-
  !, convert(Rest,List).
convert(Element,[Element]).
```

```
%% The Parser

% The top-level procedure which starts the parse.
top_parse(WordList,StartSymbol,Behaviour,Direction,All_or_First) :-
  clear_edges,
  set_edge_num_to_one,
  set_node_num_to_one,
  set_cycle_count_to_zero,
  set_parse_count_to_zero,
  init_chart(StartSymbol,WordList,Behaviour,Direction),
  parse(StartSymbol,WordList,Behaviour,Direction,All_or_First).

% Procedure which clears the chart and the agenda.
clear_edges :-
  abolish(to_process,1),
  abolish(success,0),
  abolish(packed,1),
  abolish(propose,1),
  abolish(forest,2),
  abolish(forest_error,2),
  abolish(edge,6),
  abolish(error_edge,6).


% Edge format:  edge(Category, Constituents, Expects, LeftPos, RightPos,
% Type, Num).

% Pre-terminal inactive edges are added to the agenda.
% An extra edge must be added if the rule invokation strategy is top-down.
init_chart(StartSymbol,WordList,Behaviour,top_down) :-
  init_chart(WordList,Behaviour),
  make_edge(edge(start,[],[StartSymbol],WordList,WordList),Behaviour).

init_chart(_,Wordlist,Behaviour,bottom_up) :-
  init_chart(Wordlist,Behaviour).

init_chart([],_).
init_chart([Word|Sentence],Behaviour) :-
  word_in_lexicon(Word),
  init_chart_word(Word,Sentence,Behaviour),
  init_chart(Sentence,Behaviour).

word_in_lexicon(Word):-
   rule(NonTerminal,[[Word]]).
init_chart_word(Word,Rest,Behaviour):-
  rule(NonTerminal,[[Word]]),
  make_edge(edge(NonTerminal,Word,[],[Word|Rest],Rest),Behaviour),
  fail.
init_chart_word(_,_,_).

% This procedure assigns a unique number to an edge and adds it to the top
% of the agenda.
make_edge(Edge,lifo) :-
  not(redundant(Edge)),
  Edge =.. List,
  edgenumgen(N),
  conc(List,[N],Arity_six_list),
  Real_edge =.. Arity_six_list,
  asserta(to_process(Real_edge)),!.
```

```
% This procedure assigns an unique number to an edge and adds it to the
% bottom of the agenda.
make_edge(Edge,fifo) :-
  not(redundant(Edge)),
  Edge =.. List,
  edgenumgen(N),
  conc(List,[N],Arity_six_list),
  Real_edge =.. Arity_six_list,
  assertz(to_process(Real_edge)).

make_edge(_,_).   % make_edge always succeeds.

% Adds an edge to a forest
add_to_forest(Edge,Root):-
  Edge =.. List,
  edgenumgen(N),
  conc(List,[N],Arity_six_list),
  Real_edge =.. Arity_six_list,
  retract(forest(Root,AList)),!,
  conc([N],AList,NewList),
  assert(forest(Root,NewList)),
  assert(packed(Real_edge)).

add_to_forest(Edge,Root):-
  Edge =.. List,
  edgenumgen(N),
  conc(List,[N],Arity_six_list),
  Real_edge =.. Arity_six_list,
  assert(forest(Root,[N])),
  assert(packed(Real_edge)),!.


% Selects an edge from the agenda.
select_next(Edge) :-
  !,to_process(Edge).


% Removes an edge from the agenda.
remove_current(Edge) :-
  retract(to_process(Edge)),!.


% This procedure performs the actual parsing.
% This clause succeeds when all the parses of a sentence have been found.
parse(Symbol,Wordlist,Behaviour,Direction,all) :-
  select_next(edge(Symbol,Constituents,[],Wordlist,[],Num)),
  asserta(edge(Symbol,Constituents,[],Wordlist,[],Num)),
  remove_current(edge(Symbol,Constituents,[],Wordlist,[],Num)),
  increment_cycle_count,
  fail.


% This clause succeeds when one parse of a sentence has been found.
parse(Symbol,Wordlist,Behaviour,Direction,first) :-
  select_next(edge(Symbol,Constituents,[],Wordlist,[],Num)),
  asserta(edge(Symbol,Constituents,[],Wordlist,[],Num)),
  remove_current(edge(Symbol,Constituents,[],Wordlist,[],Num)),
  increment_cycle_count.


% An edge is taken from the agenda, an attempt is made to apply the
% fundamental rule and directional strategies. The edge is added to
```

```
% the chart and removed from the agenda.
parse(Symbol,Wordlist,Behaviour,Direction,All_or_First) :-
  select_next(Edge),
  fundamental_rule(Edge,Behaviour),
  find_rules(Edge,Behaviour,Direction),
  remove_current(Edge),
  asserta(Edge),
  increment_cycle_count,
  parse(Symbol,Wordlist,Behaviour,Direction,All_or_First).

parse(_,_,_,_,_).

% This works for the case when an inactive edge meets an active edge:
fundamental_rule(edge(Cat,Const,[],Lpos,Rpos,Num),Behaviour) :-
  edge(Act_Cat,Act_Const,[Cat|New_Expects],Act_lpos,Lpos,Act_num),
  make_edge(edge(Act_Cat,[Num|Act_Const], New_Expects ,
  Act_lpos,Rpos),Behaviour),
fail.



% This works for the case when an active edge meets an inactive edge.
fundamental_rule(edge(A_Cat,A_Const,[FirstExpect|New_Expects],A_lpos,A_rpos,A_num),
Behaviour):-
  edge(FirstExpect,_,[],A_rpos,Rpos,Num),
  make_edge(edge(A_Cat,[Num|A_Const],New_Expects,A_lpos,
  Rpos),Behaviour),
 fail.


% This always succeeds.
fundamental_rule(_,_).

% Bottom-up Rule:
find_rules(edge(Cat,Const,[],Lpos,Rpos,Num),Behaviour,bottom_up) :-
  rule(Lhs,[Cat|Rest]),
  make_edge(edge(Lhs,[],[Cat|Rest],Lpos,Lpos),Behaviour),
  fail.

% Top-down Rule:
find_rules(edge(Cat,Const,[Symbol|Expects],Lpos,Rpos,Num),Behaviour,
top_down) :-
  rule(Symbol,Rhs),
  not(terminal(Rhs)),
  make_edge(edge(Symbol,[],Rhs,Rpos,Rpos),Behaviour),
fail.

find_rules(_,_,_).    %find_rules always succeeds.

% Recovery phase parsing
recover_parse(Symbol,Wordlist):-
   set_cycle_count_to_zero,
   set_parse_count_to_zero,
   abolish(success,0),
   abolish(packed_error,1),
   abolish(forest_error,2),
   abolish(error_edge,6),
   abolish(propose,1),
   recover_lexical_parse(Symbol,Wordlist).

recover_parse(Symbol,Wordlist):-
   recover_grammar_parse(Symbol,Wordlist).
```

```
% Missing/Extra word errors
recover_grammar_parse(Symbol,Wordlist) :-
  edge(Cat,Const,[],Lpos,Rpos,Num),
  increment_cycle_count,
  find_error_rule(edge(Cat,Const,[],Lpos,Rpos,Num),Wordlist),
  reparse(Symbol,Wordlist),
  recover_print_trees(Symbol,Wordlist),
  fail.

% Substitution errors
recover_lexical_parse(Symbol,Wordlist):-
    my_member(Word,Wordlist,Rest),
    increment_cycle_count,
    recover_lexical_parse1(Symbol,Wordlist,[Word|Rest]),fail.

my_member(X,[X|R],R).
my_member(X,[_|R],R1):-
  my_member(X,R,R1).

recover_lexical_parse1(Symbol,Wordlist,[Word|Rest]):-
  find_lexical_error_rule(Word,Rest,Wordlist),
  reparse(Symbol,Wordlist),
  recover_print_trees(Symbol,Wordlist).

% Reinvoke the parsing process
reparse(Symbol,Wordlist):-
        abolish(success,0),
abolish(error_edge,6),
        abolish(packed_error,1),
        abolish(forest_error,2),
reparse1(Symbol,Wordlist),!.

reparse1(Symbol,Wordlist) :-
  select_next(error_edge(Symbol,Constituents,[],Wordlist,[],Num)),
  asserta(error_edge(Symbol,Constituents,[],Wordlist,[],Num)),
  remove_current(error_edge(Symbol,Constituents,[],Wordlist,[],Num)),
  increment_cycle_count,
  assert(success),
  fail.

reparse1(Symbol,Wordlist) :-
  select_next(Edge),
  retry_fundamental_rule(Edge),
  retry_find_rules(Edge),
  remove_current(Edge),
  asserta(Edge),
  increment_cycle_count,
  reparse1(Symbol,Wordlist).

reparse1(_,_):- success,!.


% This works for the case when an inactive edges meets an active edge.
retry_fundamental_rule(error_edge(Cat,Const,[],Lpos,Rpos,Num)):-
  edge(Act_Cat,Act_Const,[Cat|Rest],Act_lpos,Lpos,_),
  make_error_edge(error_edge(Act_Cat,[Num|Act_Const],Rest,Act_lpos,Rpos)),
  fail.

% This works for the case when an inactive edges meets an active edge.
retry_fundamental_rule(error_edge(Cat,Const,[],Lpos,Rpos,Num)):-
  error_edge(Act_Cat,Act_Const,[Cat|Rest],Act_lpos,Lpos,_),
```

```
  make_error_edge(error_edge(Act_Cat,[Num|Act_Const],Rest,Act_lpos,Rpos)),
  fail.

% This works for the case when an active edge meets an inactive edge.
retry_fundamental_rule(error_edge(A_Cat,A_Const,[FirstExpect|New_Expects],A_lpos,A_rpos,_)):-
  edge(FirstExpect,Const,[],A_rpos,Rpos,Num),
  make_error_edge(error_edge(A_Cat,[Num|A_Const],New_Expects,A_lpos,Rpos)),
  fail.

% This works for the case when an active edge meets an inactive edge.
retry_fundamental_rule(error_edge(A_Cat,A_Const,[FirstExpect|New_Expects],A_lpos,A_rpos,_)):-
  error_edge(FirstExpect,Const,[],A_rpos,Rpos,Num),
  make_error_edge(error_edge(A_Cat,[Num|A_Const],New_Expects,A_lpos,Rpos)),
  fail.
retry_fundamental_rule(_).

% Bottom-up rule
retry_find_rules(error_edge(Cat,Const,[],Lpos,Rpos,Num)) :-
  rule(Lhs,[Cat|Rest]),
  error_lookahead(Rest,Rpos,Categories,Final),
  conc(Categories,[Num],NewCategories),
  make_error_edge(error_edge(Lhs,NewCategories,[],Lpos,Final)),
  fail.
retry_find_rules(_).

find_error_rule(edge(Cat,_,[],Lpos,Rpos,Num),Wordlist):-
   find_missing_error_rule(edge(Cat,_,[],Lpos,Rpos,Num),Wordlist).
find_error_rule(edge(Cat,_,[],Lpos,Rpos,Num),Wordlist):-
   find_extra_error_rule(edge(Cat,_,[],Lpos,Rpos,Num),Wordlist).

find_missing_error_rule(edge(Cat,_,[],Lpos,Rpos,N),Wordlist):-
   missing_error_rule(Lhs,[Cat|Rest]),
   error_lookahead(Rest,Rpos,Categories,Final),
   conc(Categories,[N],NewCategories),
   make_initial_error_edge(error_edge(Lhs,NewCategories,[],Lpos,Final)).

find_extra_error_rule(edge(Cat,_,[],Lpos,Rpos,N),Wordlist):-
   extra_error_rule(Lhs,[Cat|Rest]),
   extra_error_lookahead(Rest,Rpos,Categories,Final),
   conc(Categories,[N],NewCategories),
   make_initial_error_edge(error_edge(Lhs,NewCategories,[],Lpos,Final)).


% Perform forward scan
error_lookahead(Rhs,Rpos,Categories,Final):-
  error_lookahead1(Rhs,Rpos,[],Categories,Final).

error_lookahead1([F|R],Rpos,AnswerSoFar,Categories,Final):-
  edge(F,_,[],Rpos,NextPos,N),
  error_lookahead1(R,NextPos,[N|AnswerSoFar],Categories,Final).
error_lookahead1([],Rpos,Answer,Answer,Rpos).


extra_error_lookahead(Rhs,Rpos,Categories,Final):-
  extra_error_lookahead1(Rhs,Rpos,[],Categories,Final).

extra_error_lookahead1([cat|Rest],Rpos,AnswerSoFar,Categories,Final):-
edge(_,_,[],Rpos,NextPos,N),
extra_error_lookahead1(Rest,NextPos,[N|AnswerSoFar],Categories,Final).
extra_error_lookahead1([First|Rest],Rpos,AnswerSoFar,Categories,Final):-
not(First == cat),
edge(First,_,[],Rpos,NextPos,N),
```

```
extra_error_lookahead1(Rest,NextPos,[N|AnswerSoFar],Categories,Final).
extra_error_lookahead1([],Rpos,Categories,Categories,Rpos).

find_lexical_error_rule(Word,Rest,Wordlist):-
   find_spell_error_rule(Word,Rest,Wordlist).
find_lexical_error_rule(Word,Rest,Wordlist):-
   find_vform_error_rule(Word,Rest,Wordlist).
find_lexical_error_rule(Word,Rest,Wordlist):-
   find_agree_error_rule(Word,Rest,Wordlist).
find_lexical_error_rule(Word,Rest,Wordlist):-
   find_lex_error_rule(Word,Rest,Wordlist).
find_lexical_error_rule(Word,Rest,Wordlist):-
   find_poss_error_rule(Word,Rest,Wordlist).
find_lexical_error_rule(Word,Rest,Wordlist):-
   find_prep_error_rule(Word,Rest,Wordlist).
find_lexical_error_rule(Word,Rest,Wordlist):-
   find_indef_error_rule(Word,Rest,Wordlist).
find_lexical_error_rule(Word,Rest,Wordlist):-
   find_aux_error_rule(Word,Rest,Wordlist).
find_lexical_error_rule(Word,Rest,Wordlist):-
   find_infprep_error_rule(Word,Rest,Wordlist).
find_lexical_error_rule(Word,Rest,Wordlist):-
   find_conj_error_rule(Word,Rest,Wordlist).
find_lexical_error_rule(Word,Rest,Wordlist):-
   find_adj_error_rule(Word,Rest,Wordlist).

find_spell_error_rule(Word,Rest,Wordlist):-
   spell_error_rule(NonTerminal,[[Word]]),
   make_initial_error_edge(error_edge(NonTerminal,Word,[],[Word|Rest],Rest)).
find_vform_error_rule(Word,Rest,Wordlist):-
   vform_error_rule(NonTerminal,[[Word]]),
   make_initial_error_edge(error_edge(NonTerminal,Word,[],[Word|Rest],Rest)).
find_agree_error_rule(Word,Rest,Wordlist):-
   agree_error_rule(NonTerminal,[[Word]]),
   make_initial_error_edge(error_edge(NonTerminal,Word,[],[Word|Rest],Rest)).
find_lex_error_rule(Word,Rest,Wordlist):-
   lex_error_rule(NonTerminal,[[Word]]),
   make_initial_error_edge(error_edge(NonTerminal,Word,[],[Word|Rest],Rest)).
find_poss_error_rule(Word,Rest,Wordlist):-
   poss_error_rule(NonTerminal,[[Word]]),
   make_initial_error_edge(error_edge(NonTerminal,Word,[],[Word|Rest],Rest)).
find_prep_error_rule(Word,Rest,Wordlist):-
   prep_error_rule(NonTerminal,[[Word]]),
   make_initial_error_edge(error_edge(NonTerminal,Word,[],[Word|Rest],Rest)).
find_indef_error_rule(Word,Rest,Wordlist):-
   indef_error_rule(NonTerminal,[[Word]]),
   make_initial_error_edge(error_edge(NonTerminal,Word,[],[Word|Rest],Rest)).
find_aux_error_rule(Word,Rest,Wordlist):-
   aux_error_rule(NonTerminal,[[Word]]),
   make_initial_error_edge(error_edge(NonTerminal,Word,[],[Word|Rest],Rest)).
find_infprep_error_rule(Word,Rest,Wordlist):-
   infprep_error_rule(NonTerminal,[[Word]]),
   make_initial_error_edge(error_edge(NonTerminal,Word,[],[Word|Rest],Rest)).
find_conj_error_rule(Word,Rest,Wordlist):-
   conj_error_rule(NonTerminal,[[Word]]),
   make_initial_error_edge(error_edge(NonTerminal,Word,[],[Word|Rest],Rest)).
find_adj_error_rule(Word,Rest,Wordlist):-
   adj_error_rule(NonTerminal,[[Word]]),
   make_initial_error_edge(error_edge(NonTerminal,Word,[],[Word|Rest],Rest)).

member(X,[X|_]).
member(X,[_|R]):- member(X,R).
```

```
% Adds an edge to a forest during recovery module
add_to_forest_error(Edge,Root):-
  Edge =.. List,
  edgenumgen(N),
  conc(List,[N],Arity_six_list),
  Real_edge =.. Arity_six_list,
  retract(forest_error(Root,AList)),!,
  conc([N],AList,NewList),
  assert(forest_error(Root,NewList)),
  assert(packed_error(Real_edge)).

add_to_forest_error(Edge,Root):-
  Edge =.. List,
  edgenumgen(N),
  conc(List,[N],Arity_six_list),
  Real_edge =.. Arity_six_list,
  assert(forest_error(Root,[N])),
  assert(packed_error(Real_edge)),!.

% This procedure assigns a unique number to an edge and adds it to the top
% of the agenda.
make_error_edge(Edge) :-
  not(redundant(Edge)),
  Edge =.. List,
  edgenumgen(N),
  conc(List,[N],Arity_six_list),
  Real_edge =.. Arity_six_list,
  asserta(to_process(Real_edge)),!.

make_error_edge(_).

make_initial_error_edge(Edge) :-
  not(redundant(Edge)),
  Edge =.. List,
  edgenumgen(N),
  conc(List,[N],Arity_six_list),
  Real_edge =.. Arity_six_list,
  asserta(to_process(Real_edge)),
  asserta(propose(Real_edge)),!.

make_initial_error_edge(_).
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

%%
%% Utilities
%%

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% Printing

% Enumerating parses found during recovery phase
recover_print_trees(Symbol,Wordlist):-
 error_edge(Symbol,_,[],Wordlist,[],Num),
 increment_parse_count,
 print_tree(Num,0),nl,
 fail.

% Enumerating parses found during normal phase
print_trees(Symbol,Wordlist):-
 edge(Symbol,_,[],Wordlist,[],Num),
 increment_parse_count,
```

```
 %write('new parse'),nl,
 print_tree(Num,0),nl,
 fail.

print_trees(_,_):-!.

recover_print_trees(_,_):-!.

% A null tree prints nothing.
print_tree_list([],_).


% Print first the constituents which were found first.
print_tree_list([Const_head|Rest],Depth) :-
  print_tree_list(Rest,Depth),
  print_tree(Const_head,Depth).


% Given an index to an edge, finds its category symbol, prints tree under
% it. If appropriate, finds inactive edges of the same category in the same
% place in chart and prints these too.
print_tree(Num,Depth) :-
  edge(Cat,[Const_head|Rest],_,_,_,Num),!,
  assert(printed(Num)),
  find_others(Num,NumList),
  space_it(Cat,Depth),
  print_tree_list(Rest,Depth + 3),
  print_tree(Const_head,Depth+3),
  print_tree_list(NumList,Depth).

print_tree(Num,Depth) :-
  error_edge(Cat,[Const_head|Rest],_,_,_,Num),!,
  find_others(Num,NumList),
  space_it(Cat,Depth),
  print_tree_list(Rest,Depth + 3),
  print_tree(Const_head,Depth+3),
  print_tree_list(NumList,Depth).

print_tree(Num,Depth) :-
  packed(edge(Cat,[Const_head|Rest],_,_,_,Num)),!,
  space_it(Cat,Depth),
  print_tree_list(Rest,Depth + 3),
  print_tree(Const_head,Depth+3).

print_tree(Num,Depth) :-
  packed_error(error_edge(Cat,[Const_head|Rest],_,_,_,Num)),!,
  space_it(Cat,Depth),
  print_tree_list(Rest,Depth + 3),
  print_tree(Const_head,Depth+3).

% ... unless the constituent is simply a terminal symbol.
print_tree(Num,Depth) :-
  edge(Cat,Symbol,_,_,_,Num),
  terminal(Symbol),!,
  space_it(Cat,Depth),
  write(' '),
  write(Symbol).

% ... unless the constituent is simply a terminal symbol.
print_tree(Num,Depth) :-
  error_edge(Cat,Symbol,_,_,_,Num),
  terminal(Symbol),!,
```

```
  space_it(Cat,Depth),
  write(' '),
  write(Symbol).

% Finds all packed edges
find_others(N,List):-
  forest(N,EdgeList),
  forest_error(N,ErrorEdgeList),
  conc(ErrorEdgeList,EdgeList,List),!,
  increment_parse_list(List).

find_others(N,List) :-
  forest(N,List),!,
  increment_parse_list(List).

find_others(N,List) :-
  forest_error(N,List),!,
  increment_parse_list(List).

find_others(N,[]).

%Increments parse counts for all unpacked edges
increment_parse_list([]).
increment_parse_list([_|R]):- increment_parse_count,
                             increment_parse_list(R).

% Prints the node name at a suitably indented position.
% Invoked for leaves of tree.
space_it(Node_Name,Spaces) :-
 nl,
 tab(Spaces),
 write(Node_Name).

%% other stuff

%Checks if an edge or error edge needs to be created
redundant(edge(Cat,Const,X,Lpos,Rpos)) :-
  to_process(edge(Cat,Const,X,Lpos,Rpos,_)),!.
redundant(edge(Cat,Const,X,Lpos,Rpos)) :-
  edge(Cat,Const,X,Lpos,Rpos,_),!.
redundant(edge(Cat,Const,X,Lpos,Rpos)) :-
  packed(edge(Cat,Const,X,Lpos,Rpos,_)),!.
redundant(edge(Cat,Const1,[],Lpos,Rpos)) :-
  edge(Cat,Const2,[],Lpos,Rpos,N),
  chain(N,Const1),!.
redundant(edge(Cat,Const1,[],Lpos,Rpos)) :-
  to_process(edge(Cat,Const2,[],Lpos,Rpos,N)),
  chain(N,Const1),!.
redundant(edge(Cat,Const1,[],Lpos,Rpos)):-
  to_process(edge(Cat,Const2,[],Lpos,Rpos,N)),
  add_to_forest(edge(Cat,Const1,[],Lpos,Rpos),N),!.
redundant(edge(Cat,Const1,[],Lpos,Rpos)):-
  edge(Cat,Const2,[],Lpos,Rpos,N),
  add_to_forest(edge(Cat,Const1,[],Lpos,Rpos),N),!.


redundant(error_edge(Cat,Const,X,Lpos,Rpos)) :-
  propose(error_edge(Cat,Const,X,Lpos,Rpos,_)),!.
redundant(error_edge(Cat,Const,X,Lpos,Rpos)) :-
  edge(Cat,Const,X,Lpos,Rpos,_),!.
redundant(error_edge(Car,Const,X,Lpos,Rpos)) :-
  packed(edge(Cat,Const,X,Lpos,Rpos,_)),!.
```

```
redundant(error_edge(Cat,Const,X,Lpos,Rpos)) :-
  error_edge(Cat,Const,X,Lpos,Rpos,_),!.
redundant(error_edge(Cat,Const,X,Lpos,Rpos)) :-
  to_process(error_edge(Cat,Const,X,Lpos,Rpos,_)),!.
redundant(error_edge(Cat,Const,X,Lpos,Rpos)) :-
  packed_error(error_edge(Cat,Const,X,Lpos,Rpos,_)),!.
redundant(error_edge(Cat,Const1,[],Lpos,Rpos)) :-
  edge(Cat,Const2,[],Lpos,Rpos,N),
  chain(N,Const1),!.
redundant(error_edge(Cat,Const1,[],Lpos,Rpos)) :-
  error_edge(Cat,Const2,[],Lpos,Rpos,N),
  chain(N,Const1),!.
redundant(error_edge(Cat,Const1,[],Lpos,Rpos)) :-
  to_process(error_edge(Cat,Const2,[],Lpos,Rpos,N)),
  chain(N,Const1),!.
redundant(error_edge(Cat,Const1,[],Lpos,Rpos)) :-
  edge(Cat,Const2,[],Lpos,Rpos,N),
  add_to_forest_error(error_edge(Cat,Const1,[],Lpos,Rpos),N),!.
redundant(error_edge(Cat,Const1,[],Lpos,Rpos)) :-
  error_edge(Cat,Const2,[],Lpos,Rpos,N),
  add_to_forest_error(error_edge(Cat,Const1,[],Lpos,Rpos),N),!.
redundant(error_edge(Cat,Const1,[],Lpos,Rpos)) :-
  to_process(error_edge(Cat,Const2,[],Lpos,Rpos,N)),
  add_to_forest_error(error_edge(Cat,Const1,[],Lpos,Rpos),N),!.


chain(Node1,List):-
  edge(_,Daughters,_,_,_,Node2),
  member(Node1,Daughters),
  member(Node2,List),!.
chain(Node1,List):-
  error_edge(_,Daughters,_,_,_,Node2),
  member(Node1,Daughters),
  member(Node2,List),!.
chain(Node1,List):-
  to_process(error_edge(_,Daughters,_,_,_,Node2)),
  member(Node1,Daughters),
  member(Node2,List),!.


% Checks if a symbol is a terminal symbol.
terminal([[X]]) :- !.
terminal(X) :- rule(_,[[X]]).

% Concatenates two lists together.
conc([],A,A):-!.
conc([A|Rest],List,[A|New_list]) :- conc(Rest,List,New_list).


%%%%%%%%%%%%%%%%%%%%%%%%%%
%  Input routine from Pereira & Shieber:
read_sent(Words) :-
  get0(Char),             %prime lookahead
  read_sent(Char, Words). %get words


% Period terminates input.
read_sent(C,[]) :- period(C),!.


% Certain stuff is to be ignored.
ignore(C) :- space(C).
```

```
ignore(C) :- newline(C).


% Stuff is ignored.
read_sent(C,Words) :-
  ignore(C),!,
  get0(Char),
  read_sent(Char, Words).


% Everything else starts a word.
read_sent(Char, [Word|Words]) :-
  read_word(Char, Chars, Next),    %get the word
  name(Word, Chars),               %pack the characters into an atom
  read_sent(Next, Words).          %get some more words


% Space, period, and newline end a word.
read_word(C,[],C) :- space(C),!.
read_word(C,[],C) :- period(C),!.
read_word(C,[],C) :- newline(C),!.


% All other chars are added to the list.
read_word(Char, [Char|Chars], Last) :-
  get0(Next),
  read_word(Next, Chars, Last).


space(32).
newline(10).
period(46).

%%Resets counter for edge IDS
set_edge_num_to_one :- retract(edge_num(N)),assert(edge_num(1)).
edge_num(1).

% Generates unique numbers for edge IDs.
edgenumgen(N) :- retract(edge_num(N)), N1 is N + 1, assert(edge_num(N1)), !.


%%Resets counter for node names
set_node_num_to_one :- retract(node_num(N)),assert(node_num(1)).
node_num(1).
% Generates unique numbers for node names
nodenumgen(N) :- retract(node_num(N)), N1 is N + 1, assert(node_num(N1)), !.

%%Resets counter for node names
set_cycle_count_to_zero :- retract(cycle_count(N)),assert(cycle_count(0)).
cycle_count(0).
% Generates unique numbers for node names
increment_cycle_count:- retract(cycle_count(N)), N1 is N + 1, assert(cycle_count(N1)), !.

%%Resets counter for node names
set_parse_count_to_zero :- retract(parse_count(_)),assert(parse_count(0)).
parse_count(0).
% Generates unique numbers for node names
increment_parse_count:- retract(parse_count(N)), N1 is N + 1, assert(parse_count(N1)), !.
```

The following sentences were used as test data for the error grammar approach described in Chapter 5. The sentences were collected in the same way as the sentences in the main error corpus, and they are presented in the same way here.

(C.1) *A.* Not so at the Coolquay Lodge, everything was **is** perfect amounts. *B.* Not so at the Coolquay Lodge, everything was **in** perfect amounts.
(CityWide News, p.12, 10th December 2003, Vol.11, Issue 24, Northeast edition)

(C.2) *A.* You can make use of TeX related packages that are already installed in **you** network neighbourhood. *B.* You can make use of TeX related packages that are already installed in **your** network neighbourhood.
(MiKTeX Setup Wizard 2.4.1477, 4/2/04)

(C.3) *A.* Yes I have to **in** to trinity for one thing so i can see you at one. *B.* Yes I have to **go in** to trinity for one thing so i can see you at one. *C.* Yes I have to **go** to trinity for one thing so i can see you at one.
(personal text message, 4/2/04)

(C.4) *A.* I will text u later to let u know how tina got in exam. *B.* I will text u later to let u know how tina got **on** in exam.
(personal text message, 28/1/04)

(C.5) *A.* Uses do-while loops to **insure** that numbers are different. *B.* Uses do-while loops to **ensure** that numbers are different.
(comment in student programming assignment, 6/2/04)

(C.6) *A.* When it is the main method or when you want to take something **from that could be in** the main method and put it into a separate method in the application class. *B.* When it is the main method or when you want to take something **that could be in** the main method and put it into a separate method in the application class. *C.* When it is the main method or when you want to take something **from** the main method and put it into a separate method in the application class.
(lecture notes, February 2004)

(C.7) *A.* This class **is represents** a Lotto Draw Generator. *B.* This class **is** a Lotto Draw Generator. *C.* This class **represents** a Lotto Draw Generator.
(draft of own writing, February 2004)

(C.8) *A.* The **programs finals** requirement is to print statements to the screen informing the user of their winnings. *B.* The **program's final** requirement is to print statements to the screen informing the user of their winnings.
(student assignment, 13/2/04)

(C.9) *A.* One of them also **match** the bonus number. *B.* One of them also **matches** the bonus number.
(student assignment, 16/2/04)

(C.10) *A.* In this paper an evaluation method is presented which is similar to the above yet **is which** specifically designed to measure a parser's accuracy in the face of input which is ungrammatical. *B.* In this paper an evaluation method is presented which is similar to the above yet **which is** specifically designed to measure a parser's accuracy in the face of input which is ungrammatical.
(draft of own writing, 20/2/04)

(C.11) *A.* It doesn't make sense to release anything at the moment but hopefully when the Mark McCabe thing comes **our** I'll be able to release stuff. *B.* It doesn't make sense to release anything at the moment but hopefully when the Mark McCabe thing comes **out** I'll be able to release stuff.
(University Record, 19/1/04, p.19)

(C.12) *A.* Results are shown for features selected by **a a** single decision tree, and an ensemble of 11 decision tree models based on different feature subspaces as described above. *B.* Results are shown for features selected by **a** single decision tree, and an ensemble of 11 decision tree models based on different feature subspaces as described above.
(Toutanova and Manning 2002)

(C.13) *A.* When a sentence **is** identified as ungrammatical is added to Corpus1 the sentence is corrected. *B.* When a sentence is identified as ungrammatical **and** is added to Corpus1 the sentence is corrected. *C.* When a sentence identified as ungrammatical is added to Corpus1 the sentence is corrected.
(draft of own writing, 22/1/04)

(C.14) *A.* I have a work lunch tomorrow and I'm meeting Pat another day and **today** I have to get some stuff in town **today**. *B.* I have a work lunch tomorrow and I'm meeting Pat another day and **today** I have to get some stuff in town. *C.* I have a work lunch tomorrow and I'm meeting Pat another day and I have to get some stuff in town **today**.
(personal email, 3/2/04)

(C.15) *A.* The Every Day Learning group are looking **from** an experienced Java programmer to do some part-time work. *B.* The Every Day Learning group are looking **for** an experienced Java programmer to do some part-time work.
(personal email, 5/2/04)

(C.16) *A.* The rest might find it **as** a useful reminder. *B.* The rest might find it a useful reminder.
(http://bobi.ifs.hr/ mpinter/miktex.html, 4/2/04)

(C.17) *A.* Please note, some basic knowledge of **computer** is necessary. *B.* Please note, some basic knowledge of **computers** is necessary.
(http://bobi.ifs.hr/ mpinter/miktex.html, 4/2/04)

(C.18) *A.* The superfluous "to" has little **affect** on the other constituents in the sentence. *B.* The superfluous "to" has little **effect** on the other constituents in the sentence.
(draft of own writing, 6/2/04)

(C.19) *A.* In Java, errors are handled by special **kind** of classes known as exception classes. *B.* In Java, errors are handled by special **kinds** of classes known as exception classes.
(lecture notes, 6/2/04)

(C.20) *A.* Their research focuses on the use **off** artificial intelligence and machine learning techniques for Web search. *B.* Their research focuses on the use **of** artificial intelligence and machine learning techniques for Web search.
(email newsletter, 3/2/04)

(C.21) *A.* The program starts by printing a welcome message giving the user 2 options: to select their own numbers, or to have **to** numbers generated for them. *B.* The program starts by printing a welcome message giving the user 2 options: to select their own numbers, or to have **the** numbers generated for them.
(student assignment, 18/2/04)

(C.22) *A.* It has a method to calculate the winnings depending on the numbers **in on** the ticket passed to it. *B.* It has a method to calculate the winnings depending on the numbers **on** the ticket passed to it.
(student assignment, 18/2/04)

(C.23) *A.* To test the program, I inputted random numbers, and used the quick pick option, and it appeared to work properly, but this method only really showed whether the program correctly identified that a ticket did not win a prize, as **the** it is very unlikely to draw 6 numbers which match the six numbers entered. *B.* To test the program, I inputted random numbers, and used the quick pick option, and it appeared to work properly, but this method only really showed whether the program correctly identified that a ticket did not win a prize, as it is very unlikely to draw 6 numbers which match the six numbers entered.
(student assignment, 18/2/04)

(C.24) *A.* The Object r has **being** created using the operator NEW. *B.* The Object r has **been** created using the operator NEW.
(student assignment, 18/2/04)

(C.25) *A.* Thanks for your help **in in** this. *B.* Thanks for your help **in** this.
(personal email, 19/2/04)

(C.26) *A.* It promises to be a special evening, in which you can recover from the madness of Patrick's Day in **a** peace. *B.* It promises to be a special evening, in which you can recover from the madness of Patrick's Day in peace.
(personal email, 24/2/04)

(C.27) *A.* Some tags, e.g. @param and @return, are only used with methods since **it they** only apply to methods. *B.* Some tags, e.g. @param and @return, are only used with methods since they only apply to methods.
(lecture notes, 6/4/02)

(C.28) *A.* **Inherently robustness** *B.* **Inherently robust** *C.* **Inherent robustness**
(draft of own writing, 10/4/04)

(C.29) *A.* This class takes the French word which has been inputted and prints out **it's** translation in English. *B.* This class takes the French word which has been inputted and prints out **its** translation in English.
(student assignment, 4/4/04)

(C.30) *A.* This int determines which file is to used throughout the program depending on information provided by the user. *B.* This int determines which file is to **be** used throughout the program depending on information provided by the user.
(student assignment, 14/4/04)

(C.31) *A.* This yields a measure from which one can predict the average number of (undisambiguated) parses for a sentence **of of** length n words in a particular corpus and analysed with that grammar/parser. *B.* This yields a measure from which one can predict the average number of (undisambiguated) parses for a sentence **of** length n words in a particular corpus and analysed with that grammar/parser.
(Carroll and Briscoe 1996)

(C.32) *A.* Her comments come after the Beckhams instructed lawyers to look into the allegations **surround** his private life. *B.* Her comments come after the Beckhams instructed lawyers to look into the allegations **surrounding** his private life. *C.* Her comments come after the Beckhams instructed lawyers to look into the allegations **which surround** his private life.
(http://www.ananova.com/entertainment/story/sm_922790.html?menu=, 15/4/04)

(C.33) *A.* If the word is not present in the string a question mark is returned **is returned**. *B.* If the word is not present in the string a question mark is returned.
(student assignment, 15/4/04)

(C.34) *A.* This will indicate words not in the **programs** vocabulary. *B.* This will indicate words not in the **program's** vocabulary.
(student assignment, 15/4/04)

(C.35) *A.* Two Strings are passed in as parameters, one intended to **equal to** an individual word (the gender variable), the other intended to represent a filename. *B.* Two Strings are passed in as parameters, one intended to **be equal to** an individual word (the gender variable), the other intended to represent a filename. *C.* Two Strings are passed in as parameters, one intended to **equal** an individual word (the gender variable), the other intended to represent a filename.
(student assignment, 15/4/04)

(C.36) *A.* Following on from **a** emotional reunion **for** the lads insisted they would not get back together as a band. *B.* Following on from **an** emotional reunion **for** the lads **they** insisted they would not get back together as a band. *C.* Following on from **an** emotional reunion the lads insisted they would not get back together as a band.
(http://www.ananova.com/entertainment/story/sm_923889.html?menu=, 16/4/04)

(C.37) *A.* Suddenly I felt someone take my hand and squeeze it gently as **it** to reassure me that I shouldn't be overawed by all the hype. *B.* Suddenly I felt someone take my hand and squeeze it gently as **if** to reassure me that I shouldn't be overawed by all the hype.
(http://www.ananova.com/entertainment/story/sm_929999.html?menu=entertainment.celebrities, 22/4/04)

(C.38) *A.* Parallel texts have become a vital **resources** for efficiently deriving many multi-lingual text processing tools. *B.* Parallel texts have become a vital **resource** for efficiently deriving many multi-lingual text processing tools. *C.* Parallel texts have become vital **resources** for efficiently deriving many multi-lingual text processing tools.
(email newsletter, 22/4/04)

(C.39) *A.* The dependency-based method allows us to evaluate not only the overall performance of the parser, but also its performance with respect different grammatical relationships and phenomena. *B.* The dependency-based method allows us to evaluate not only the overall performance of the parser, but also its performance with respect **to** different grammatical relationships and phenomena.
(Lin 1998)

(C.40) *A.* If a dependency in the answer is only **consider** to be correct if it has the same label as the head, the recall will be 33.3% and the precision will be 50%. *B.* If a dependency in the answer is only **considered** to be correct if it has the same label as the head, the recall will be 33.3% and the precision will be 50%.
(Lin 1998)

(C.41) *A.* Most probabilistic parsing research – including, for example, work **by by** Collins (1997), Charniak (1997), and Manning and Carpenter (2000) – is based on branching process models (Harris, 1963). *B.* Most probabilistic parsing research – including, for example, work **by** Collins (1997), Charniak (1997), and Manning and Carpenter (2000) – is based on branching process models (Harris, 1963).
(Oepen *et al.* 2002)

(C.42) *A.* The proposed method is applicable to most NLP **system**, but parsers will be used here for the clarity of exposition. *B.* The proposed method is applicable to most NLP **systems**, but parsers will be used here for the clarity of exposition.
(Bigert *et al.* 2004)

(C.43) *A*. This mail contains a list of entries that were submitted to the HPSG Bibliography during the last two **month**. *B*. This mail contains a list of entries that were submitted to the HPSG Bibliography during the last two **months**.
(email newsletter, 2/6/04)

(C.44) *A*. Successful exploitation requires that a user can be tricked into following a link or **view** a malicious HTML document.*B*. Successful exploitation requires that a user can be tricked into following a link or **viewing** a malicious HTML document.
(email newsletter, 11/6/04)

Appendix D

Tree Transformation Code

This appendix contains the code which I developed to transform a grammatical *corpus2* sentence phrase structure parse tree into the structurally identical parse tree(s) for the corresponding ungrammatical *corpus1* sentence. This is carried out so that test parses for a *corpus1* sentence can be compared to a parse tree with the same yield. Thus they are indirectly rather than directly compared to the reference grammatical sentence parse. The code was used in the evaluation of the two broad-coverage probabilistic parsers (Collins 1999, Charniak 2000) using the evaluation method presented in Chapter 6.

The code works by assuming that an ungrammatical sentence which has the same length as its correction contains a substitution error(s), that an ungrammatical sentence which has more words than its correction contains a deletion error(s), and that an ungrammatical sentence which contains fewer words than its correction contains an insertion error(s). This will be the correct assumption to make for all the *corpus1*/*corpus2* sentence pairs with a minimum edit distance of one (the majority according to the analysis carried out in Section 5.2), and for those containing composite errors involving sequences of deleted or inserted words. However, it is not guaranteed to work on sentences containing other types of composite errors, for example, the ungrammatical sentence, *The following roadmap for the has been derived derived for this presentation* (A.797) which has the same number of words as its corrected version but the errors are not corrected by substitution but by insertion and deletion. Thus, the parses of all *corpus2* sentences with a minimum edit distance of more than one from their *corpus1* counterpart were manually checked and manually transformed if necessary. This is not an ideal solution but the transformation could easily be carried out fully automatically if, for each *corpus1*/*corpus2* sentence pair, the exact transformations that would need to be applied to transform the parse of the *corpus2* sentence into a perfect parse of the *corpus1* sentence were explicitly specified and available to the code.

```
import java.util.Stack;
import java.util.ArrayList;
import java.io.PrintWriter;
class ParseTree
{
  private String category;
  private ArrayList daughters;
  private String word; /*every tree will have a null value for either
                         daughters or word depending on whether its
                         category is a pre-terminal
                        */
  private static int count;        /*these are used to keep track of how
                                    * many words and subtrees
                                    * in a tree have been
                                    * checked.
                                    */

  private static int treeCount;


  /* Usual constructor
     */
    public ParseTree(String word, String category, ArrayList daughters)
    {
        this.word = word;
        this.category = category;
        this.daughters = daughters;
    }

  /* Takes a Penn-Treebank like parse tree in the form of a string
   * and creates a new parse tree object
   */
  public ParseTree(String s)
    {
      int index =1;
      int forwardIndex = 0;
      Stack stack = new Stack();

      //get the category symbol first of all
      category = "";
```

```
//the category is always in second position after the bracket
while (s.charAt(index) != ' ')
 {
  category += s.charAt(index);
  index++;
 }


//now get the daughters or the word
//the index is now at the space after the category symbol
//what comes after the space will either be an opening bracket or
//a word
  while (s.charAt(index) == ' ')
   {
     index++;
   }
  //if the next symbol is not an opening bracket then it is a word
  if (s.charAt(index) != '(')
  {
    word = "";
    word += s.charAt(index);
    index++;
    while (s.charAt(index) != ')')
      {
        word += s.charAt(index);
        index++;
      }
  }
//if the word is an opening bracket then it is the first daughter
else
{
    daughters = new ArrayList();
    while (index < s.length())
     {
         if (s.charAt(index) == '(')
           {
     //find its closing bracket - when you have it
             //create a new parse tree with this constituent
             //and add it to the array of daughters

            forwardIndex = index;
            stack.push(new Integer(index));
            forwardIndex++;
            while (! stack.empty())
            {
               if (s.charAt(forwardIndex) == '(')
               {
                 stack.push(new Integer(forwardIndex));
                 forwardIndex++;
               }
              else if (s.charAt(forwardIndex) == ')')
              {
                 stack.pop();
                 forwardIndex++;
              }
             else
              {
                forwardIndex++;
              }
         }
    //when we get this far the forwardIndex is one after the relevant
    //closing bracket
```

```
         daughters.add(new ParseTree(s.substring(index,forwardIndex)));
         index = forwardIndex;
               }
              else
                {
                  index++;
                }
        }
     }
 }

//getter and setter methods
  public String getCategory()
  {
     return category;
  }

  public void setCategory(String c)
  {
     category = c;
  }

 public ArrayList getDaughters()
  {
     return daughters;
  }

 public ParseTree getDaughter(int i)
  {
     return (ParseTree) daughters.get(i);
  }

 public void setDaughters(ArrayList d)
  {
     daughters = d;
  }

 public void addDaughter(ParseTree p)
  {
    daughters.add(p);
  }

 public String getWord()
  {
    return word;
  }

 public void setWord(String w)
  {
     word = w;
  }

/* return a parse tree as a Penn-Treebank like string
 */
public String toString()
  {
    String treeInfo="("+ category + " ";
    ParseTree daughter = null;
    //check if category is a pre-terminal
    if (word != null)
      {
          treeInfo += word + ")";
```

```
          return treeInfo;
      }
    else
      {
        //the category is not a preterminal
        //go through all the daughters and append
        //each of their information strings to treeInfo
        for (int i=0; i < daughters.size(); i++)
          {
            daughter = this.getDaughter(i);
            treeInfo += daughter.toString() + " ";
          }
         treeInfo += ")";
        return treeInfo;
      }
  }

  /* method to return the set of words dominated by this tree
 */
public ArrayList getWords()
 {
    ArrayList words = new ArrayList();
    ParseTree daughter = null;
    //if the category is a pre-terminal then there will only be one
    // word
    if (word != null)
      {
         words.add(word);
         return words;
      }
    else
     {
        //get all the words returned by each daughter and add to
        //list
        for (int i=0; i < daughters.size(); i++)
          {
            daughter = this.getDaughter(i);
            words.addAll(daughter.getWords());
          }
         return words;
      }
  }

 /* method which transforms a tree so that it dominates
 * the same set of words passed in as a parameter
 */
public ArrayList transform(ArrayList words)
  {
    ArrayList list = null;
    ArrayList originals = this.getWords();
    if (words.size() == originals.size())
     {
        //if the word sets are the same size, then
        //the method to replace words needs to be called
        list = this.replaceWords(words);
      }
    else if (words.size() < originals.size())
    {
        //if the current tree has more elements
        //then the extra elements need to be removed
        list = this.deleteWords(words);
    }
```

```
    else if (words.size() > originals.size())
    {
        //if the current tree has less elements
        //than the missing elements need to be added
        list = this.addWords(words);
    }
    return list;
  }

/* method which takes in a list of words of the same length as
 * the words dominated by the current parse tree
 * for all words which differ between the two lists, the words
 * in the current tree are replaced by the words in list taken
 * in as a parameter.
 */
public ArrayList replaceWords(ArrayList newWords)
 {
     ArrayList list = new ArrayList();
     ArrayList originalWords = this.getWords();
     String original;
     String newWord;
     ParseTree newTree = this.copy();
     for (int i=0; i < originalWords.size(); i++)
      {
         original = (String) originalWords.get(i);
         newWord = (String) newWords.get(i);
         //if the two words aren't the same then a replacement
         //must take place
         if (!original.equals(newWord))
           {
             newTree = newTree.replace(original,newWord,i);
           }
       }
     list.add(newTree);
     return list;
  }

 /* method which replaces one word in the tree at a certain position
 * with a new word.
 */
public ParseTree replace(String original, String newWord, int pos)
  {
     count = 0;
     ParseTree copy = this.copy();
     copy.replace2(original, newWord, pos);
     return copy;
  }

public boolean replace2(String original, String newWord, int pos)
  {
    //see if the category is a pre-terminal
    //if it is check the word
    if (word != null)
      {
         //if the word is equal to the original word and is in the
        //correct position
        if (word.equals(original) && count == pos)
          {
            //change the value of word
            word = newWord;
            return true;
          }
```

```
        else
          {
            //increment the count and return false
            count++;
            return false;
          }
       }
     else
       {
         //check the words of the daughters
         ParseTree daughter = null;
         boolean found = false;
         for (int i=0; i < daughters.size() && !found; i++)
          {
             daughter = this.getDaughter(i);
             found = daughter.replace2(original,newWord,pos);
           }
         return found;
       }
  }

/* method which takes a set of words which contain
 * all the words in the current tree, plus extra ones
 * this method deletes the subtrees containing these
 * extra words from the current tree
 */
public ArrayList deleteWords(ArrayList words)
{
    ArrayList list = new ArrayList();
    ArrayList originalWords = this.getWords();
    String original;
    String newWord;
    int pos=0;
    ParseTree newTree = this.copy();
    for (int i=0,j=0; i < originalWords.size();i++)
     {
         original = (String) originalWords.get(i);
         newWord = (String) words.get(j);
         //if the two words aren't the same then the word
         //in the current tree needs to be deleted
         if (!original.equals(newWord))
           {
                     newTree = newTree.delete(original,pos);
           }
         else
         {
             j++;
             pos++;
         }
      }
    list.add(newTree);
     return list;
}

 /* method which takes a word and deletes the subtree dominating this word
 * from the tree
 * second parameter is the position of the word in the word set
 */
public ParseTree delete(String deleteWord, int pos)
{
    count=0; //keeps track of how many words have been encountered
    ParseTree copy = this.copy();
```

```
    copy.delete2(deleteWord,pos);
    //now go through the tree and actually delete the subtree
    copy.deleteSubtree();
    return copy;
}

public boolean delete2(String deleteWord, int pos)
{
    if (word != null)
     {
        //check if the current word is the word to be deleted
        //if so, delete it, otherwise increment word count
        if (word.equals(deleteWord) && (pos == count))
        {
            //set the word to be empty for now
              this.setWord("");
              return true;
        }
        else
        {

            count++;
            return false;
        }
    }
    else
    {
        ParseTree daughter;
        boolean found=false;
        //go through the daughters
        for (int i=0; i< daughters.size() && !found;i++)
        {
            daughter = this.getDaughter(i);
            found = daughter.delete2(deleteWord, pos);
        }
        return found;
    }
}

/* method which deletes a subtree
 * the subtree which has the value "" is the one to be deleted
 */
public boolean deleteSubtree()
{
    if (daughters != null)
    {
        ParseTree daughter;
        boolean found = false;
        for (int i=0; i < daughters.size()&& !found;i++)
        {
            daughter = this.getDaughter(i);
            if (daughter.getWord() != null)
            {
                if (daughter.getWord().equals(""))
                {
                    //remove this from the list of daughters
                    daughters.remove(i);
                    found = true;
                }
            }
            else
            {
```

```
                //call the method again on this daughter
                found = daughter.deleteSubtree();
            }
        }
        return found;
    }
    else
    {
        return false;
    }
}

 /* method which adds extra words to the current tree
 */
public ArrayList addWords(ArrayList words)
{
    ArrayList list = new ArrayList();
    ArrayList originals = this.getWords();
    String original;
    String newWord;
    list.add(this);
    ArrayList newList = null;
    ParseTree tree = null;
    boolean replace = false;
    for (int i=0,j=0; i < words.size(); i++)
    {
        original = (String) originals.get(j);
         newWord = (String) words.get(i);
         //if the two words aren't the same then the word
         //in the list passed in as a parameter needs to be added
         //at position i
         if (!original.equals(newWord))
         {
                   newList = new ArrayList();
                   replace = false;
                   for (int k=0; k < list.size(); k++)
                    {
                           tree = (ParseTree) list.get(k);
                           if (tree.getWords().size() == words.size())
                                {
                                   newList.add(tree.replace(original,newWord,i));
                                   replace = true;
                                }
                           else
                               {
                                  newList.addAll(tree.add(newWord,i));
                               }
                    }
                   if (replace)
                       {
                          j++;
                       }
                   list = newList;

         }
         else
         {
             j++;
         }
      }
        return list;
}
```

```
/* method which creates a clone of the current parse tree
  */
 public ParseTree copy()
 {
     String newWord = null;
     if (word != null)
         {
             newWord = new String(word);
         }
     String newCat = new String(category);
     ArrayList newD = null;
     if (daughters != null)
        {
              newD = new ArrayList();
              ParseTree daughter = null;
              for (int i=0; i < daughters.size(); i++)
               {
                 daughter = (ParseTree) daughters.get(i);
                 newD.add(daughter.copy());
              }
        }
     return new ParseTree(newWord, newCat, newD);
 }


/ * Adds word to current tree - word will be in
  * position pos in overall sentence
  * There may be more than one place for the word
  * to be added so this method returns a list of
  * trees which are the same as the current tree
  * except that they have the added word at various
  * appropriate positions
  */
 public ArrayList add(String word, int pos)
 {
     ArrayList list = new ArrayList();
     ArrayList subTrees = getNewSubTrees(word,pos);
     ParseTree copy = null;
     int overallTreePosition = 0;
     ParseTree subTree = null;
     ArrayList subTreeInfo = null;
     for (int i=0; i < subTrees.size(); i++)
         {
             copy = this.copy();
             subTreeInfo = (ArrayList) subTrees.get(i);
             overallTreePosition = ((Integer) subTreeInfo.get(0)).intValue();
             subTree = (ParseTree) subTreeInfo.get(1);
             treeCount=1;
             copy.addSubTree(overallTreePosition,subTree);
             list.add(copy);
         }
      return list;
 }

 /* Find the new subtrees which result when the
  * when the word is added at various points in the
  * tree.
  */
 public ArrayList getNewSubTrees(String word, int pos)
 {
```

```
treeCount=0;
        count=0;
        return getNewSubTrees2(word,pos);
    }

    public ArrayList getNewSubTrees2(String word, int pos)
    {
        ArrayList list = new ArrayList();
        treeCount++;
        if (this.getWord() != null)
            {
                count++;
                return list;
            }
        else
            {
                ParseTree daughter;
                int daughterWordCount=0;
                ArrayList list2 = null;
                ParseTree newSubTree = null;
                int overallWordCount =0;
                boolean found = false;
                //look at the daughters of this subtree to see if it is suitable
                 //if it is add the word to a copy of this subtree,
                //keep a note of this new subtree and its position in the
                //original tree
                for (int i=0; i < daughters.size() && count <= pos && !found; i++)
                  {
                        daughter = (ParseTree) daughters.get(i);
                        daughterWordCount =  daughter.getWords().size();
                        overallWordCount += daughterWordCount;
                        if (overallWordCount <= pos && overallWordCount + count == pos)
                           {
                                list2 = new ArrayList();
                                list2.add(new Integer(treeCount));
                                list2.add(getNewSubTree(word,i+1));
                                list.add(list2);
                                found = true;
                            }
                        else if (count == pos)
                           {
                                list2 = new ArrayList();
                                list2.add(new Integer(treeCount));
                                list2.add(getNewSubTree(word,i));
                                list.add(list2);
                                found = true;
                            }
                    }
                //try another subtree to see if it is suitable
                for (int i=0; i < daughters.size() && count <= pos; i++)
                    {
                        daughter = (ParseTree) daughters.get(i);
                        list.addAll(daughter.getNewSubTrees2(word,pos));
                    }
                return list;
        }
    }

    /* Actually creates a new subtree for the word
     * to be added and adds it to a particular tree
     * this tree is returned
     */
```

```
    public ParseTree getNewSubTree(String word, int pos)
    {
        ParseTree newTree = new ParseTree("(CAT " + word + ")");
        ParseTree copy = this.copy();
        copy.getDaughters().add(pos,newTree);
        return copy;
    }

    /* Method which adds a subTree to the current tree at
     * a particular position in the current tree
     */
    public void addSubTree(int pos, ParseTree subTree)
    {
        if (this.daughters != null)
        {
              ParseTree daughter = null;
              for (int i=0; i < daughters.size() && treeCount <= pos; i++)
              {
                   daughter = (ParseTree) daughters.get(i);
                   treeCount++;
                   if (treeCount == pos)
                       {
                           daughters.set(i,subTree);
                       }
                   else
                       {
                           daughter.addSubTree(pos,subTree);
                       }
               }
        }
    }
}
```

Appendix E

LKB Code Changes

The purpose of this appendix is to highlight the main code changes that were made to the LKB *dag.lsp* file in order to allow agreement error constraints to be relaxed. The function `unify-dags` takes two dags `dag1` and `dag2`, unifies them and returns `dag1` with relevant pointers set beteen its nodes and the nodes in `dag2` (see Section 7.6.2). It calls the function `unify1` which calls `unify2` where the type unification takes place. This code listing shows the changes made to these functions, plus any extra functions which were written. The dag data structure was expanded to include an extra *robust agreement identity* pointer. At the moment, the agreement paths in the grammar are hardcoded into the `agree-path` and `find-path-pairs` functions, obviously an unsatisfactory solution which will have to be changed. This listing shows the agreement paths for a tiny toy grammar which comes with the LKB. The lines of code which I added are flagged by comments in uppercase letters.

```
(defvar *agree-error-path* nil) ;PATH IN DAG1 WHERE AGREEMENT CLASH OCCURS
(defvar *agree-error-dag* nil)  ;THE DAG WHICH IS RETURNED IF AGREEMENT ERROR OCCURS
(defvar *agree-error-type* nil) ;THE TYPE IN DAG2 AT ROOT OF CLASH
(defvar *during-sentence-parse* nil) ;THIS IS USED TO ENSURE THAT AGREEMENT
                                     ;CONSTRAINTS ARE ONLY RELAXED DURING
                                     ;PARSING, NOT DURING GRAMMAR COMPILIATION

(defstruct (dag
             (:constructor make-dag-x (type arcs))
             (:copier copy-dag-x))
   (type nil)
   (arcs nil)
   ;BEGIN MODIFIED SECTION
   (inconsistent nil)
   ;END MODIFIED SECTION
   ;; generation counter for all temporary slots
   (x-generation 0 :type fixnum)
   ;; new type computed during a unification
   (x-new-type nil)
   ;; new arcs computed during a unification
   (x-comp-arcs nil)
   ;; pointer to a copy of this fs
   (x-copy nil)
   ;; pointer to representative for this fs's equivalence class
   (x-forward nil)
   ;; flag used when traversing a fs doing interleaved unifications
   (x-visit-slot nil))

(defun unify-dags (dag1 dag2)
  (if *within-unification-context-p*
      (#+:uprofile
       prof:with-sampling #+:uprofile nil
       #-:uprofile
       progn
       #+(and mcl powerpc)(decf bb (CCL::%HEAP-BYTES-ALLOCATED))
       (prog1
           (catch '*fail*
             (progn
               (unify1 dag1 dag2 nil)
               (when (or *unify-debug* *unify-debug-cycles*)
                 (if (cyclic-dag-p dag1)
                   ;;
                   ;; for the (eq *unify-debug* :return) variant the
                   ;; %failure% value is determined in cyclic-dag-p()
                   ;; already; hence suppress printed output here.  the
                   ;; baroque conditionals preserve the original LKB
                   ;; behaviour.                   (19-mar-99  -  oe)
                   ;;
                   (unless (and (null *unify-debug-cycles*)
                                (eq *unify-debug* :return))
                     (format t "~%Unification failed - cyclic result"))
```

```
                  (when (and *unify-debug*
                             (not (eq *unify-debug* :return)))
                    (format t "Unification succeeded"))))
              ;dag1))
   ;BEGIN MODIFIED VERSION
   ;IF AN AGREEMENT ERROR HAS OCCURRED RETURN VALUE OF AGREE-ERROR-DAG
   (if (eq *AGREE-ERROR-DAG* nil)
      dag1 *AGREE-ERROR-DAG*
   )))
   ;END MODIFIED VERSION
         #+(and mcl powerpc)(incf bb (CCL::%HEAP-BYTES-ALLOCATED))))
    (with-unification-context (dag1)
      (when (unify-dags dag1 dag2) (copy-dag dag1)))))

(defun unify1 (dag1 dag2 path)
  (setq dag1 (deref-dag dag1))
  (setq dag2 (deref-dag dag2))
  (cond
   ((eq (dag-copy dag1) :inside)
    (when (or *unify-debug* *unify-debug-cycles*)
      (if (eq *unify-debug* :return)
        (setf %failure% (list :cycle (reverse path)))
        (format
         t
         "~%Unification failed: unifier found cycle at < ~{~A ~^: ~}>"
         (reverse path))))
    (throw '*fail* nil))
   ((not (eq dag1 dag2)) (unify2 dag1 dag2 path)))
   ;BEGINNING OF MODIFIED SECTION
   ;CHECKS WHETHER AN AGREEMENT ERROR HAS OCCURRED
   (unless (eq *agree-error-path* NIL)
       (progn
        (unless (eq (existing-complete-dag-at-end-of dag1 (reverse *agree-error-path*)) NIL)
         (progn
            ;ESTABLISH WHETHER IT IS A TRUE AGREEMENT ERROR
            (setf second-agree-paths (find-path-pairs *agree-error-path*))
             (setf is-agree-error nil)
             (dolist (second-agree-path second-agree-paths is-agree-error)
                (cond
                  ((eq
                    (existing-complete-dag-at-end-of dag1 (reverse second-agree-path))
                    (existing-complete-dag-at-end-of dag1 (reverse *agree-error-path*)))
                        (setf is-agree-error t))))
             ;IF IT IS NOT AN AGREEMENT ERROR RETURN FAIL
             (if (eq is-agree-error nil)
                (throw '*fail* nil))
            ;GET THE DAG THAT WOULD HAVE BEEN RETURNED AND
            ;MAKE CHANGES TO IT
            (setf *agree-error-dag* (copy-dag dag1))
            ;NEW-DAG CORRESPONDS TO F4 IN DEFINITION OF ROBUST
            ;AGREEMENT UNIFICATION
            (setf new-dag
                  (build-new-agree-dag *agree-error-dag*
                  (reverse *agree-error-path*) (list (first (reverse *agree-error-path*))) *agree-error-type*))
            ;CHANGE THE APPROPRIATE PART OF *AGREE-ERROR-DAG*
            (setf (dag-arc-value (find-arc-given-path  *agree-error-path* *agree-error-dag*)) new-dag)
            ;THIS IS DONE TO PREVENT FURTHER COPYING OF STRUCTURE
            (setf (dag-copy *agree-error-dag*) *agree-error-dag*)
        ))
       ));END OF MODIFIED SECTION
   dag1)
```

```
(defun unify2 (dag1 dag2 path)
  (multiple-value-bind (new-type constraintp)
      (greatest-common-subtype (unify-get-type dag1) (unify-get-type dag2))
    (if new-type
        (progn
          (setf (dag-new-type dag1) new-type)
          ;; theory is that the atomic type check is spurious because
          ;; an atomic type can't have a gcsubtype with a type
          ;; that has features.  Removing this check will mean that
          ;; type constraints etc which specify features on atomic types
          ;; are only found when we try and make them well-formed

        ;; unify in constraints if necessary - may have to copy them to
        ;; prevent separate uses of same constraint in same unification
        ;; becoming reentrant
        (when (and constraintp *unify-wffs*)
          (let ((constraint (if *expanding-types*
                                (possibly-new-constraint-of
                                 new-type)
                              (may-copy-constraint-of new-type))))
            ;;                (when *recording-constraints-p*
            ;;                (pushnew new-type *type-constraint-list* :test #'eq))
            (if  *unify-debug*
                (let ((res
                       (catch '*fail* (unify1 dag1 constraint path))))
                  (unless res
                    (if (eq *unify-debug* :return)
                        (setf %failure%
                          (list :constraints
                                (reverse path) new-type nil nil))
                      (progn (when *expanding-types*
                               (format
                                t "Problem in ~A" *expanding-types*))
                             (format
                              t
                              "~%Unification with constraint
                    of type ~A failed ~
                    at path < ~{~A ~^: ~}>"
                              new-type (reverse path))))
                    (throw '*fail* nil)))
              (unify1 dag1 constraint path)))
          ;; dag1 might just have been forwarded so dereference it again
          (setq dag1 (deref-dag dag1)))
        ;; cases for each of dag1 and dag2 where they have no arcs just
        ;; considering straightforward use of unify1: if we've
        ;; previously visited a node with no arcs then it must have got
        ;; forwarded then so we won't ever visit it again - so no need
        ;; to test for presence of any comp-arcs BUT:
        ;; unify-paths-dag-at-end-of1 adds to comp-arcs independently so
        ;; we do need the additional tests
        (cond
         ((and (null (dag-arcs dag1)) (null (dag-comp-arcs dag1)))
          (setf (dag-new-type dag2) new-type)
          ;BEGINNING OF MODIFIED SECTION
          ;PROPOGATING INCONSISTENT IDENTITY RELATION
          (unless (eq (dag-inconsistent dag1) nil)
           (setf (dag-inconsistent dag2) (dag-inconsistent dag1)))
          ;END OF MODIFIED SECTION
          (setf (dag-forward dag1) dag2))
         ((and (null (dag-arcs dag2)) (null (dag-comp-arcs dag2)))
```

```
           ;BEGINNING OF MODIFIED SECTION
           ;PROPOGATING INCONSISTENT IDENTITY RELATION
          (unless (eq (dag-inconsistent dag2) nil)
           (setf (dag-inconsistent dag1) (dag-inconsistent dag2)))
            ;END OF MODIFIED SECTION
          (setf (dag-forward dag2) dag1))
         (t
          (setf (dag-forward dag2) dag1)
          (setf (dag-copy dag1) :inside)
          (unify-arcs dag1 dag2 path)
          (setf (dag-copy dag1) nil))))
      ;BEGINNING OF MODIFIED SECTION
      ; TYPE CLASH AT THE END OF AN AGREEMENT PATH
      (if (and (equal *during-sentence-parse* t)
             (agree-path path)
             )
         (progn
           ;RECORD AGREEMENT ERROR INFO
           ;IN GLOBAL VARIABLES AND PROCEED
           ;AS NORMAL
           (setf *agree-error-path* path)
           (setf *agree-error-type* (dag-type dag2))
           (setf (dag-new-type dag1) (dag-type dag1))
           (setf (dag-forward dag2) dag1)
         )
      ;END OF MODIFIED SECTION
      (progn
        (when *unify-debug*
          (if (eq *unify-debug* :return)
           (setf %failure%
            (list :clash (reverse path)
                  (unify-get-type dag1) (unify-get-type dag2)))
        (progn (with-open-file (stream "debug" :direction :output)
        (format stream "~%Unification of ~S and ~S failed at path < ~{~S ~^: ~}>"
         (unify-get-type dag1) (unify-get-type dag2)
         (reverse path))))))
    (throw '*fail* nil))))))

;ALL THE FOLLOWING FUNCTIONS WERE ADDED

;BUILDS A NEW DAG WITH AGREEMENT INFO FROM DAG2 - F4 IN DEFINITION OF
;ROBUST AGREEMENT UNIFICATION
(defun build-new-agree-dag (dag path subpath newtype)
  (setf agree-dag (find-agree-dag dag (rest path) subpath))
 (setf new-atomic-dag (make-dag :type newtype :arcs nil))
 (if (equal path *agree-error-path*)
      ;THIS IS THE SIMPLEST CASE WHERE AGREEMENT TYPE CLASH OCCURS AT
      ;SAME POINT AS AGREEMENT STRUCTURE-SHARING CONSTRAINT
      ;JUST RETURN AN ATOMIC TYPE IN THIS CASE
      (progn
       (setf new-dag new-atomic-dag)
       (setf (dag-inconsistent new-dag) agree-dag)
       (setf (dag-inconsistent agree-dag) new-dag)
      )
      ;THIS IS MORE COMPLICATED CASE WHERE AGREEMENT TYPE CLASH OCCURS AT
      ;POINT LOWER DOWN IN DAG THAN AGREEMENT STRUCTURE-SHARING CONSTRAINT
      (progn
           (setf new-dag (copy-dag-completely agree-dag))
           (setf new-path  (reverse (set-difference path *agree-error-path*)))
           (setf (dag-arc-value (find-arc-given-path new-path new-dag)) new-atomic-dag)
           ;ESTABLISHING INCONSISTENT IDENTITY RELATION
           (setf
```

```
                (dag-inconsistent (dag-arc-value (find-arc-given-path new-path new-dag)))
                 (dag-arc-value (find-arc-given-path new-path agree-dag)))
            (setf
                  (dag-inconsistent (dag-arc-value (find-arc-given-path new-path agree-dag)))
                  (dag-arc-value (find-arc-given-path new-path new-dag)))
     ))
  new-dag)

;USED IN BUILD-NEW-AGREE-DAG
;CHECKS IF THE AGREEMENT CONSTRAINT IS HIGHER UP DAG1 THAN WHERE CLASH OCCURRED
;IF SO RESETS AGREE-ERROR-PATH VARIABLE
;RETURNS DAG AT END OF AGREE-ERROR-PATH
(defun find-agree-dag (dag path subpath)
   (cond ((agree-path (reverse subpath))
           (progn
                  (setf *agree-error-path* subpath)
                  (existing-complete-dag-at-end-of dag subpath)))
         ((eq path nil) nil)
         (t (find-agree-dag dag (rest path) (append subpath (list (first path)))))
))

;FIND THE DAG AT THE END OF PATH LABELS-CHAIN IN REAL-DAG
(defun existing-complete-dag-at-end-of (real-dag labels-chain)
   (cond
      ((null labels-chain) real-dag)
      (t
         (let ((one-step-down
                  (get-complete-dag-value real-dag
                     (car labels-chain))))
            (if one-step-down
               (existing-complete-dag-at-end-of one-step-down
                  (cdr labels-chain))
               nil)))))

;USED IN EXISTING-COMPLETE-DAG-AT-END-OF
(defun get-complete-dag-value (dag attribute)
   (dolist (arc (append (dag-arcs dag) (dag-comp-arcs dag)) nil)
      (when (eq attribute (dag-arc-attribute arc))
         (return-from get-complete-dag-value (dag-arc-value arc)))))

;THIS FUNCTION FINDS THE ARC CORRESPONDING TO PATH IN DAG
(defun find-arc-given-path (path dag)
  (setf dag1 dag)
  (setf arc1 NIL)
  (dolist (attribute path arc1)
    (setf arc1 (unify-arcs-find-arc attribute (dag-arcs dag1)
                                               (dag-comp-arcs dag1)))
    (setf dag1 (dag-arc-value arc1))
  )
)

;HARD-CODED AGREEMENT PATHS IN GRAMMAR BEING USED WITH LKB
;THIS EXAMPLE IS A TINY GRAMMAR WHICH COMES WITH LKB
(defun agree-path (path)
  (cond
        ((equal path '(AGR FIRST ARGS)) t)
        ((subpath (reverse '(AGR FIRST ARGS)) (reverse path)) t)
        ((equal path '(AGR FIRST REST ARGS)) t)
        ((subpath (reverse '(AGR FIRST REST ARGS)) (reverse path)) t)
          (t nil)))

; GIVEN ONE POTENTIAL AGREEMENT PATH, FINDS ITS MATCHING RE-ENTRANT PATH
```

```
; HARD-CODED - NEEDS TO BE CHANGED SO THIS INFORMATION CAN BE COMPILED
; FROM THE GRAMMAR
(defun find-path-pairs (path)
  (cond
;tiniest/grammar3
((subpath (reverse '(AGR FIRST ARGS)) (reverse path))
  (list (append (set-difference path '(AGR FIRST ARGS)) '(AGR FIRST REST ARGS))))
;tiniest/grammar3
        ((subpath (reverse '(AGR FIRST REST ARGS)) (reverse path))
  (list (append (set-difference path '(AGR FIRST REST ARGS)) '(AGR FIRST ARGS))))
        (t nil)))


;DETERMINES IF PATH1 IS A SUBPATH OF PATH2
;USED IN FIND-PATH-PAIRS AND AGREE-PATH
(defun subpath (path1 path2)
    (cond ((eq path1 nil) t)
          ((equal (first path1) (first path2)) (subpath (rest path1) (rest path2)))
          (t nil)))
```