Identifying and resolving *one*-anaphora

Mary Gardiner

November 2003

Thesis submitted in partial fulfilment of the requirements for the degree of Bachelor of Science (Honours).

Department of Computing, Division of ICS, Macquarie University

Abstract

This thesis describes the results of a largely data-driven investigation of *one*-anaphora. *One*-anaphora is a little investigated variety of anaphora where the anaphor and its antecedent partly share sense as well as or instead of reference. An example of a *one*-anaphor is *the large one* in the sentence "My drink is the large one," and the resolution problem is that of determining that *the large one* means a large drink.

The problem of identifying and resolving *one*-anaphora has three parts: distinguishing *one*-anaphors from other uses of *one* in English, locating the antecedent of each *one*-anaphor, and determining which parts of the sense of the antecedent are shared by the *one*-anaphor. This thesis presents results illuminating part of each of these problems.

Many of the results of this project were derived from analysis of *one*-anaphors in a corpus of English text. The major computational outcomes of this project are: a program that identifies *one*-anaphors in text, a program that identifies the antecedents of *one*-anaphors in text, and a program that uses web results to resolve *one*-anaphors. The major results of the corpus analysis and of the programs are: analysis of the distribution of *one*-anaphors in text; analysis of the grammatical function of *one*-anaphors as compared to other uses of the word *one*; analysis of the grammatical function of *one*-anaphors and their antecedents; and analysis of shallow semantic techniques for resolving *one*-anaphors.

Contents

1 Introduction **5**
- 1.1 Anaphora resolution and language progressing . . . 6
- 1.2 *One*-anaphora . . . 7
- 1.3 Scope of this project . . . 9
  - 1.3.1 Identifying *one*-anaphors . . . 9
  - 1.3.2 Finding antecedents of *one*-anaphors . . . 9
  - 1.3.3 Finding the sense of *one*-anaphors . . . 10
- 1.4 The British National Corpus . . . 10
  - 1.4.1 Description of the British National Corpus . . . 10
  - 1.4.2 *One* in the British National Corpus . . . 11
- 1.5 Thesis Outline . . . 11

2 Literature Review **13**
- 2.1 Non-anaphoric uses of *one* . . . 13
  - 2.1.1 Numeral *one* . . . 14
  - 2.1.2 Generic *one* . . . 15
  - 2.1.3 Object *one* . . . 15
  - 2.1.4 "Pro-noun" *one* . . . 15
  - 2.1.5 Criteria for distinguishing uses of *one* . . . 16
- 2.2 *One*-anaphora . . . 16
  - 2.2.1 Discourse models and anaphora resolution . . . 17
  - 2.2.2 The definition of and discourse function of *one*-anaphora . . . 17
- 2.3 Resolution of *one*-anaphora . . . 22
  - 2.3.1 Techniques for finding antecedent noun phrases . . . 22
  - 2.3.2 Techniques for finding the meaning of a *one*-anaphor . . . 25
- 2.4 Conclusions . . . 25

3 Identifying *one*-anaphora **27**
- 3.1 Uses of *one* in English . . . 27
  - 3.1.1 Example uses of *one* in English . . . 28
  - 3.1.2 Describing uses of *one* in English . . . 28
- 3.2 Approaches to distinguishing uses of *one* in English . . . 29
  - 3.2.1 Distinguishing uses of *one* by part of speech . . . 30

3.2.2 Distinguishing uses of *one* by grammatical role . . . 31
3.2.3 Distinguishing uses of *one* using semantic information . . . 33
3.3 System overview . . . 33
3.4 Results . . . 34
3.4.1 Test data sets . . . 34
3.4.2 Analysis of the distribution of grammatical roles among the uses of *one* . . . 35
3.4.3 Distinguishing uses of *one* in English . . . 37
3.5 Error analysis . . . 37
3.6 Conclusions . . . 38

4 Resolving *one*-anaphora 39
4.1 Test data set . . . 40
4.2 Finding the antecedent of a *one*-anaphor . . . 40
4.2.1 The nearness of *one*-anaphors and their antecedents . . . 41
4.2.2 The syntactic role of *one*-anaphors and their antecedents . . . 41
4.2.3 Finding antecedents using syntactic information . . . 44
4.2.4 Finding antecedents using semantic information . . . 45
4.3 Finding the sense of a *one*-anaphor . . . 47
4.3.1 Contrast between *one*-anaphor and antecedent . . . 49
4.3.2 Discovering contrast using Google . . . 52
4.4 Conclusions . . . 53

5 Conclusions 56
5.1 Identifying *one*-anaphora . . . 56
5.2 Resolving *one*-anaphora . . . 57
5.3 Future work . . . 57
5.3.1 Corpus analysis of *one*-anaphora . . . 57
5.3.2 Finding antecedents of *one*-anaphora . . . 58
5.3.3 Resolving *one*-anaphora . . . 59
5.3.4 Summary of future work . . . 60

A Uses of the word *one*: Taxonomy 63
A.1 Non-anaphoric uses of *one* . . . 63
A.1.1 Numeric *one* . . . 64
A.1.2 Generic *one* . . . 64
A.1.3 Object *one* . . . 65
A.1.4 Partitive *one* . . . 66
A.1.5 Idiomatic uses of *one* . . . 67
A.2 Anaphoric uses of *one* . . . 68
A.2.1 Antecedent is a type or class . . . 68
A.2.2 Antecedent is a set . . . 69
A.2.3 Antecedent is an instance . . . 70
A.3 Conclusions . . . 71

B Uses of *one* in English: Empirical evidence **72**
B.1 Evidence From Part of Speech Tags 72
B.2 Evidence from Part of Speech Bigrams 75
B.3 Evidence From Lexical Bigrams and Trigrams 76
B.4 Conclusion 78

C Code documentation **80**
C.1 Corpus analysis tools 80
C.2 Corpus annotation tools 81
C.3 Anaphora resolution system 81
C.3.1 Hand annotation results 82

Chapter 1

Introduction

This document describes the results of my honours project, an investigation of computational approaches to a particular problem that is part of the semantic interpretation of text: the problem of resolving *one*-anaphora.

One-anaphors are anaphoric noun phrases headed by the word *one*—that is uses of *one* like that in example 1.1 in which the noun phrase *the brown one* cannot be correctly interpreted as referring to a brown kitten without deducing its relationship to the phrase *the black kitten*.

(1.1) I'd rather the black kitten than *the brown one*.

I have designed and implemented several systems that solve several parts of the problem of resolving *one*-anaphora. In the case of example 1.1, the system would determine that *the brown one* is in fact a *one*-anaphor, locate the noun phrase *the black kitten* from which the *one*-anaphor derives its meaning, and analyse how *the brown one* relates to *the black kitten*.

In order to resolve *one*-anaphors, I developed algorithms that first identify *one*-anaphors and then determine their referents. I have implemented a system that resolves *one*-anaphors, using the algorithms that I develop.

My project builds on existing theories of the various uses of *one*, *one*-anaphora in particular, and will focus on developing and using algorithms that are computationally achievable.

This chapter introduces anaphora resolution in general in Section 1.1, the particular problem of *one*-anaphora resolution in Section 1.2, the scope of my project in Section 1.3, the source of data I used in Section 1.4 and concludes by outlining the remainder of this thesis in Section 1.5.

1.1 Anaphora resolution and language progressing

The problem of anaphora resolution is far larger than the particular problem discussed in this thesis, and in turn it is only part of the problem of natural language understanding. The general problem is that of determining what a text is saying. Example 1.1 can be paraphrased in a number of ways, an example of such a paraphrase is:

(1.2) There are two kittens. One of the kittens is black. One of them is brown. If I was to choose between them, I would choose the black kitten, not the brown kitten.

Example 1.1 directly asserted that the speaker preferred the black kitten to the brown one, but indirectly asserted a number of other propositions. I have made some of these propositions clearer in the paraphrase, but in particular, example 1.1 implied the existence of three entities: the speaker; a black kitten; and a brown kitten, and described some relationships between those entities.

In computational terms, a model of the world described by example 1.1 would contain three entities, each of which would have certain properties inferred from the sentence—one entity is the speaker of a sentence, another is a black kitten, and the third a brown kitten, and would also contain the relationship between the three entities.

A model of example 1.3, on the other hand, would contain only one entity—representing the first edition—but with properties supplied by two noun phrases, *[t]he first edition* and *the one without illustrations*:

(1.3) The first edition was *the one without illustrations*.

The problem of natural language understanding can be modelled as a function from an utterance in natural language to model of the meaning of the text such as those described above.

There are many problems with developing such a function. For example, a model of 1.1 contains three entities, but there is no intrinsic reason why this is so. There are three *references* to entities made in the sentence, but there is no reason that each reference must be to a distinct entity. For example

(1.4) *I* hate *myself*.

contains two references, but both references are to the same entity, the speaker.

Our understanding that *I*, *the black kitten* and *the brown one* refer to distinct things comes from our understanding of language and the world. In example 1.1, we understand that the semantics of the word *rather* (in the sense of preferring) normally requires that the preferred thing and the less preferred thing aren't one and the same. A second way that this might be inferred is by using our knowledge that *black* and *brown* contrast.

On the other hand, *[t]he first edition* and *the one without illustrations* in example 1.3 are describing the same entity, and part of the way we determine this is by using our knowledge of books to determine that a first edition may well be without illustrations.

The problem of determining when two references co-refer—refer to the same thing—is part of the problem of anaphora resolution. Anaphora are pieces of a text that cannot be fully interpreted without using other pieces of the text, and sometimes also knowledge gained from the context. For example, a model of example 1.5 would contain an entity which was a dentist *and* played golf on weekends. The fact that *she* refers to an entity already mentioned is determined by looking at foregoing text, and taking into account various correspondences between *Jane* and *she*.

(1.5) Jane's a dentist, and *she* plays golf on weekends.

However, co-reference is not the only problem in anaphora resolution, as example 1.1 illustrated. In particular, while some *one*-anaphors do co-refer with their antecedent, this is not always the case. A *one*-anaphora resolution system is primarily concerned with modelling the sense of the *one*-anaphor, rather than with linking it to co-referents earlier in the text.

Anaphora resolution is crucial to any natural language system that relies on a detailed model of the meaning of a text. Examples of natural language systems that use anaphora resolution are machine translation systems and question answering systems.

1.2 *One*-anaphora

This section discusses the problem of resolving *one*-anaphora as addressed in this thesis using some examples of *one*-anaphors. It concludes by identifying the three main subproblems of *one*-anaphora resolution.

One-anaphors are anaphoric noun phrases headed by *one*. Determining the sense of a *one*-anaphor requires using at least one other noun phrase in the text. Examples 1 to 1.10 are all examples of *one*-anaphors.

(1.6) There's **an Imperial Airways Heracles**, *the one that flies to India*.[1] (BNC file ALL sentence 2028)[2]

(1.7) The examination, held at the Freemasons' tavern, was entirely viva voce and the complaint was made that the candidates did not even have to cast **a horse**, or operate on *a dead one*, still less on *a live one which might have benefited*. (BNC file B2W sentence 1072)

[1] Throughout this thesis, I use the following typographical conventions for *one*-anaphora and their antecedents: an antecedent noun phrase is highlighted **like this** and a *one*-anaphoric noun phrase *like this*.

[2] The citation indicates that this example can be found in the British National Corpus, in the file named ALL, sentence number 2028.

(1.8) The escorting Macchi pilots attacked **the Hurricanes**, claiming **four** shot down, *one* by all the pilots jointly, and *one* each by Magg. (BNC file CA8 sentence 656)

(1.9) How will your book give **a different image of Modigliani** from *the usual one in already published works*? (BNC file CKT sentence 887)

(1.10) And since for the successful confection and storage of many, although not all, potted meats and fish, clarified butter is a necessary adjunct, it seems only fair to warn readers that the process does involve **a little bother**, although *a trifling one* compared to the services rendered by a supply of this highly satisfactory sealing, mixing, and incidentally, frying ingredient. (BNC file EFU sentence 998)

The problem of resolving *one*-anaphora is the problem of mapping from their surface form (*one* in the above examples) to a model of their referent. For example, in the case of example 1, *one* refers to the Imperial Airways Hercules already mentioned, but in example 1.9, *one* refers to the usual image of Modigliani, rather than the different image of Modigliani referred to by its antecedent.

Consider the problem represented by example 1.9 in more detail. Given the sentence, the first problem is to determine whether or not the use *one* in the sentence is in fact anaphoric. Contrast example 1.9 with example 1.11 below, in which the properties of the noun phrase *one pint* can be determined from the phrase itself:

(1.11) If a normal 7-inch or bigger dish is used, *one pint* of fondue is the right quantity. (BNC file EFU, sentence 59)

Having determined that the phrase *the usual one* in example 1.9 is anaphoric, the next problem is to determine that *a different image of Modigliani* is the antecedent of *one*—that is, that the meaning of *one* can be inferred from the phrase *a different image of Modigliani*.

The final problem is *how* the referent of *the usual one* can be determined using the antecedent *a different image of Modigliani*. Assume that the system understands *a different image of Modigliani* to be an image with two properties:

1. being of Modigliani; and
2. being different from the referent of *the usual one*.

The system must then be able to infer that *the usual one* does not refer to a "different usual" image of Modigliani, while ideally also recognising that *the usual one* does not refer to a usual image of something that is *not* Modigliani. A perfect solution to the problem should be able to select the maximal set of properties that apply to the referent of the *one*-anaphor while excluding properties of the antecedent that do not apply to the referent of the *one*-anaphor.

So the problem of mapping the surface form of a *one*-anaphor to a representation of its meaning devolves into three problems:

1. that of identifying *one*-anaphors;
2. that of finding the antecedents of *one*-anaphors; and
3. that of using the antecedent to find the properties of the referent of the *one*-anaphor.

1.3 Scope of this project

The project described in this thesis explored aspects of all three problems listed at the end of Section 1.2: identifying *one*-anaphors, finding the antecedents of *one*-anaphors, and using the antecedent to find the properties of *one*-anaphors. This section describes the project's exploration of each of these subproblems in more detail.

1.3.1 Identifying *one*-anaphors

As part of this project I unified existing descriptions of *one*-anaphora, explored the problem of finding sources of *one*-anaphors in the British National Corpus, explored the problem of distinguishing *one*-anaphors from other uses of *one*, and developed a system which computationally distinguished uses of *one*.

Various authors have described anaphoric and non-anaphoric uses of *one*. Although their categories substantially overlap there is no single set of names for each of the uses of *one* used in the literature. Section 2.1 of this thesis describes the different taxonomies of non-anaphoric *one* usage and discusses where they overlap and where they differ. Section 2.2 describes the different taxonomies of *one*-anaphors and discusses where they overlap and where they differ. Appendix A describes my own taxonomy of uses of *one*.

Given that many uses of *one* are not anaphoric, one of the earliest problems for the project was that of finding a source of *one*-anaphors to analyse. Appendix B is the result of this investigation.

There are several conceivable ways to distinguish different uses of *one*. This project explored two of them: the grammatical role of uses of *one* and some shallow semantic information about uses of *one*. The result of this exploration was a system that distinguishes uses of *one*. Chapter 3 gives the results of grammatical analysis of uses of *one* and describes the program that distinguishes uses of *one*.

1.3.2 Finding antecedents of *one*-anaphors

As part of this project I described existing methods of finding the antecedents of *one*-anaphors, investigated several features of the antecedents of *one*-anaphors, and developed several programs that use different methods to identify the antecedents.

While the literature on *one*-anaphors is very small, Section 2.3.1 discusses the extent of existing work on finding the antecedents of *one*-anaphors and describes existing systems that have been developed to find the antecedents of *one*-anaphors.

I annotated a set of 495 *one*-anaphors and their antecedents, and used this annotation to explore two sets of relationships between antecedents and *one*-anaphors: their nearness, and their grammatical function. The annotated data is described in Section 4.1 and the results of this annotation in Sections 4.2.1 and 4.2.2.

I developed several systems that locate antecedents of *one*-anaphors. Most of these systems use grammatical information, but two use web searches as a source of semantic information about likely antecedents. These systems and their results are discussed in Sections 4.2.3 and 4.2.4.

1.3.3 Finding the sense of *one*-anaphors

As part of this project I annotated a set of *one*-anaphors and their antecedents in order to quantitatively evaluate the different relationships between antecedents and *one*-anaphors. I also developed a small system which uses web searches to evaluate different properties of the antecedent and the likelihood of those properties also being a property of the *one*-anaphor. The analysis of the relationship between *one*-anaphors and their antecedents is described in Section 4.3.1 and the use of web searches to evaluate properties of the antecedent is discussed in Section 4.3.2.

1.4 The British National Corpus

This section describes the format and contents of the British National Corpus, a corpus of English texts from a variety of sources which I used as the source of sentences containing both *one*-anaphors and non-anaphoric uses of *one* for this project. Section 1.4.1 describes the format of the corpus and sources of the text contained within it, and Section 1.4.2 briefly describes the places in which *one* appears in the corpus.

1.4.1 Description of the British National Corpus

The British National Corpus contains over one hundred million words of British English, drawn from written and spoken sources. It is drawn from books, periodicals, advertising material, letters, natural conversations, and spoken monologues. (British National Corpus 1995)

The BNC is annotated in SGML. The annotations represent the part of speech tags assigned by the CLAWS part-of-speech tagger, and also indicate structural features such as lists and paragraphs, and information such as the last update of the data. (British National Corpus 1995). Here is a sample of the text in BNC file A00:

```
<s n="1"><w NP0>HIV <w NN1>IT<w VBZ>'S <w DPS>YOUR <w NN1>CHOICE
</head>
<pb n=1> <gap desc="Newspaper cutting omitted" resp=OUP>
<pb n=2>
<p>
<s n="2"><w AT0>Every <w NN1>day <w AT0>the <w NN1>virus
<w VVG>causing <w NN1>AIDS <w VBZ>is <w VVG>infecting
<w AV0>more <w AJ0>young <w NN0>people<c PUN>.
<s n="3"><w AT0>A <w NN1>friend <w VM0>can <w VVI>infect
<w PNP>you <w PRP>without <w DPS>your <w NN1>knowing<c PUN>.
<pb n=3>
</p>
```

Each sentence is tagged using the `<s>` tag, and assigned a number. Each word is tagged with a `<w>` tag, and the tag's attribute represents the part of speech assigned by the CLAWS part of speech tagger. The CLAWS tagger is based upon a Hidden Markov statistical model, but is designed to be robust over the wide variety of texts that the BNC contains. (Leech, Garside & Bryant 1994)

Leech et al. (1994) determined that the system tagged 1.5% of the corpus incorrectly, and could not assign a tag to a further 3.3% of the corpus. Error and ambiguity rates rise to 1.78% and 4.60% respectively as the system tries to assign subcategories (such as common noun) to words.

1.4.2 *One* in the British National Corpus

This subsection briefly discusses the distribution of the word *one* in the BNC. For a more complete account, see Appendix B.

The word *one* accounts for 0.26% of all the words in the BNC. It is assigned twelve parts of speech by the CLAWS tagger, but since five of these twelve account for over 99.99% of all occurences of *one*, the other seven are presumably errors.

By far the most common part of speech tag assigned to *one* is the **cardinal numeral** tag, which is assigned to 70.20% of all uses of *one*. The next most common tag is the **indefinite pronoun** tag, assigned to 21.19% of all occurences of *one*.

Appendix B examines part of speech bigrams—that is, the part of speech of the word *one*, together with that of the word preceding and following it. If uses of *one* are divided based upon the part of speech of the word preceding them, as well as their own assigned part of speech, the sets of uses become significantly smaller.

1.5 Thesis Outline

This thesis is structured as follows:

Chapter 2 reviews existing work on *one*-anaphora, and uses of *one* in English in general;

Chapter 3 discusses the problem of distinguishing uses of *one* in English, and describes the approach that this project took;

Chapter 4 discusses the problems of finding antecedents of *one*-anaphora, and of using those antecedents to model the referents of *one*-anaphors;

Chapter 5 draws conclusions from the work done in this project and describes possible future avenues of research;

Appendix A describes the uses of *one* in English, including the differing types of *one*-anaphora;

Appendix B describes an empirical investigation into the distribution of the uses of *one* in English; and

Appendix C documents the programs developed in the course of this project.

Chapter 2

Literature Review

This chapter describes previous work on resolving *one*-anaphors and related work on resolving other noun phrase anaphora.

This chapter is divided into three sections, each discussing a separate problem related to the resolution of *one*-anaphora. Section 2.1 discusses existing descriptions of non-anaphoric uses of *one*, and describes the criteria used in the literature to distinguish different uses of *one*. Section 2.2 describes existing models of *one*-anaphora and their role in discourse. Section 2.3 discusses techniques for resolving *one*-anaphora. Section 2.4 discusses the results of this review in terms of the three problems listed in Section 1.3 that this project addresses.

2.1 Non-anaphoric uses of *one*

It is clear that not every occurrence of *one* in English is anaphoric. In fact, that there are several other uses of *one* in English, some of which are more frequently used than the anaphoric *one*, and many of which do not refer at all.

Consider these examples:

(2.1) I'll have *one* pencil.

(2.2) *One* prefers that *one*'s minions write in pencil.

In example 2.1, the word *one* is a cardinal number modifying the word *pencil* rather than serving any referential purpose at all. In example 2.2, despite *one* being at the head of a noun phrase like a *one*-anaphor and despite the fact that it may in fact refer to a person, it is not *one*-anaphoric since the reader does not need preceding text to derive its sense.

Non-anaphoric uses of *one* are described in the literature to varying extents but there is no single widely-used taxonomy and no attempts to unify the various descriptions.

This section discusses previous work on describing the various uses of *one* in English. Such work is generally not computationally oriented, so criteria described here for determining whether a use of *one* falls into a particular category is not always computationally usable.

The purpose of this review is twofold. Firstly, it informs my taxonomy in two ways: the criteria and collection of examples used by each author informs my taxonomy; and the examples and criteria allow me to compare and contrast my taxonomy with those developed by others. Secondly, it reviews existing criteria for distinguishing non-anaphoric uses of *one* from one another.

2.1.1 Numeral *one*

Several authors identify this use of *one*. They variously call it the **cardinal numeral *one*** (Halliday & Hasan 1976), the **determiner *one*** (Luperfoy 1991), and the **cardinal number *one*** (Dahl 1985).

Halliday & Hasan (1976), Dahl (1985) and Luperfoy (1991) all provide examples of the numeral *one* modifying a head noun:

(2.3) He made *one* very good point. (Halliday & Hasan 1976, p. 98)

(2.4) The *one* friend who never let her down was Enid. (Halliday & Hasan 1976, p. 99)

(2.5) You've already got *one* red one. (Halliday & Hasan 1976, p. 99)

(2.6) It took *one* man to lift the piano. (Luperfoy 1991, p. 115)

(2.7) I have *one* dog. (Dahl 1985, p. 4)

Luperfoy's (1991) characterisation of this use of *one* as a determiner would seem to exclude sentences like that in example 2.8, whereas Halliday & Hasan's (1976) and Dahl's (1985) characterisations do not.

(2.8) Bill was the *one* exception to the general trend.

Halliday & Hasan (1976) describe the numeral *one* as contrasting with *two*, *three* and so on and Luperfoy (1991) makes a similar observation when she shows that in a plural noun phrase, the plural of the numeral *one* cannot be *ones* by contrasting the unacceptable sentence in example 2.9 with the acceptable sentences in example 2.10.

(2.9) * It took *ones* man to lift the piano. (Luperfoy 1991, p. 115)

(2.10) It took $\left\{ \begin{array}{c} \textit{four} \\ \textit{several} \\ \textit{many} \end{array} \right\}$ men to lift the piano. (Luperfoy 1991, p. 115)

Halliday & Hasan (1976) are alone in arguing that uses of *one* like that in example 2.11 are numeral uses of *one*.

(2.11) Ten set out, but only *one* came back. (Halliday & Hasan 1976, p. 98)

They explain that *one* in example 2.11, while anaphoric, is intended to substitute for *one man* rather than simply for *man* and therefore functions exactly like a "cardinal numeral *one*" with an elided head. See Section 2.2.2 for a longer description of Halliday & Hasan's (1976) idea of substitution.

2.1.2 Generic *one*

The generic *one* is variously called the **generalised person *one*** (Halliday & Hasan 1976), the **generic personal pronoun *one*** (Luperfoy 1991) and the **impersonal *one*** (Dahl 1985). Each provide an example of the generic *one*:

(2.12) *One* never knows what might happen. (Halliday & Hasan 1976, p. 98)

(2.13) *One* need only enter the lobby to feel the presence of wealth. (Luperfoy 1991, p. 115)

(2.14) *One* never knows what will happen. (Dahl 1985, p. 4)

Luperfoy (1991, p. 115) observes that there is no separate generic plural pronoun, and that therefore speakers use the second person plural pronoun *you* instead:

(2.15) *You* need only gather in the lobby to be arrested for rioting. (Luperfoy 1991, p. 115)

2.1.3 Object *one*

This use of *one* is only described by Luperfoy (1991) and is intended to account for sentences like those in example 2.16, where *one one* in 2.16a refers to a single one dollar bill, and *ten ones* in 2.16b refers to ten one dollar bills.

(2.16) a. I'm getting low on small bills. I only have one *one*. (Luperfoy 1991, p. 115)
b. I can trade two fives and ten *ones* for your twenty. (Luperfoy 1991, p. 115)

2.1.4 "Pro-noun" *one*

Halliday & Hasan (1976) identify a use of *one* they call the **"pro-noun" *one***. It has uniquely human referents, but is not anaphoric—it does not rely on other parts of the text for meaning. It always means "a person." They give the following examples:

(2.17) If such a *one* be fit to govern, speak. (Halliday & Hasan 1976, p. 102)

(2.18) The *ones* she really loves are her grandparents. (Halliday & Hasan 1976, p. 102)

(2.19) The *one* he needs is his lawyer. (Halliday & Hasan 1976, p. 103)

Halliday & Hasan (1976) observe that the plural of the pro-noun *one* is *ones*, as in example 2.20.

(2.20) Now, my dearest *ones*; gather round. (Halliday & Hasan 1976, p. 103)

2.1.5 Criteria for distinguishing uses of *one*

Halliday & Hasan (1976), Dahl (1985) and Luperfoy (1991) identify two main criteria that distinguish the non-anaphoric uses of *one* from each other: their syntactic properties and their possible plurals. They identified three main syntactic functions performed by the word *one*. *One* occurs as the head of a noun phrase (called a **nominal group** in Halliday & Hasan's (1976) grammar); as the modifier of a noun phrase; and as a determiner of a noun phrase. They identify three possible sets of plurals for non-anaphoric uses of *one*:

1. the other cardinal numbers, *two*, *three* etc, and the equivalent indefinite plurals *many* and *several*;
2. *ones*; and
3. *you*.

Unfortunately, while pluralising a sentence in order to determine how *one* is being used would be a useful technique for a human annotator, it is not a computationally useful technique because the state of the art in natural language understanding systems cannot easily decide upon the acceptability of any particular sentence.

The grammatical function of a particular word can be computationally determined by a parser like that provided by Connexor, but as *one*-anaphors, the generic *one* and the object *one* all occur as the head of a noun phrase, this technique is also not sufficient for computationally distinguishing uses of *one*.

2.2 *One*-anaphora

This section describes the discourse role of *one*-anaphora. Section 2.2.1 briefly describes the common model of discourse. Section 2.2.2 describes the roles that *one*-anaphora play in discourse, describing the three main roles of *one*-anaphora identified in the literature.

2.2.1 Discourse models and anaphora resolution

This section describes the common model of anaphora resolution. In this model the aim of anaphora resolution is to derive not only the relationship between the anaphor and its antecedent, but also the relationship between the anaphor and the description of the world that the text is describing. The purpose of this section is to provide necessary background for models of *one*-anaphora given in Section 2.2.2.

Anaphoric noun phrases rely on information outside the noun phrase to determine their meaning. This information might be extra-linguistic or obvious from context, for example a speaker might point at a loaf of break and ask:

(2.21) Give me *that*!

In many cases however, especially in written text, the meaning can be derived from a noun phrase elsewhere in the text called the **antecedent**, which usually precedes the anaphoric noun phrase.

Much of the existing work on anaphora resolution is on reference resolution—finding out what entity an anaphor refers to—rather than sense resolution—finding out the properties of the entity the anaphor refers to.

A frequently used definition of anaphora is that of Halliday & Hasan (1976). **Cohesion**, in their account, is a relation of semantic dependency between elements of a discourse. In some way, the interpretation of one of the elements depends on the other. Anaphora is a type of cohesion where the dependent element (the anaphor) occurs in the text after the element it depends on. They describe the relationship between anaphor and antecedent as being one where the anaphor "points back" in the text to its antecedent.

In most accounts, however, while discovery of this cohesive relationship is part of the problem of resolving anaphora, the goal of anaphora resolution is not to develop a model of the cohesion relations in the text, but rather to develop a model of the world that the text talks about.

Consider Figures 2.1 and 2.2. Both are a model of example 2.22, but Figure 2.1 models the internal relationships between textual elements, and Figure 2.2 the relationship between the text and the world it is describing.

(2.22) The robot will always obey orders. He will tear off a limb if ordered to.

A model like that in Figure 2.2 is usually called a **discourse model**. Anaphora resolution is part of the process of building a discourse model.

2.2.2 The definition of and discourse function of *one*-anaphora

The term ***one*-anaphor** is often not restricted to anaphoric uses of the word *one*, or even anaphoric uses of *one* and its various plurals. Most descriptions do not agree with my definition of *one*-

Figure 2.1: A model of cohesive relationships in example 2.22

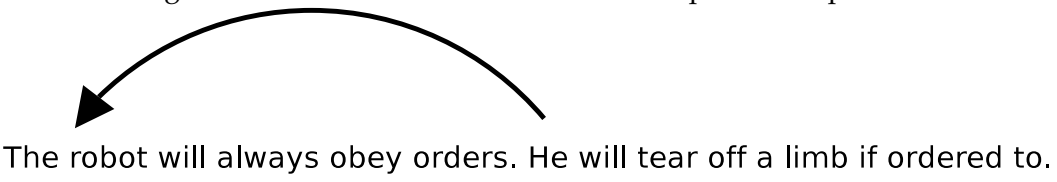
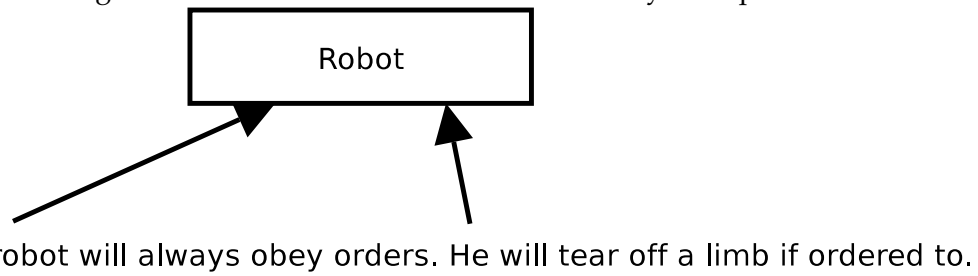


Figure 2.2: A model of the world described by example 2.22



anaphoras simply a anaphoric use of the English word *one*. Instead, a *one*-anaphor is usually defined in terms of its semantic function rather than its lexical form.

This section introduces, and gives examples of, the various phenomena described as *one*-anaphora. It also discusses the four main discourse functions assigned to *one*-anaphora by various authors: contrasting with their antecedent; belonging to a set introduced by their antecedent; sharing a sense introduced by their antecedent; and substituting for an existing piece of text.

Contrastive *one*-anaphors

The function of contrastive *one*-anaphora is to contrast the sense or reference of the *one*-anaphor with its antecedent in some way. In example 2.23, a contrast between small dogs and large dogs is effected by the antecedent *a small dog* and the *one*-anaphoric noun phrase *a large one*.

(2.23) I prefer a small dog to *a large one*.

Halliday & Hasan (1976) discuss this use of *one*-anaphora as part of a phenomenon they call **comparative reference**, and use the following examples to illustrate a *one*-anaphor used for comparative reference:

(2.24) It's the same cat as *the one we saw yesterday*. (Halliday & Hasan 1976, p. 78)

(2.25) It's a similar cat to *the one we saw yesterday*. (Halliday & Hasan 1976, p. 78)

(2.26) It's a different cat to *the one we saw yesterday*. (Halliday & Hasan 1976, p. 78)

(2.27) The little dog barked as noisily as *the big one*. (Halliday & Hasan 1976, p. 82)

The definition of reference used by Halliday & Hasan (1976) encompasses both relationships between textual items like that illustrated in Figure 2.1 on page 18 and relationships between textual items and the world like those illustrated in Figure 2.2 on page 18, so neither the *one*-anaphor anaphor nor the antecedent need to be referential in the conventional sense for there to be a relationship of comparative reference between them.

Luperfoy (1991) describes contrastive *one*-anaphora as those where the *one*-anaphor is modified, partitioning a set and comparing two subsets of it. An example of a contrastive *one*-anaphor is shown in example 2.28:

(2.28) When I find red hats I always buy *the one that fits*. (Luperfoy 1991, p. 118)

The second set may be a strict subset of the first as in examples 2.29 and 2.30, or both the first and second subset might be part of a larger set, as in example 2.31.

(2.29) Kate has 14 black Labrador puppies. Jack has reserved $\left\{ \begin{array}{c} \textit{the playful ones} \\ \textit{the ones who are playful} \end{array} \right\}$. (Luperfoy 1991, p. 123)

(2.30) Kate has a black Labrador Retriever. Jack raises $\left\{ \begin{array}{c} \textit{big ones} \\ \textit{ones that weigh 50lb or more} \end{array} \right\}$. (Luperfoy 1991, p. 123)

(2.31) Kate has a black Labrador Retriever. Jack raises $\left\{ \begin{array}{c} \textit{yellow ones} \\ \textit{ones with yellow fur} \end{array} \right\}$. (Luperfoy 1991, p 123)

Dahl (1985) also gives examples of contrastive *one*-anaphors:

(2.32) Bill said you wanted a cinnamon donut. No, I wanted *a chocolate one*. (Dahl 1985, p. 131)

(2.33) Leptons are particles that respond to the weak nuclear force but not *the strong one*. (Dahl 1985, p. 132)

Luperfoy (1991) observes that contrastive *one*-anaphors have the plural *ones*, as illustrated in the following examples:

(2.34) When I find red hats I always buy *the ones that fit*. (Luperfoy 1991, p. 118)

(2.35) When I find red hats I always buy *large ones*. (Luperfoy 1991, p. 188)

Furthermore, *ones* cannot occur at the head of an unmodified noun phrase, as shown in the unacceptable sentence in example 2.36.

(2.36) * When I find red hats I always buy *ones*. (Luperfoy 1991, p. 118)

Member *one*-anaphors

This use of *one*-anaphora indicates that the referent of the *one*-anaphor is a member of a set introduced by the antecedent. In example 2.37, the *one*-anaphor selects a single tree from the set of trees introduced by the antecedent *fourteen trees*.

(2.37) **Fourteen trees** were cut down, we started with *the one at the bottom of the hill*.

Luperfoy (1991) calls this use of *one*-anaphora the **representative sampling** use. She gives the following examples:

(2.38) Howard collects starfish. Thelma gave him *one*. (Luperfoy 1991, p. 131)

(2.39) Dogs are on my lawn. And *one* is barking from across the street. (Luperfoy 1991, p. 133)

Dahl (1985, p. 97) also identifies this function of unmodified noun phrases headed by *one*, but considers that modified *one*-anaphors have essentially the same function, as in this example:

(2.40) Dogs eat too much. I know. I just finished feeding *an especially hungry one*. (Dahl 1985, p. 97)

Luperfoy (1991) observes that while there are many acceptable plurals of the member *one*-anaphor as illustrated in the sentences in example 2.41, *ones* is not among them, as illustrated by the unacceptable sentence in example 2.42.

(2.41) Howard collects starfish. Thelma gave him $\left\{\begin{array}{c} \textit{one} \\ \textit{some} \\ \textit{several} \\ \textit{fourteen} \\ \textit{very few} \\ \textit{four pounds} \\ \textit{a handful} \end{array}\right\}$. (Luperfoy 1991, p. 131)

(2.42) Howard collects starfish. * Thelma gave him *ones*. (Luperfoy 1991, p. 131)

Halliday & Hasan (1976) seem to consider examples like these substitute *one*-anaphora rather than member *one*-anaphora. They consider that the relationship is between the word *one* and the head of the antecedent, rather than between the referent of the *one*-anaphor and a set introduced by its antecedent. The following discussion of substitute *one*-anaphors will describe this further.

Sense sharing *one*-anaphors

This is a category described by Luperfoy (1991), and she places in it many *one*-anaphors that other authors would classify as substitute *one*-anaphors (see the following discussion of substitute *one*-anaphors). A sense sharing *one*-anaphor, which Luperfoy (1991) calls **a new specimen of a salient kind** introduce a new entity, which has some sense introduced into the discourse earlier.

Luperfoy (1991) gives the following examples of sense sharing *one*-anaphors:

(2.43) All the officers wore hats so Joe wore *one* too. (Luperfoy 1991, p. 142)

(2.44) Kelly is seeking a unicorn and Millie is seeking *one* too. (Luperfoy 1991, p. 142)

(2.45) Dad drank a beer slowly and I drank *one* fast. (Luperfoy 1991, p. 143)

Observe that in example 2.45, the beer that the speaker drinks is a different beer from the one that the speaker's father drinks. This differs from Luperfoy's (1991) representative sample *one*-anaphora, because the antecedent does not introduce a set of which the *one*-anaphor's referent is a part.

Luperfoy (1991) observes that the plural of the sense sharing *one*-anaphoris *them*, but not *ones*, as illustrated by the acceptable sentence in example 2.46 and the unacceptable sentence in example 2.47.

(2.46) Dad drank beers slowly, and I drank *them* fast. (Luperfoy 1991, p. 143)

(2.47) * Dad drank beers slowly, and I drank *ones* fast. (Luperfoy 1991, p. 143)

Substitute *one*-anaphors

Some authors describe the function of *one*-anaphora, or at least of some *one*-anaphors, as being one of substitution. The *one*-anaphor takes the place of certain pieces of text, and resolving a *one*-anaphor is the process of finding out which piece of text the *one*-anaphor is taking the place of.

In this model, the process of resolving the *one*-anaphor *a small one* in example 2.48 is simply the process of transforming it into the sentence in example 2.49.

(2.48) I need a large hammer and *a small one*.

(2.49) I need a large hammer and a small hammer.

Halliday & Hasan (1976) describe a subset of *one*-anaphora as **nominal substitutes**. A substitute *one*-anaphor is simply the use of *one* to "stand in" for a meaningful head. In example 2.48, *one* is simply standing in for *hammer*. Halliday & Hasan (1976) argue that a nominal substitute *one* can only ever stand in for the head property of its antecedent. They would argue, for example, that *one* in example 2.50 substitutes for *lollipop* and not for *red lollipop*. Therefore, in the substitute model of *one*-anaphora, the correct resolution of *one* in example 2.50 is example 2.51, not example 2.52.

(2.50) I was handing out **red lollipops** and only gave a single *one* to Tom.

(2.51) I was handing out **red lollipops** and only gave a single *lollipop* to Tom.

(2.52) I was handing out **red lollipops** and only gave a single *red lollipop* to Tom.

Dahl (1985) proposes a similar use of *one*-anaphora when she describes *one*-anaphors with unmodified heads that do not refer as acting simply as "an NP place-holder", although she does not suggest that they are placeholders only for the head of a noun phrase. In example 2.53, neither *a job* nor *one* has a referent in the world described by the sentence, and Dahl (1985) therefore considers the word *one* to be simply standing in for the word *job*.

(2.53) Bill doesn't have a job and Ann doesn't have *one* either. (Dahl 1985, p. 97)

Webber (1979) places all *one*-anaphors in this class by defining *one*-anaphors as phrases which a speaker substitutes for a description. For instance Webber (1979) analyses *one* as substituting for the phrase "cotton T-Shirt" in example 2.54:

(2.54) Some cotton T-Shirts are expensive but not the *one* Wendy gave Bruce yesterday. (Webber 1979, p. 3-1)

2.3 Resolution of *one*-anaphora

This section briefly reviews some of the body of literature on anaphora resolution techniques. This review illustrates general anaphora resolution techniques that are potentially useful for *one*-anaphora resolution, as well as techniques specifically designed to resolve *one*-anaphors. Section 2.3.1 considers techniques used for finding antecedent noun phrases, and Section 2.3.2 considers techniques used for relating the chosen antecedent to the meaning of the *one*-anaphor.

2.3.1 Techniques for finding antecedent noun phrases

This section reviews techniques for finding the antecedent of a *one*-anaphor. The techniques reviewed here can be divided into two kinds: techniques that arise from resolution of other types

of noun phrase anaphora, notably pronouns; and techniques that are specifically suggested for finding the antecedents of *one*-anaphors.

Pronouns, like *one*-anaphors, and unlike definite noun phrase anaphora, lack a head property that could be used to find the antecedent noun phrase. Consider the relationship between the antecedent and anaphor in example 2.55 and that in examples 2.56 and 2.57:

(2.55) I suspected **the local zoo** would be closed, and indeed a sign informed me that *the zoo* was closed.

(2.56) I suspected **the zoo** would be closed, and indeed a sign informed me that *it* was closed.

(2.57) **The doors** were closed and *one* bore a notice of the zoo's opening hours.

Determining that *the zoo* in example 2.55 co-refers with its antecedent *the local zoo* is made easier by the fact that they share the head *zoo*—Vieria & Poesio (2000) showed that 30% of definite noun phrases in their corpus shared a head with their antecedent (50% definite noun phrases in their corpus had no antecedent at all). Resolving *it* in example 2.56 and *one* in example 2.57 is more difficult: they both have less information about their type because they both lack a meaningful head.

There is extensive literature on techniques for resolving pronominal anaphora. Some techniques for finding the antecedents of pronouns exist that require only comparatively shallow analysis of the text, and which therefore may be suitable for finding antecedents of other information poor anaphora, such as *one*-anaphora. Such analysis might be purely syntactic, like that of Lappin & Leass (1994), or even shallower, based purely on part of speech tags, like that of Kennedy & Boguraev (1996).

Lappin & Leass (1994) describe a set of rules for assigning scores to noun phrases based upon their syntactic role in the text, together with their proximity to the pronoun itself. High scoring noun phrases are more likely to be the antecedent of the pronoun than lower scoring noun phrases. Their scoring system attempts to model the **salience** of each noun phrase—its prominence in the discourse—without needing to build a model of the discourse itself.

For example, Lappin & Leass's (1994) algorithm would assign a score of 210 to the noun phrase *Angela* and 180 to the noun phrase *noises* in example 2.58, because *Angela* is a non-adverbial head noun in subject position, and *noises* a non-adverbial head noun in object position.

(2.58) *Angela* heard *noises*.

Kennedy & Boguraev (1996) use a scoring system based on that of Lappin & Leass (1994), but required less analysis of the text. Rather than requiring a full syntactic analysis of the text, Kennedy & Boguraev's (1996) scoring system requires a simpler analysis based on part of speech tags and morphosyntactic tags, rather than a full parse tree.

The scoring systems of Lappin & Leass (1994) and Kennedy & Boguraev (1996) are intending to model discourse salience—the most prominent concepts in the discourse at the time the anaphor

needs to be resolved. The most salient noun phrase is assumed to be the antecedent. Dahl (1985) uses a similar concept—the concept of **topic** in her discussion of finding the set to which the referent of a *one*-anaphor belongs.

Dahl (1985) shows that the referent of a *one*-anaphor is a member of the set which is the current topic. The topic is the current focus of the discourse—the aim of the discourse is for the speaker to tell the listener more about whatever the topic is. Consider Dahl's (1985) example 2.33, an example of a contrastive *one*-anaphor, again:

(2.33) Leptons are particles that respond to the weak nuclear force but not *the strong one*. (Dahl 1985, p. 132)

Here *the strong one* refers to a subset of the set of nuclear forces, which is the topic at the time.

Unfortunately, the most reliable indicators of topic, as Dahl (1985, p. 114) points out, are stress and intonation, features that are only available if the utterance was spoken and you have access to either the original speech or to a transcription with intonation markings.

Webber's (1979) work is on developing a representation of *one*-anaphors that would facilitate finding antecedents, rather than on the process of finding antecedents itself. She imposes the following requirements on a representation of candidate antecedents:

noun phrase preservation since the antecedent of a *one*-anaphor need not be a referential noun phrase, all noun phrases must be accessible to the antecedent finding algorithm;

head distinction since the head property of the antecedent is always inherited by the *one*-anaphor, the representation must at least preserve the special status of the head;

word sense disambiguation semantically ambiguous items should be represented by their sense as well as their surface form. For example, the word *plant* would need to be marked as either the botanical sense of *plant* or the industrial sense of *plant*;

representations of definite pronouns in both resolved and unresolved form

Webber (1979) suggests two possible representations that fulfil these criteria: a representation based upon surface syntactic structure of the antecedent, and a formal semantic representation. Ferrández, Palomar & Moreno (1998) propose another representation using their Slot Unification Grammar (Ferrández, Palomar & Moreno 1997) which is a level of representation somewhere between the two discussed by Webber (1979).

Luperfoy (1991, pp. 255-263) gives an algorithm for *one*-anaphora interpretation. The antecedent finding algorithm she uses is constraint based, and eliminates candidates which are inappropriate antecedents for various reasons. The *one*-anaphor antecedent finding constraints are a subset of those Luperfoy (1991) uses for finding the antecedents of pronouns and definite noun phrase. She excludes constraint checking based on number and gender agreement, but other constraints used

by the system include: eliminate candidates that are not in focus in the local context; semantic type checking; and checking that the candidate antecedent is not embedded in inappropriate logical structures.

2.3.2 Techniques for finding the meaning of a *one*-anaphor

This section discusses techniques for using a chosen antecedent to determine the meaning of a *one*-anaphor. The problem of determining the meaning of a *one*-anaphor is that of determining which of the properties of the antecedent apply to the *one*-anaphor.

The antecedent of *a brown one* in example 2.59 is *several speckled red hens*. *[A] brown one* inherits the head sense ('hen') from *several speckled red hens*, but is the brown hen 'red', 'several' or 'speckled'?

(2.59) There were **several speckled red hens** and *a brown one*.

The solution to this problem is not extensively discussed in the literature. Halliday & Hasan (1976) give one possible method for resolving at least their substitute *one*-anaphors: simply use the head property of the antecedent as the head property of the *one*-anaphor, and discard the remainder of the properties. While this produces a sensible result—in the case of example 2.59 it simply suggests that *a brown one* means a brown hen—it may not produce a complete model of the properties.

Luperfoy (1991, p. 214) simply states that this problem is non-linguistic, and seems to suggest that a reasoning approach using whatever knowledge base the system has available is the natural approach.

2.4 Conclusions

This chapter has reviewed descriptions of non-anaphoric uses of *one*, of *one*-anaphors and their function, and of methods of resolving *one*-anaphors. The three problems for this project identified in Section 1.3 were : identifying *one*-anaphors, finding antecedents of *one*-anaphors, and finding the meaning of *one*-anaphors. In each case, the existing literature suggests some approaches to the problem but does not suggest a full solution.

The accounts of non-anaphoric uses of *one* have a large amount of overlap, and strongly suggest that at least the numeric *one*, generic *one* and object *one* are distinct uses of the word *one*. However, the criteria for distinguishing them is non-computational in nature, and so the problem of developing a program that distinguishes *one*-anaphors from other uses of *one* remains.

The accounts of finding an antecedent suggest several approaches to the problem, mainly based on checking the semantic compatibility of various candidate antecedents. However, the literature does not describe either the design or implementation of a knowledge base structure that makes this possible.

There is very little work on using linguistic resources for sense resolution.

Chapter 3

Identifying *one*-anaphora

This chapter discusses the problem of identifying *one*-anaphora in real text. The problem of identifying *one*-anaphora is the problem of determining for any given use of the word *one* in text whether that use of *one* is a *one*-anaphor or if it has a different, non *one*-anaphoric role in the text.

As seen in the previous chapter, several authors have attempted to describe the various uses *one* might have in English text. It should be clear at least that there are several non-anaphoric uses of *one*. Further discussion of non-anaphoric uses of *one* can be found in Appendix A, Section A.1.

This chapter discusses the process of distinguishing *one*-anaphora from other uses of *one* in English. Section 3.1 discusses the several common uses of *one* in English. Section 3.2 discusses approaches to the problem of distinguishing these use of *one*. Section 3.3 describes a prototype system which uses syntactic information provided by the Connexor parser together with some limited semantic information to distinguish uses of *one* in text from one another. Section 3.4 provides the results of testing this system on approximately 700 uses of *one* in the British National Corpus. Section 3.5 concludes the chapter with a discussion of some of the errors made by this system, and the type of knowledge the system would need to successfully classify those particular uses of *one*.

3.1 Uses of *one* in English

The first part of distinguishing uses of *one* in English is that of describing what these uses actually are. This section gives an overview of the non-anaphoric uses of *one* in English—a more complete analysis of the English uses of *one* is given in Appendix A.

Section 3.1.1 is a brief discussion of some examples of *one* used in the British National Corpus. Section 3.1.2 outlines the taxonomy of uses of *one* described in Appendix A.

3.1.1 Example uses of *one* in English

Here are three sets of sentences taken from the British National Corpus: set 3.1 contains uses of *one* that are clearly not *one*-anaphoric; set 3.2 contains uses of *one* that are *one*-anaphoric; and set 3.3 contains uses of *one* where additional context is needed for a reader to decide whether or not they are *one*-anaphoric:

(3.1)
a. Each string portrays *one* state of a target system... (BNC file FNR sentence 2309)
b. This wall ran along *one* side of the Stratford Road and the residents called it The Front. (BNC file AMC sentence 185)
c. A chorus of fists, plates, and in *one* case, teeth, greeted his rise. (BNC file ALL sentence 344)
d. *One* might be excused for wondering what the aged poet really believed. (BNC file CAW sentence 843)
e. The European Script Fund is the only *one* of the MEDIA programme initiatives to be based in the UK (in London). (BNC file A0E sentence 629)

(3.2)
a. He's not a carpenter, dear, he's **an architect**, and a highly respected *one* at that. (BNC file H8H sentence 411)
b. For most of us , our day to day experience of the environment has changed from **an essentially static, rural condition** to a highly mobile technological *one* in less than two centuries. (BNC file CL0 sentence 895)

(3.3)
a. Only *one* of these, that of the designed environment, has been adopted in Britain, and infrequently at that. (BNC file C8F sentence 984)
b. Be careful lads not to let this *one* slip away! (BNC file J1H sentence 3423)
c. And secondly she's actually recently had *one* so we need to make sure we completely eradicate it. (BNC file GYD sentence 94)

Given the limited context provided above, it is hard to tell if the sentences in set 3.3 have antecedents. For example, it is possible that example 3.3a occurs in a context where *one* has no antecedent, but is intended to be a generic reference to a person or type of person.

The aim of the work described in this chapter is to distinguish uses of *one* like those in 3.1 from uses of *one* like those in sets 3.2 and 3.3. Once this identification has been done, the anaphoric examples can be processed by a system like that described in Chapter 4.

3.1.2 Describing uses of *one* in English

This section provides an overview of what *one*-anaphora and the other uses of *one* in English are. It is possible to distinguish some of these uses syntactically, others must be distinguished by their semantic function.

Section 2.1 provided an overview of existing categorisations of the word *one*. Features used in the literature to distinguish uses of *one* in English include: various types of cohesive function; plurals; word class; contrasting words; and various semantic functions. My own taxonomy of uses of the word *one*, which uses many of these same features to produce a more detailed taxonomy of uses of *one*, is presented in detail in Appendix A. In summary, there are four major non-anaphoric uses of *one* in English:

Numeric the "*one* ball" usage, which indicates singularity, "one ball" as opposed to "two balls";

Generic the "*one* doesn't like that" usage, sometimes referred to as the **impersonal pronoun**, used to refer to a generic person or to the speaker of a sentence;

Object the "I'll have *a one* and a ten please" usage, used, for example, when referring to a one dollar bill as "a one"; and

Partitive the "*one* of the animals bit me" usage, used to select an individual from a set.

Here are the non-anaphoric examples from the introduction categorised by type:

Numeric: Each string portrays *one* state of a target system...
This wall ran along *one* side of the Stratford Road and the residents called it The Front.
A chorus of fists, plates, and in *one* case, teeth, greeted his rise.

Generic: *One* might be excused for wondering what the aged poet really believed.

Partitive: The European Script Fund is the only *one* of the MEDIA programme initiatives to be based in the UK (in London).

3.2 Approaches to distinguishing uses of *one* in English

This section discusses approaches to the problem of distinguishing *one*-anaphora and other uses of *one* in English from each other. There are five possible results of such a distinction: the use of *one* might be numeric, generic, objective, anaphoric, or partitive.

In practise, the "object" use of *one* is so rare—it occurred in none of the data sets annotated—that I did not develop heuristics that uniquely identify it, so these heuristics distinguish only four uses of *one*: numeric, generic, anaphoric, and partitive.

The first part of this section describes the different parts of speech tags assigned by the CLAWS tagger to different uses of *one* and discusses the extent to which these allow uses of *one* to be distinguished. The second part describes the different grammatical functions assigned to the word *one* by the Connexor parser, and discusses the extent to which these can be used to distinguish uses of *one*. The third section discusses some addition semantic and pragmatic cues that my system makes use of when syntactic cues are insufficient.

3.2.1 Distinguishing uses of *one* by part of speech

Part of speech assignment is a common pre-processing technique used by natural language systems. The British National Corpus, from which I drew my experimental data, provides automatically assigned part of speech tags for every word in the corpus. (See Section 1.4.1 for a more detailed description of the BNC.) Part of speech tags are a useful way of gaining some understanding of the roles words play in a text without needing a deep understanding of the meaning of the text. Hence, when seeking heuristics for identifying *one*-anaphora in a text, I began by using rules based on part of speech tags.

A perfect assignment of parts of speech would always correctly identify the numeric uses of *one*, as they are the only adjectival uses of *one* in English. Most parts of speech taggers use a sufficiently fine-grained tag set that numeric adjectives (*one*, *two*, *three*. . .) are marked with a special tag. Unfortunately, part of speech taggers do not assign tags completely accurately, so the part of speech tag assigned to a use of *one* cannot be taken as completely reliable evidence of the function of that use of *one*.

Example 3.4 shows a use of *one* that is mis-tagged: *one* in example 3.4 is tagged by the CLAWS tagger as PNI (indefinite pronoun) when it should have been tagged as CRD (cardinal numeral):

(3.4) "Higher quality at lower prices" is the reassuring promise made in their advertisements and a ten year guarantee—as opposed to the normal *one* year guarantee—is what Wilson & Glick are committed to. (BNC file AAY, sentence 210)

Even a perfect assignment of part of speech tags would not suffice to distinguish all uses of *one* from each other. For example, the generic use of *one* and *one*-anaphors PNI (indefinite pronoun) tag even if the tagging is perfect. Example 3.5 shows an example of a generic use of *one* that is impossible to distinguish from a *one*-anaphor on the basis of its part of speech tag, even though it is correctly assigned a PNI tag by the CLAWS tagger:

(3.5) To be really modern *one* should have no soul. (BNC file A6D, sentence 1077)

Even given that the tags are not completely reliable, they are a useful way of subdividing the quarter of a million sample uses of *one* in the BNC. In Appendix B I discuss in detail the distribution of parts of speech assigned to *one* in the BNC corpus, and identify several sets of uses of *one* in the BNC which are likely to be rich in *one*-anaphora:

1. a word marked as a general determiner followed by *one* marked as indefinite pronoun;
2. a word marked as an article followed by *one* marked as an indefinite pronoun;
3. a word marked as an adjective followed by *one* marked as an indefinite pronoun;
4. a word marked as an ordinal followed by *one* marked as an indefinite pronoun;

5. a word marked as a wh-determiner followed by *one* marked as an indefinite pronoun;
6. *one* marked as an indefinite pronoun followed by a word marked as a modal auxiliary verb;
7. *one* marked as an indefinite pronoun followed by a word marked as a wh-determiner; and
8. *one* marked as an indefinite pronoun followed by a word marked as a wh-pronoun.

Hand annotation of 579 occurrences of *one* in set 3 shows that a large majority of occurrences of *one* in this set are in fact *one*-anaphors (see Table 3.1).

Type	Percentage of occurrences
Numeric	1.04 %
Generic	2.42 %
Part of	2.25 %
Idiomatic	2.94 %
Anaphor	85.66 %
Unclassifiable	4.66 %

Table 3.1: Success at identifying uses of *one* English based on part of speech

The error rate of 14.34% is caused by the two factors discussed above: errors in the part of speech tagging; and the fact that part of speech tagging is insufficient for perfect discriminating types of *one*. So part of speech tagging provides a useful way to identify the numeric uses of *one* and to extract data sets likely to be rich in *one*-anaphora, but is insufficient for distinguishing all the uses of *one* described in the previous section.

3.2.2 Distinguishing uses of *one* by grammatical role

Given that part of speech tags alone are insufficient for distinguishing use of *one* in English, this section discusses a further set of heuristics that can be used to distinguish uses of *one* in English: grammatical role. These heuristics are based on the results of grammatical analysis by the Connexor English parser. (Connexor 2003)

Again, as when part of speech tags are used to discriminate uses of *one*, the numeric use of *one* is easy to distinguish by grammatical role. The numeric use of *one* is the only use of *one* that is not found at the head of a noun phrase. However, unlike when part of speech tags are used, some other uses of *one* have distinct grammatical roles assigned by Connexor.

A more complete analysis of functional roles assigned to uses of *one* is presented later in the chapter (see Table 3.3 on page 36) but the following grammatical roles distinguish uses of *one*:

the quantifier dependency relation is used to distinguish numeric uses of *one*;

the pre-modifying quantifier functional role is used to distinguish numeric uses of *one*;

the post-modifying prepositional phrase role is used to partially distinguish partitive and *one*-anaphoric uses of *one* from other uses; and

the pronoun morphological tag is used to distinguish anaphoric and generic uses of *one*.

These heuristics are better than part of speech based heuristics, because they give us the ability to discriminate partitive and generic uses of *one*. However, the problem of separating anaphoric uses of *one* from either partitive or generic uses remains.

One particularly notable problem with these heuristics is distinguishing the partitive use of *one* like that in example 3.6 from an anaphoric use of *one* like that in example 3.7.

(3.6) Beside the Cross *one* of the men had torn a paper into little bits and scattered them, to a groaning catcall from the crowd. (BNC file `A0N`, sentence 350)

(3.7) The problem was the unusual *one* of a warmish (well, for the plants anyway), wet spring. (BNC file `A0G`, sentence 1689)

In example 3.6 the referent of *one* is a man and we can deduce the sense **man** from within the noun phrase *one of the men*. In example 3.7, *one* has the sense **problem**, and we need to use the antecedent noun phrase *the problem* to deduce this. But the fact that example 3.6 is partitive and example 3.7 *one*-anaphoric is not obvious from the syntax: in both cases, *one* is post-modified by a prepositional phrase beginning with *of*—*of the men* in the case of example 3.6 and *of a warmish... spring* in the case of example 3.7.

There is a simple, but not foolproof, syntactic test that can be used here. Observe that *men*, head of the noun phrase embedded in *of the men* is plural whereas *spring*, head of the noun phrase embedded in *of a warmish... spring*, is singular. Partitive uses of *one* are selecting an individual from a set, and therefore the set, embedded in the prepositional phrase, needs to be plural. *One*-anaphora, by definition, never select an individual from a set that is referred to in the same noun phrase, and therefore the head of the embedded noun phrase need not be plural.

Another similar, but also not foolproof, test allows some *one*-anaphors to be distinguished from generic *ones*. Generic *one*, as shown in the analysis in Table 3.3 on page 36 almost always occurs in subject position in a sentence. *One*-anaphors do occur in subject position, but also occur frequently in other contexts: for example, as the subject complement. This suggests the following heuristic: *one* marked as a pronoun that does **not** have a subject role is a *one*-anaphor, not a generic *one*.

Grammatical roles are a better way of distinguishing uses of *one* in English than part of speech tags are. The largest hole in the set of heuristics outlined in this section is this: if *one* is tagged as a pronoun by Connexor and is in subject position it is impossible to tell from the grammatical role whether *one* is a *one*-anaphor or a generic *one*.

3.2.3 Distinguishing uses of *one* using semantic information

Given the heuristics described in previous sections, the remaining problem is that of deciding whether a use of *one* in subject position is generic or *one*-anaphoric. My heuristics for deciding this rely on certain discourse features that tend to occur with use of the generic *one*.

I use two pieces of information about the generic *one* to distinguish it from *one*-anaphors in subject position. The first is the the generic *one* refers, at least abstractly, to a person. The second is that the generic *one* is normally used in very formal contexts, and tends to be used in particular discourse structures.

The fact that the generic *one* refers to a person and is often in subject position suggests the following test: use the main verb of the sentence to identify generic uses of *one*. In particular, the generic use of *one* will tend to be the subject of sentences that have a verb requiring an animate subject, such as *say* and *think*.

The approximation used in the system described below is very rough. It doesn't draw on a systemic verb ontology to decide whether or not the main verb of the sentence requires an animate subject. Instead, I built a short list of such verbs, and the system checks if the main verb is in that list.

The tendency of the generic *one* to be used in formal contexts means that it quite frequently occurs in sentences like the following two examples:

(3.8) In total *one* can only marvel at the numerous stations opened in the 1980s and it is pleasing to note that many more are in the pipeline. (BNC file `A11` sentence 1459)

(3.9) One's overall sense of The Possessed absolutely refuses to confirm any such duality, and *one* can pay the novel no simpler or fuller tribute than by saying so. (BNC file `A18` sentence 1202)

In both these examples, a generic use of *one* is the subject of a main verb which is the head of a verb chain. The auxiliary verb (*can* in these two examples) is the immediate parent of the generic use of *one*. The generic *one* is frequently headed by qualifiers like *can*, *could*, and *would*, and thus an additional heuristic for identifying a generic *one* is testing whether its immediate head is an auxiliary verb such as these.

3.3 System overview

Based on the reasoning given in the previous sections, I developed a system which distinguishes different usages of *one* in English, which is outlined here.

For a given sentence containing the word *one*, the system I developed uses the following heuristics to classify the use of *one*:

1. if the functional role of *one*'s immediate dependency is a post-modifying phrase beginning with *of*, and the embedded noun phrase has a plural head, the use of *one* is **partitive**;
2. if the functional role of *one*'s immediate dependency is a post-modifying phrase beginning with *of*, and the embedded noun phrase has a singular head, the use of *one* is **anaphoric**;
3. if the functional role of *one* is that it is a pre-modifying quantifier, the use of *one* is **numeric**;
4. if the *one* is marked as a numeric adjective, the use of *one* is **numeric**;
5. if the *one* is in subject position and:
   (a) it depends on one of the words *might*, *may*, *should*, *could* or *must*; or
   (b) it depends a verb which is part of a verb chain; or
   (c) it depends on a verb which is one of the animate verbs listed in Section 3.2.3;

   it is generic; otherwise
6. it is *one*-anaphoric.

3.4 Results

This section gives three sets of results testing the ability of techniques described in this section to discriminate different uses of *one* in English. Section 3.4.1 discusses the distribution of types of *one* in three test data sets drawn from the British National Corpus. Section 3.4.2 discusses the distribution of grammatical roles among the uses of *one*. The results from this part provided many of the heuristics described in Section 3.2.2. Section 3.4.3 gives the result of testing the system of heuristics described in Section 3.3 on the data sets described below.

3.4.1 Test data sets

I tested the system by comparing its results with the results of annotating the type of each use of *one* myself. This subsection describes the data used to test the system detailed above, and gives the results of the hand-annotation, listing the distribution of types of *one* in each of the three data sets used to test the system

Appendix B discusses different subsets of the British National Corpus likely to contain *one*-anaphora. One of these sets—the set of *one* tagged as an indefinite pronoun, preceded by a word tagged as an adjective—was discussed in Section 3.2.1, and is among the sets named in Appendix B as containing many *one*-anaphors. However, precisely because this set is rich in *one*-anaphora, it doesn't provide many examples of numeric, generic, or partitive uses of *one* to test the above system on. Therefore, I selected two more sets of data with a more even distribution of types of *one*.

All three sets of data were drawn from the British National Corpus. The sets were:

1. 573 uses of *one* from the set where *one* is marked as an indefinite pronoun and is preceded by an adjective;
2. 100 uses of *one* from the set where *one* is marked as a cardinal numeral and is preceded by a coordinating conjunction; and
3. 100 uses of *one* from the set where *one* is marked as an indefinite pronoun and is preceded by a coordinating conjunction.

The first set is the one chosen specifically to yield a high number of *one*-anaphors. Because a coordinating conjunction (the word *and* is an example of a coordinating conjunction) is essentially free of meaning, the latter two sets yield a more even distribution of generic, numeric, partitive and anaphoric uses of *one*.

The results of the hand annotation of the distribution of uses of *one* in each of the sets are summarised in Table 3.2.

Data set	Numeric		Generic		Partitive		Anaphoric		Idiomatic		Uncertain	
	#	%	#	%	#	%	#	%	#	%	#	%
Adjective + indefinite pronoun *one*	6	1.05	15	2.62	13	2.27	495	86.39	17	2.97	27	4.71
Conjunction + cardinal numeral *one*	51	51.00	1	1.00	41	41.00	7	7.00	0	0.00	0	0.00
Conjunction + indefinite pronoun *one*	2	2.00	45	45.00	0	0.0	52	52.00	0	0.00	0	0.00

Table 3.2: Distribution of uses of *one* in the test data.

3.4.2 Analysis of the distribution of grammatical roles among the uses of *one*

This subsection describes the grammatical roles played by different types of *one* in the sentences in which they occur. Table 3.3 shows the distribution of grammatical roles among each of the uses of *one* in each of the data sets.

Some of the more indicative grammatical roles are:

pre-modifying quantifier which occurs 60 times, and which indicates a numeric use of *one* 88% of the time;

post-modifier which occurs 31 times, and which indicates a partitive use of *one* 52% of the time and a *one*-anaphor 42% of the time; and

subject which occurs 300 times, and which indicates a *one*-anaphor 73% of the time and a generic *one* 17% of the time.

Adjective+ indefinite pronoun *one*					
	Numeric	Generic	Partitive	Anaphoric	Idiomatic
Pre-modifying quantifier	2	0	0	3	1
Object	1	0	4	62	3
Preposition complement	2	2	0	80	4
Subject complement	1	2	0	138	1
Post-modifier	0	1	6	6	0
Subject	0	10	2	189	7
Stray NP	0	0	1	2	0
Adverbial	0	0	0	1	1
Apposition	0	0	0	14	0
Conjunction+ cardinal numeral *one*					
	Numeric	Generic	Partitive	Anaphoric	Idiomatic
Pre-modifying quantifier	50	0	0	1	0
Subject	1	1	18	2	0
Stray NP	0	0	2	2	0
Adverbial	0	0	1	0	0
Object	0	0	2	0	0
Post-modifier	0	0	10	0	0
Post-modifier (beg. *of*)	0	0	1	0	0
Subject complement	0	0	7	2	0
Conjunction+ indefinite pronoun *one*					
	Numeric	Generic	Partitive	Anaphoric	Idiomatic
Pre-modifying quantifier	1	2	0	1	0
Subject complement	1	0	0	14	0
Subject	0	42	0	28	0
Post-modifier	0	1	0	7	0
Object	0	0	0	2	0

Table 3.3: Grammatical function assigned by Connexor to different uses of *one*

3.4.3 Distinguishing uses of *one* in English

This subsection gives the results of using the system described in Section 3.3 to distinguish uses of *one* in English. Recall and precision figures for each of the four categories: numeric, generic, partitive and anaphoric are given in Table 3.4.

	Number in data set	Number identified	Number correct	Recall	Precision
Numeric	59	81	54	91.5%	66.7%
Generic	61	75	37	60.6%	49.3%
Partitive	55	66	46	83.6%	69.7%
Anaphoric	553	543	472	85.4%	86.9%

Table 3.4: Accuracy of the system that identifies uses of *one* in English

3.5 Error analysis

This section shows the number of erroneous classifications by each of the rules given in Section 3.3. Table 3.5 shows the number and percentage of erroneous classifications that each rule made. Table 3.6 shows which uses of *one* were being misclassified by each rule.

Rule matching...	Identifies	Total identified	# incorrect	% incorrect
Quantitative function or numeral morphology	Numeric	60	10	16.7
Postmodifier, singular embedded NP	Anaphor	36	9	25.0
Postmodifier, no singular embedded NP	Partitive	74	28	38.0
Subject headed by aux. verb	Generic	34	25	73.5
Subject headed by 'animate' verb	Generic	13	3	23.0
Subject headed by *might* etc	Generic	22	4	18.1
Default	Anaphor	453	8	1.8

Table 3.5: Error rates by rule

The error analysis in Table 3.5 shows that the default classification of uses of *one* as *one*-anaphors is the most successful rule. One of the rules that is designed to identify generic *one* is by far the least successful. By absolute number of mis-classifications, this rule and the single rule which attempts

	# wrongly identified			
Rule matching...	Numeric	Generic	Partitive	Anaphor
Quantitative function or numeral morphology	n/a	1	0	9
Postmodifier, singular embedded NP	0	0	9	n/a
Postmodifier, no singular embedded NP	0	1	n/a	27
Subject headed by aux. verb	0	n/a	0	25
Subject headed by 'animate' verb	0	n/a	0	3
Subject headed by *might* etc	0	n/a	0	4
Default	4	4	0	n/a

Table 3.6: Erroneous classification by rule

to identify partitive uses of *one* are the most error-prone rules. A large majority of errors that these two rules make are misclassifying *one*-anaphors as partitive and generic *one* respectively, suggesting that the system is better at identifying partitive *one* and generic *one* than it is at identifying *one*-anaphors.

3.6 Conclusions

This chapter presented four sets of results: the results of identifying *one*-anaphors based upon the part of speech, the results of identifying *one*-anaphors based upon grammatical roles, the results of a program which identified *one*-anaphors using a combination of grammatical role and shallow semantic information, and an error analysis of this system.

The main findings of this chapter were details of the grammatical role of the various non-anaphoric uses of *one*. The main computational outcome of this chapter was a system which uses these grammatical roles and some semantic information to distinguish the uses of *one* from each other.

Chapter 4

Resolving *one*-anaphora

This chapter discusses the problem of resolving a *one*-anaphor. The problem of resolving *one*-anaphora can be divided into two subproblems. The first of these is finding the antecedent of a *one*-anaphor: locating a preceding noun phrase from which at least one property can be inferred. The second of these problems is using the antecedent to actually infer the sense of the *one*-anaphor.

Consider the problem of deducing the meaning of *a naughtier new one* in example 4.1.

(4.1) Kylie had cathartically climbed out of **her demure old skin** into *a naughtier new one*—with the world watching. (BNC file ADR, sentence 1599)

The first problem is correctly selecting *her demure old skin* from the preceding text as the antecedent. The second problem is deciding that *a naughtier new one* has the property 'skin' but not the properties 'demure' or 'old'.

This can be a difficult problem, requiring in some cases semantic knowledge which cannot be drawn from existing sources. This chapter explores some properties of *one*-anaphora, their antecedents, and their relationships with their antecedents. I developed systems that solve parts of these problems, and these are also described in this chapter.

This chapter is divided into three sections. Section 4.1 discusses the data set used in the remainder of the chapter, and the other two sections explore the two major problems of resolving *one*-anaphora: Section 4.2 explores the problem of finding the antecedent noun phrase and Section 4.3 explores the problem of determining which properties of the antecedent apply to the referent of the *one*-anaphor.

Each section presents an analysis of the problem based upon the 495 annotated *one*-anaphors from the set of uses of *one* described in Section 4.1—the set where the use of *one* itself was tagged by the CLAWS tagger as a indefinite pronoun and the previous word as an adjective.

4.1 Test data set

This section discusses the test data set—451 examples of *one*-anaphors taken from the British National Corpus with hand-annotated antecedents, and a further 43 examples of *one*-anaphors with no clear single antecedent. These 495 *one*-anaphors are the 495 *one*-anaphors which are marked as indefinite pronouns preceded by an adjective, and are a subset of the 573 uses of *one* so marked that were described in Section 3.4.1.

I annotated these 495 *one*-anaphors a second time, marking the antecedents of each *one*-anaphor. Of the 495 *one*-anaphors, only 451 had an identifiable antecedent. The remainder required that the sense of the *one*-anaphor be inferred from several previous noun phrases. Normally, this occurs when several members of a set are mentioned, followed by a *one*-anaphor which the reader must infer to be part of the same set. Example 4.2 shows an example of a *one*-anaphor with no specific antecedent:

(4.2) For **Diana**, it was like having **an ally** in **the camp**, and **Balmoral** was much more enjoyable **that year** with **Fergie** about. **She** was not alone any more in feeling oppressed by **the strict formality**, **the strict time-keeping**; **after-dinner games** were more lively—and **she** was no longer *the only one* who wanted to giggle at the sound of the bagpipes that played them out of the dining-room after dinner every night. (BNC file A7H, sentence 1361–1362)

The inferred sense of *the only one* in Example 4.2 is something like "the only person at Balmoral." but none of the candidate antecedents highlighted have that sense—in particular, none of the candidates have the head property 'person' or any other head sense that could possibly apply to Diana.

Much of the analysis in the remainder of this chapter focuses on the 451 *one*-anaphors which do have a single identifiable antecedent, rather than attempting to solve the problems of deciding when there is no such antecedent, or of finding the sense of the *one*-anaphor when there is no single antecedent to derive the sense from.

4.2 Finding the antecedent of a *one*-anaphor

The problem of finding the antecedent of a *one*-anaphor is the problem of choosing the noun phrase which gives the *one*-anaphor meaning from all the noun phrases that precede the *one*-anaphor in the text. Example 4.3 shows that, even when limiting the context to the sentence in which the *one*-anaphor occurs, and the previous sentence, there are a lot of noun phrases that could be antecedents.

(4.3) Also, there can be **no doubt** that **the South** would intervene in **the North** on behalf of **the catholic – nationalist minority** if **Britain** withdrew and **civil war** broke out. Equally,

> **Southern catholic nationalists** prefer **a unified Irish state**, though **a majority of 54 per cent** would accept *a federal one* based on the present two state units. (BNC file A07, sentence 68–69)

In this chapter, an antecedent is defined as the noun phrase from which the head sense of the *one*-anaphor can be determined. In example 4.3 the antecedent of *a federal one* is *a unified Irish state*, and its head noun *state* confers 'state' as the head sense of *a federal one*. (For discussion of whether the senses 'unified' and 'Irish' are conferred as senses of *a federal one*, see Section 4.3.)

This section discusses two possible solutions to the problem of identifying the antecedent of a *one*-anaphor: relatively knowledge poor solutions based upon syntactic information and solutions based upon semantic information. Section 4.2.1 discusses the distance in sentences between *one*-anaphors and their antecedents. Section 4.2.2 discusses the syntactic role that *one*-anaphors and their antecedents play in the structure of the sentences they are part of. Section 4.2.3 discusses several systems that use these grammatical roles to choose antecedents for *one*-anaphors. Section 4.2.4 discusses using web searches as a source of semantic information for finding antecedents.

4.2.1 The nearness of *one*-anaphors and their antecedents

One-anaphors and their antecedents are quite close in the text. In the 451 cases where an antecedent existed, that antecedent was in either the sentence containing the *one*-anaphor or the three sentences immediately preceding it and Table 4.1 shows that a large majority of antecedents are in exactly the same sentence as the *one*-anaphor. The average distance between a *one*-anaphor and its antecedent in sentences was only 0.19 sentences, and the median distance was 0 sentences.

This result suggests that as for other anaphora, recency is an important factor when finding antecedents for *one*-anaphor.

Distance	# of occurrences	% of occurrences
0	378	83.8
1	62	13.7
2	9	2.0
3	2	0.4

Table 4.1: Distance in sentences between *one*-anaphor and antecedent

4.2.2 The syntactic role of *one*-anaphors and their antecedents

This subsection discusses the syntactic roles assigned to *one*-anaphors and their antecedents by the Connexor parser, in order to assess whether there are useful regular patterns that would allow

me to develop heuristics for finding antecedents based on syntactic roles similar in spirit to the heuristics developed for finding the *one*-anaphors themselves in Chapter 3.

The five highest frequency assignments of grammatical function to the antecedent of the 451 *one*-anaphors with identifiable antecedents is given in Table 4.2. When the results are split by distance in sentences between the *one*-anaphor and its antecedent, distance does not appear to have a marked effect on the grammatical function of the antecedent, although possibly 451 examples is not a sufficiently large sample size to reveal any patterns of this kind.

	# of occurrences at distance			% of occurrences at distance		
Antecedent function	Any	0	1	Any	0	1
Subject	174	156	13	38.6	41.3	21.0
Preposition complement	90	71	15	20.0	18.8	24.2
Object	89	67	20	19.7	17.7	32.3
Subject complement	46	37	0	10.2	9.8	0.0
(Lone) head of noun phrase	19	17	2	4.2	4.5	3.2

Table 4.2: The five grammatical functions Connexor most frequented assigned to antecedents

Table 4.2 shows that four grammatical roles: subject, preposition complement, object and subject complement account for the grammatical role of the antecedent in 90% of cases. However, a simple rule that stated, for example, "use the preceding subject as the antecedent" would only be successful at most 40% of the time.

The five highest frequency assignments of grammatical function to the 451 *one*-anaphors themselves are shown in Table 4.3.

	# of occurrences with distance			% of occurrences with distance		
One-anaphor function	Any	0	1	Any	0	1
Subject	180	136	39	39.9	36.0	14.5
Subject complement	117	109	6	25.9	28.8	9.7
Preposition complement	77	73	6	17.0	19.3	9.7
Object	55	42	10	12.2	11.1	16.1
Apposition	11	9	2	2.4	2.4	3.2

Table 4.3: The five grammatical functions Connexor most frequented assigned to *one*-anaphors

Table 4.4 shows the most frequent ten pairings of particular antecedent grammatical functions

with particular *one*-anaphor grammatical functions.

		# of occurrences at distance			% of occurrences at distance		
Antecedent function	*One*-anaphor function	Any	0	1	Any	0	1
Subject	Subject complement	79	77	1	17.5	20.4	1.6
Subject	Subject	57	45	9	12.6	11.9	14.5
Preposition complement	Subject	39	29	9	8.6	7.7	14.5
Object	Subject	40	26	13	8.6	6.9	21.0
Preposition complement	Preposition complement	25	22	2	5.5	5.8	3.2
Subject complement	Subject	23	18	5	5.1	4.8	8.0
Object	Object	22	17	5	4.9	4.5	8.0
Object	Preposition complement	21	20	1	4.7	5.3	1.6
Subject	Preposition complement	19	19	0	4.2	5.0	0.0
Subject	Object	12	9	2	2.7	2.4	3.2
Preposition complement	Object	12	8	2	2.7	2.0	3.2

Table 4.4: The ten most frequent pairings of antecedent grammatical functions with *one*-anaphor grammatical functions

Again, Table 4.4 suggests that the success rate of rules that locate likely antecedents based simply on their grammatical function is not likely to be high, with one exception: the strong association between a subject antecedent and a subject complement *one*-anaphor. Of the 117 total subject complement *one*-anaphors, 77, or 65.8%, have an antecedent which is marked as a subject, and is in the same sentence. Three such sentences are given in example 4.4:

(4.4) a. **The Act** is *an enabling one* and does not affect the existing system unless the leadership of the churches wish to co-operate in a reshaping of the system or of schools in any particular area. (BNC file `A07` sentence 1403)

b. **The summer**, which was *a glorious one* that year, gave way to a gusty autumn, and, as is the way with these things, after the autumn came the winter. (BNC file `A08` sentence 1379)

c. . . . the former friends he left behind in the Church of England replied that **the early undivided Church** had been *the only one* wholly to contain this supernatural essence. (BNC file `A1T` sentence 174)

However, this pattern is not necessitated by English syntax, as the *one*-anaphors marked as subject complement that *don't* have subject antecedents shown in example 4.5 illustrate:

(4.5) a. However tough things seem, it's vital that you should keep your spirits up, and remember that acting is **an art** and *a thrilling one*... (BNC file `A06` sentence 1571)

b. . . . and the pilgrimage festival known as Succoth , also known as the Feast of Booths (or Tabernacles or Ingathering) which is **a harvest festival**, *an especially colourful and joyous one*. . . (BNC file A0P sentence 323)

c. In addition to **the secular names**, there is *an additional one* by which the person is known in the synagogue, by which he is called to the Torah. (BNC file A0P sentence 274)

Sentence 4.5c also illustrates an exception to another possible rule that might be inferred from the other sentences in examples 4.4 and 4.5: the noun phrase immediately before the *one*-anaphor (*there*) is not the antecedent of the *one*-anaphor.

The analysis in this subsection shows that rules that would rely on syntactic information alone to find the antecedents of *one*-anaphors are not obvious and would have fairly low success rates. The following subsection confirms this by giving the results of some systems that used various syntactic rules to find the antecedents of *one*-anaphors.

4.2.3 Finding antecedents using syntactic information

This subsection describes several attempts to find a rule set based on syntactic information that finds the antecedents of *one*-anaphora. The systems were:

baseline 1 a system which used the noun phrase immediately preceding the *one*-anaphor as the antecedent;

baseline 2 a system which used the first noun preceding the *one*-anaphor that was marked as the subject as the antecedent;

heuristic a system which used a set of scores based on grammatical functions to find the antecedent; and

l&l a system based on the rules given in Lappin & Leass (1994) to find the antecedent.

A summary of the performance of all systems on the data set of 451 *one*-anaphors which had an identifiable antecedent is given in Table 4.5.

The rules given by Lappin & Leass (1994) assign scores to candidate antecedent noun phrases based on factors like: whether or not they are the subject position, whether they are in the same sentence as the anaphor, and whether or not they are embedded in another noun phrase. They are an attempt to model "structural salience", or prominence in the discourse.

The heuristics I developed were not designed to model the salience of various senses in order to determine whether a noun phrase with that sense is likely to be the antecedent of a *one*-anaphor, they are simply designed to take advantage of some observed patterns in the grammatical relations of antecedent and *one*-anaphor.

System	# antecedents found	% antecedents found
Baseline 1 (previous NP)	133	29.5
Baseline 2 (previous Subj.)	140	31.0
Heuristic	151	33.5
L&L	153	33.9

Table 4.5: Summary of the performance of the semantic knowledge based antecedent algorithms

The complete set of heuristics used in the system I call **heuristic** above is:

- if the *one*-anaphor is marked by Connexor as a modifier of the candidate, add 100 to the candidate's score;
- if the candidate is marked as a subject, add 100 to the candidate's score;
- if the *one*-anaphor is marked by Connexor as a complement, and the candidate itself is marked as a subject, add 50 to the candidate's score;
- if the *one*-anaphor and the candidate are joined by a coordinating conjunction (such as *and*), add 100 to the candidate's score;
- if the *one*-anaphor and the candidate are in the same sentence, add 100 to the candidate's score; and
- if the function of the candidate and the *one*-anaphor are the same (an indicator of syntactic parallelism) add 100 to the candidates score.

As in the system described by Lappin & Leass (1994), the highest scoring noun phrase is selected as the antecedent.

None of the four systems described here are very effective. A reasonable performance rate for antecedent location would be approximately 85%, the result achieved by Lappin & Leass (1994) for pronominal anaphora resolution.

4.2.4 Finding antecedents using semantic information

This section describes a small experiment involving finding the antecedent of a *one*-anaphor using semantic information. This experiment attempted to model the appropriateness of the combination of head sense of the antecedent with the sense of the *one*-anaphor.

In order to illustrate this process, consider example 4.3 again:

(4.3) Also, there can be **no doubt** that **the South** would intervene in **the North** on behalf of **the catholic – nationalist minority** if **Britain** withdrew and **civil war** broke out. Equally, **Southern catholic nationalists** prefer **a unified Irish state**, though **a majority of 54 per cent** would accept *a federal one* based on the present two state units. (BNC file `A07` sentence 68–69)

The choice between the candidate antecedents is something like the choice of the following meanings of *a federal one*: a federal doubt, a federal South, a federal North, a federal minority etc. A substantial proportion of these potential meanings of *a federal one* are nonsensical, and some of the remainder unlikely.

Ideally, given either a sufficiently large corpus or a sufficiently good way of modelling the relative probabilities of having a federal doubt or a federal minority, it should be possible to assign relative weights to the different antecedents based on how likely they are to occur.

One way of estimating the relative likelihood of a text describing a federal doubt as opposed to a federal catholic-nationalist minority is to use the Google search engine to estimate the total number of documents containing the phrase "federal doubt" and the phrase "federal minority".

Since Google exposes an API (see Google (2003)) for conducting search queries, I developed a system that tries to estimate scores for candidate antecedents based upon Google search results. For each candidate antecedent, the system uses this process to assign the score:

1. For each property of the *one*-anaphor find its semantic head.

   For the purposes of this system, a **property** of a *one*-anaphor is simply any phrase that syntactically depends on the head *one* and isn't a determiner. For example, the properties of *the very large one with red wings* are 'very large' and 'with red wings'.

   The **semantic head** of a property is the head of any embedded noun phrase, or the syntactic head of the property if there is no embedded noun phrase. Property 'very large' has no embedded noun phrase, so the semantic head is the same as the syntactic head: *large*. Property 'with red wings' has an embedded noun phrase *red wings*, so the semantic head is the head of the noun phrase: *wings*.

2. Construct search queries from the permutations of the head of the candidate antecedent, and the semantic heads of the *one*-anaphor's properties. If we were evaluating the candidate antecedent *small snail* for the *one*-anaphor *the very large one with red wings*, the search queries would be "wings large snail", "large wings snail", "wings snail large", "snail wings large", "large snail wings" and "snail large wings". Stop words are filtered out of the queries at this stage, since Google will not filter them out of phrase queries.

3. Find the base number of search results for the head of the candidate antecedent. Using the example above, the base number of search results would be the number of results for the query "snail".

4. Perform a Google search using each query string constructed at step 2. Sum the estimated total results for each of these queries, and divide by the base number of search results to get the score for this candidate antecedent.

Like the previous systems when the scores for the candidate antecedents are computed this system takes the highest scoring candidate as the antecedent. I developed a second system which is extremely similar to this one, except that it also gives a bonus score of 1 to any candidates that are in the same sentence as the *one*-anaphor, and to any candidates which are the subject of the sentence.

Since the number of search queries for any given antecedent is quite high, and Google restricts users to 1000 queries in a single day, I tested this system on the first 50 of the 451 annotated *one*-anaphors with antecedents. The number of antecedents found by this system. and the other systems, on these 50 *one*-anaphors are shown in Table 4.6. The two systems which use Google are called **google** and **google + gram**.

System	# antecedents found	% antecedents found
Baseline 1 (previous NP)	18	36.0
Baseline 2 (previous Subj.)	18	36.0
Heuristic	19	38.0
L&L	21	42.0
Google	12	24.0
Google + gram	21	42.0

Table 4.6: Comparison of the Google-based antecedent finders with the other systems

Table 4.6 shows that semantic information, at least of the sort used by this search, is a fairly poor way to find antecedents, since the **google** test was outperformed by both the base systems. On the other hand, the addition of two very simple tests which model recency and make the subject a more plausible antecedent candidate, the performance of this system is equivalent to the best performing system on this data set.

4.3 Finding the sense of a *one*-anaphor

This section discusses the problem of finding the sense of a *one*-anaphor. This is the problem of selecting which of the antecedent properties apply to the *one*-anaphor. It is the problem of working out that *a tumultuous one* in example 4.6 means a tumultuous year, and of working out that *the new one* in example 4.7 refers to a world, but a world that is neither 'old', nor does it have 'aristocratic values'.

(4.6) **This last year** has been *a tumultuous one* for the Dame. (BNC file ARJ sentence 1661)

(4.7) Amateurism provided a bridge between **the old world of aristocratic values** and *the new one of bourgeois exertion and competitiveness*. (BNC file A6Y sentence 1207)

For a fuller account of the possible relationships between *one*-anaphor and antecedent, see Section A.2 in Appendix A. This chapter does not attempt to discuss finding all the relationships discussed in Section A.2 computationally, but attempts to address a simpler version of this problem: that of identifying contrastive properties.

The main problem considered in this section is that of determining which properties of the antecedent actually contrast with a stated property of the antecedent. In some cases, such as those in example 4.8, none of the properties of the antecedent contrast with properties of the *one*-anaphor. In other cases, such as those in example 4.9, one or more properties of the antecedent do contrast with a property of the *one*-anaphor. In yet other cases, the contrast cannot be inferred from a direct contrast between two properties of the antecedent, but instead must be inferred from by other means, as is the case with the sentences in example 4.10.

(4.8) a. **The third area of UK concern**, and *one which has not yet been resolved*, relates to frontier controls. (BNC file HH2 sentence 1289)

b. Out on to the stones of the terrace there fell **a thick metal spike**, not at all dissimilar to *the blood-smeared one* on which Lord Woodleigh had not fallen... (BNC file A0D sentence 436)

(4.9) a. And the Epilogue also points forward in its closing words to "**a new tale**" because "*our present one* is ended"... (BNC file A18 sentence 47)

b. Because southern prices rose faster than those in the regions in the mid-1980s, for instance, by last year **an average house in Yorkshire and Humberside**, which in 1983 had been worth 69 per cent of *a similar one in London*... (BNC file A5T sentence 149)

(4.10) a. Needless to say, **this new world** always bore an uncanny resemblance to *the one* I had so recently abandoned. (BNC file FR3 sentence 2011)

b. ... she combines, in a radically new way, **two disparate American voices**—*the immediate one of the oral culture, the homespun, the unadorned*, and *the complex, civilised and allusive one of international modernism*. (BNC file A5F sentence 200)

Note that even identifying the explicit contrast between a new tale and the present tale requires understanding that in this context a tale cannot be both a new one and the one under discussion, and identifying that the same houses cannot be both in Yorkshire and London. Interpreting the *one*-anaphors in example 4.10 is even harder, as it requires inferring that a recently abandoned world isn't the same as a new world; and that a single voice cannot be disparate, even though neither 'immediate', or 'of the oral culture' directly contrast with 'disparate'

The harder problem of inferring a sense from context when there is no single antecedent (see the discussion of Example 4.2 in Section 4.1 will not be addressed in this section.

A related problem to that of determining sense shared between *one*-anaphor and antecedent is that of determining shared reference. The reference problem is not discussed further here but is the problem of determining whether the *one*-anaphor and its antecedent refer to the same entity (if both indeed refer to an entity at all). Ideally, it should be possibly to determine that *a sea level* and *the present one* in example 4.11 refer to two distinct sea levels, but this chapter does not discuss the problem of deciding the co-referentiality or otherwise of the *one*-anaphor and its antecedent.

(4.11) The theory thus involves a long phase of stillstand with **a sea level** considerably lower than *the present one* in the latter part of the Tertiary period immediately prior to the Ice Age. (BNC file GV0, sentence 709)

This section has two parts: the first discussing the distribution of *one*-anaphors that have a sense which contrasts with their antecedent among those that do not contrast, and the second discussing some preliminary results of an attempt to use web results to judge the likelihood of two properties contrasting.

4.3.1 Contrast between *one*-anaphor and antecedent

Not all antecedents and *one*-anaphors contrast. Examples 4.12 and 4.13 show two other possible relationships between antecedent and *one*-anaphor:

(4.12) . . . remember that acting is **an art** and *a thrilling one*. . . (BNC file A06 sentence 1571)

(4.13) They can be cut in two separate operations or , if **three alternating veneers** are taped together, *the top one* bearing the pattern, both halves of the design can be cut in one operation. (BNC file A0X sentence 834)

Example 4.12 adds an extra property to the referent of *an art*, the property 'thrilling'. Example 4.13 selects a member of the set referred to by *three alternating veneers*, the member with the property of 'top'

This section discusses three sets of annotation results: the type of antecedents that *one*-anaphors have; the relationships between antecedents and *one*-anaphors, and the combination of type of antecedent and type of relationship.

Type of antecedent

I annotated the 451 *one*-anaphors described in Section 4.1 to determine the number of examples that had the following types of antecedents:

kind where the *one*-anaphor has an antecedent that doesn't refer to anything, but which does have properties inherited by the *one*-anaphor's referent;

set where the *one*-anaphor has an antecedent that refers to a set of entities with a common property inherited by the *one*-anaphor's referent; and

individual where the *one*-anaphor has an antecedent that refers to a single entity with a common head property inherited by the *one*-anaphor's referent.

[C]rime in example 4.14 is a **kind** antecedent, referring neither to a particular crime nor to a set of crimes.

(4.14) The draft letter to Katkov merely claims that **crimes** like *this fictional one* can be found in the newspapers. . . (BNC file A18 sentence 269)

[T]he secular names in example 4.15 is a **set** antecedent, since *the secular names* refers to a set of names. In this case, the additional name is not a member of the set of secular names, but it does inherit the head property 'name' from the antecedent noun phrase.

(4.15) In addition to **the secular names**, there is *an additional one* by which the person is known in the synagogue. . . (BNC file AOP sentence 274)

This theme in example 4.16 is a **individual** antecedent, since *this theme* refers to a single theme individual.

(4.16) **This theme** is *a useful one* for assessing the quality of a critic 's writing. . . (BNC file A04 sentence 632)

Table 4.7 shows the number of occurrences of each type of antecedent among the 451 annotated *one*-anaphor with identifiable antecedents. Individual antecedents are by far the most common, and kind antecedents relatively rare.

Antecedent type	# of occurrences	% of occurences
Kind	26	5.8
Set	77	17.1
Individual	348	77.2

Table 4.7: Antecedent type

Relationship between antecedent and *one*-anaphor

I annotated the 451 *one*-anaphors described in Section 4.1 to determine the number of *one*-anaphors that had the following property sets, as compared to that of their antecedent:

identical properties where the antecedent has a property set identical to that of the *one*-anaphor;

contrasting properties where the antecedent and the *one*-anaphor contrast in at least one property; and

additional properties where the *one*-anaphor has properties additional to that of the antecedent.

The antecedent's properties and the *one*-anaphor's properties are identical in example 4.17: *one* simply refers to "a wee helping hand".

(4.17) I 'd have been glad to give **a wee helping hand** there myself, and I 'm sure *one* was needed. (BNC file AB9 sentence 2516)

[A] full spin and *an incipient one* contrast in example 4.18. The spin that *an incipient one* describes is not both 'full' and 'incipient', it is incipient but not full.

(4.18) However, whether it is **a full spin** or just *an incipient one* is academic if the glider stalls a few hundred feet up. (BNC file A0H sentence 869)

[A]n enabling one in example 4.19 adds a property to the properties of *[t]he Act*.

(4.19) **The Act** is *an enabling one* and does not affect the existing system unless the leadership of the churches wish to co-operate in a reshaping of the system or of schools in any particular area . (BNC file A07 sentence 1403)

Table 4.8 shows the number of occurrences of each antecedent-*one*-anaphor relationship among the 451 annotated *one*-anaphors. *One*-anaphors that add properties are about twice as common as *one*-anaphors that have a property that contrasts with their antecedent, and *one*-anaphors that have the same property set as their antecedent are rare.

Relationship	# of occurrences	% of occurences
Identical properties	6	1.3
Contrasting properties	133	29.5
Additional properties	312	69.2

Table 4.8: Relationship between the properties of antecedent and *one*-anaphor

Type of antecedent and relationship between antecedent and *one*-anaphor

The relationship between *one*-anaphor and antecedent, and the type of the antecedent are orthogonal. However, some combinations of relationship and antecedent type were found very infre-

quently in the 451 annotated examples and some combinations not at all. The number of occurrences of *one*-anaphors that have any given combination of type and relationship with antecedent are shown in Table 4.9.

Type	Relationship	# of occurrences	% of occurrences
Kind	Identical	1	0.2
Kind	Contrasting	0	0.0
Kind	Additional	25	5.5
Set	Identical	1	0.2
Set	Contrasting	12	2.7
Set	Additional	64	14.2
Individual	Identical	4	0.9
Individual	Contrasting	121	26.8
Individual	Additional	223	49.4

Table 4.9: Type of antecedent and relationship between the properties of antecedent and *one*-anaphor

4.3.2 Discovering contrast using Google

In order to derive the fact that the assembly referred to by *the wet one* in example 4.20 is a wet assembly and not an assembly that is both dry and wet, we need a source of semantic knowledge that informs us that something cannot be both dry and wet.

(4.20) **The dry assembly** need not be cramped tight , but *the wet one* obviously must. (BNC file A0X sentence 1374)

One way of determining that the assembly referred to by *the wet one* is not dry would be to model the semantic likelihood (or semantic compatibility) of the two possible interpretations of *the wet one* given the antecedent *the dry assembly*: 'wet' and 'assembly'; and 'dry', 'wet' and 'assembly'. The desired result is that the second of the possibilities gets a significantly lower score than the first.

As in Section 4.2.4, where I used web results to score candidate antecedents by modelling the likelihood that the head of the antecedent was semantically compatible with the properties of the *one*-anaphor, it should be possible to use web results to model the likelihood that any given property of the antecedent is compatible with the properties of the *one*-anaphor.

I designed a system to compute this score. It uses this scoring mechanism for any given property from the antecedent:

1. Find the head of the antecedent. For example, if the antecedent was *a ghastly fat bird*, the head would be *bird*.
2. Find the semantic heads of the properties of the *one*-anaphor. For example, if the *one*-anaphor was *the very large one with red wings*, the semantic heads of the properties are *large* and *wings*, as in Section 4.2.4.
3. construct a set of queries based upon the head of the antecedent, the semantic heads of the properties of the *one*-anaphor and the property being queried. If we were querying the property 'ghastly' of *a ghastly fat bird*, the queries would be all 24 permutations of 'ghastly', 'large', 'wings' and 'bird'.
4. Execute all of the queries, and sum the total number of results returned for any given score.

Using this scoring mechanism for each property of the antecedent, improbable or impossible combinations of properties, like 'dry' and 'wet' should be able to be eliminated on the basis of their low scores.

I developed a system which computes the property scores using the above technique. The complete system works as follows:

1. Compute a baseline score for the *one*-anaphor assuming that none of the antecedent's properties are inherited. In the case of the antecedent being *a ghastly fat bird* and the *one*-anaphor being *the very large one with red wings* the baseline property set is 'large', 'wings' and the head of the antecedent, 'bird'.
2. For each of the antecedent's properties, compute a score using the above mechanism.

I didn't develop a technique for comparing the scores and eliminating contrasting properties using the scores due to time constraints, but there is a small set of examples in Table 4.10.

The results in Table 4.10 for the antecedent-*one*-anaphor pairs that do contrast suggest that this scoring mechanism is inaccurate even as an approximation. Observe, for example, that the combination of '39', 'single' and 'player' scores much more highly than the combination of 'British' 'single' and 'player', even though '39' and 'single' are in contrast. Even in the cases where the desired result—a contrasting property receives a low score—is achieved, it is not clear how to quantify "low score".

4.4 Conclusions

This chapter presented six sets of results: the nearness of *one*-anaphors and their antecedents; the syntactic role of *one*-anaphors and their antecedents; the semantic relationship between *one*-anaphor and antecedents; and the results of three sets of systems: one which found antecedents using grammatical information, one which found antecedents using grammatical and semantic

information, and one which scored properties of an antecedent based on whether they were likely to contrast with the properties of the *one*-anaphor.

The main computational results of this chapter were: several systems that identify candidate antecedents, and a system which evaluates properties of a chosen antecedent based on whether or not they are likely to contrast with properties of the *one*-anaphor.

Antecedent	*One*-anaphor	Contrasting	Antecedent property scores	
			Property	Score
a longing for Gavin's wife	a more urgent one for a teacher at school	for Gavin's wife	baseline	27480
			for Gavin's wife	1152
the theatrical experience	a very concentrated one	None	baseline	1096000
			theatrical	47580
a unified Irish state	a federal one	unified	baseline	8110000
			unified Irish	1000000
the evangelical version	the basic one in the North of Ireland	None	baseline	49300000
			evangelical	138000
the spirit of the laws of this Irish state	a religious one	None	baseline	3800000
			laws	1774000
the concept of the majority	an important one	None	baseline	7430000
			majority	6400000
a contact insecticide	a systemic one	contact	baseline	82200
			contact	53450
the smouldering coals	a red one	None	baseline	89800
			red	5292
the secular names	an additional one	secular	baseline	22560000
			secular	747200
39 British players	a single one	39	baseline	12900000
			British	4006000
			39	6109000

Table 4.10: Examples of using Google results for various antecedent properties

Chapter 5

Conclusions

This chapter discusses the achievements of this project and work that could be built on this project. This project has several significant parts: the results of the corpus analysis, programs that process *one*-anaphors, and theoretical exploration of the problem space. Significant further work in each of these areas is possible.

Many of the significant results from this thesis arise from the corpus analysis: in particular the quantification of the grammatical roles of antecedents, and of the relationship between *one*-anaphor and antecedent. Examination of the data has led to the taxonomies developed in Appendix A.

The experiments with using semantic data suggest that further work on the potential uses of semantic information in *one*-anaphora resolution could yield a good resolution system.

5.1 Identifying *one*-anaphora

The major results of the investigation of identifying *one*-anaphora are threefold: an exploration of the uses of the word *one*, an analysis of the occurrence of these uses in real text, and a program which distinguishes uses of *one*.

Previous work on *one*-anaphora has tended to leave aside the problem of computationally distinguishing uses of *one*, even where the work was computational. This project developed a system which does exactly this, using a combination of grammatical and semantic knowledge.

In addition to developing this system, this project has contributed a unified taxonomy of non-anaphoric uses of *one* based on the work in the literature, and a taxonomy of anaphoric uses of *one*.

5.2 Resolving *one*-anaphora

The major results of the investigation into finding antecedents of *one*-anaphors were threefold: an analysis of the grammatical roles of antecedents, systems that identify antecedents using grammatical roles, and a system that identifies antecedents using a web search for semantic knowledge.

The two most important results showed that grammatical information alone is a fairly poor way of locating the antecedent of a *one*-anaphor, and that combining grammatical information with semantic knowledge about likely antecedents is likely to provide better results if explored further.

Like the problem of locating *one*-anaphors in text, the relationship between *one*-anaphor and antecedent has not been comprehensively explored by existing work. The major results of the investigation into the relationship between antecedent and *one*-anaphor in this project were: quantitative results showing that most antecedents do not contrast with *one*-anaphors, and a system attempting to model likely contrast.

5.3 Future work

This project has raised a number of questions about *one*-anaphors and their relationship with their antecedent. In particular, this project did not even scratch the surface of computationally deriving the relationship between *one*-anaphors and their antecedents, nor of modelling the *one*-anaphor's part in the larger discourse structure. A related, and large, problem, is that of modelling the sense relationships in a discourse in the way that models of the reference relationships in a discourse have been modelled as part of solving the problem of anaphors that co-refer with their antecedent.

This section discusses a number of interesting problems that could be investigated as part of this goal. It is divided into four parts, and discusses in turn further corpus analysis of *one*-anaphora; the problem of distinguishing *one*-anaphors from other uses of *one*; the problem of finding the antecedent of a *one*-anaphor; and the problem of finding the sense of a *one*-anaphor.

5.3.1 Corpus analysis of *one*-anaphora

There are several improvements that could be made on the results of this project by improving or extending the corpus analysis. The most obvious two are analysing a larger or more varied data set, and improved analysis.

Larger annotation

Analysing a larger data set might give sufficient data to uncover patterns in the relationship between *one*-anaphor and antecedent that were not evident after the annotation of 500 examples. A larger data set might also be a suitable training set for machine learning techniques.

The annotations of *one*-anaphors in this project were of *one*-anaphors in a particular context: namely *one*-anaphors preceded by adjectives. There are many contexts other than this in which *one*-anaphors can occur, annotation and analysis of these data sets may yield either confirmation of the results of this project, or demonstrate exceptions to some of its findings.

Some improvements on the existing annotation are possible. In particular, producing the syntactic data by hand might provide more reliable data about the grammatical function of uses of *one* and the antecedents of *one*-anaphors.

Improved extraction of *one*-anaphor from corpora

The techniques used to select the data from the British National Corpus relied on a probability that particular part of speech bigrams contained a large number of *one*-anaphors among other uses of *one* in English. This proved to be correct in the case of the data set chosen here, but a system which extracted *one*-anaphors without relying on subdividing the corpus first.

Ideally, the system developed in Chapter 3 could be developed to the point where it could extract *one*-anaphora from free text.

5.3.2 Finding antecedents of *one*-anaphora

This subsection discusses avenues of future research into finding the antecedents of *one*-anaphora. The single clearest problem emerging from the analysis in Section 4.2 is that of finding a good measure for the likelihood of any noun phrase being the antecedent of a *one*-anaphor. The only clear result emerging from this section was that most *one*-anaphors, like most pronouns and most definite noun phrases, normally have an antecedent within a few sentences and often within the same sentence.

Syntactic antecedent finding

Further work could be done into syntactic features that make a noun phrase likely to be the antecedent of a *one*-anaphor. A possible approach that is more rigorous than my attempts to develop heuristics for finding antecedents would be using machine learning techniques to derive a weighting of different grammatical features of candidate antecedents.

Mining large corpora for sense information

It is more likely that good results could be obtained by using semantic knowledge of some kind. Since the head sense is always inherited from the antecedent, attempting to measure the likelihood that a particular head sense could co-occur with the other sense of the *one*-anaphor is probably a reasonable approach.

There are several avenues of research: specification of the ideal semantic information source for resolving *one*-anaphors; capabilities of existing semantic information sources; and using existing semantic information sources to approximate an ideal data source. A sub-problem of this is using existing web search technologies to model an ideal source of semantic information.

Modelling sense salience

A related avenue of research is that of developing a more general model of **sense salience**, like the models variously called 'centering' and 'focus' in the pronoun resolution literature, which attempt to model the relevance of particular entities to the discourse. A sense salience model would attempt to quantify the senses which are most relevant to the discourse.

Presumably, a noun phrase with a more salient sense is more likely to the antecedent of a *one*-anaphor. Also, an antecedent with other salient senses may also be more likely to be the antecedent of a *one*-anaphor.[1]

An additional problem, not addressed at all in this project, is the significant minority of *one*-anaphors that have no single antecedent. There are two related problems here: locating the multiple noun phrases from which the head sense of the *one*-anaphor must be inferred, and determining a head property from them.

5.3.3 Resolving *one*-anaphora

This project barely scratched the surface of the relationship between an antecedent and a *one*-anaphor. Further work includes developing a better model of the sense of antecedent and *one*-anaphor, a fuller analysis of the types of *one*-anaphor, the specification of a semantic data source that would aid the resolution of *one*-anaphora, and the development of such a data source.

Sense representation

One of the first open problems in determining the full sense of a *one*-anaphor is developing a more useful or complete representation of that sense. The representation in the body of the thesis simply considers every syntactic dependency of the head noun of the antecedent a property, and distinguishes them using a concept of 'semantic head'.

A more sophisticated representation, perhaps like the representation in Webber (1979), would aid both a better corpus analysis and development of better resolution techniques by allowing a much richer annotation and much fuller analysis of antecedents and *one*-anaphors. Development

[1] It is difficult to tell, however, what **sense salience** might mean outside the context of resolving *one*-anaphora and related phenomena like anaphoric use of the word *other*, so this argument may be nothing more than a second call for a better measure of antecedent likelihood.

of better resolution techniques would require the development of techniques for computing the representation given the text.

Discourse representation

While this project did investigate the contrast between properties of antecedents and properties of *one*-anaphors, it did not investigate the use of this contrast in the discourse. In other words, it did not answer the question "what are *one*-anaphors used for?"

There are two possible approaches to this problem: analysis of discourse functions in text based on examples of *one*-anaphors in a corpus, and development of computational approaches to understanding these functions and modelling them given text.

5.3.4 Summary of future work

There are two avenues of further research that this project suggests. The first is fuller data analysis, using larger or more varied data sets. The second is better theorectical modelling of *one*-anaphors and of their role in discourse. More data analysis would extend the results of this project by discovering whether its results are applicable to *one*-anaphors in more contexts, and by uncovering existing relationships between *one*-anaphors and their antecedents and *one*-anaphors and their discourse context that this project did not explore. A better model of *one*-anaphora would allow the development of generalisable computational approaches to *one*-anaphora resolution—perhaps to the point of being generalisable to some other problems in anaphora resolution and natural language understanding.

Bibliography

British National Corpus (1995), 'What is the BNC?'. Available from `http://www.natcorp.ox.ac.uk/what/`.

Connexor (2003), 'Machinese syntax'. Available from `http://www.connexor.com/m_syntax.html`.

Dahl, D. A. (1985), The structure and function of *one*-anaphora in English, PhD thesis, University of Minnesota.

Ferrández, A., Palomar, M. & Moreno, L. (1997), Slot unification grammar and anaphora resolution, *in* 'Recent Advances in Natural Language Processing (RANLP'97)', Bulgaria.

Ferrández, A., Palomar, M. & Moreno, L. (1998), A computational approach to pronominal anaphora, one-anaphora and surface-count anaphora, *in* 'Proceedings of the Discourse Anaphora and Resolution Colloquium', pp. 117–128.

Google (2003), 'Google web apis'. Available from `http://www.google.com/apis/`.

Halliday, M. A. K. & Hasan, R. (1976), *Cohesion in English*, Longman, London and New York.

Kennedy, C. & Boguraev, B. (1996), Anaphora for everyone: pronominal anaphora resolution without a parser, *in* 'Proceedings of the 16th International Conference on Computational Linguistics (COLING'96)', Copenhagen, Denmark, pp. 113–118.

Lappin, S. & Leass, H. (1994), 'A syntactically based algorithm for pronominal anaphora resolution', *Computational Linguistics* **20**, 535–561.

Leech, G., Garside, R. & Bryant, M. (1994), CLAWS4: The tagging of the British National Corpus, *in* 'Proceedings of the 15th International Conference on Computational Linguistics (COLING 94)', Kyoto, Japan, pp. 622–628.

Luperfoy, S. (1991), Discourse Pegs: A Computational Analysis of Context-Dependent Referring Expressions, PhD thesis, University of Texas at Austin.

Vieria, R. & Poesio, M. (2000), Processing definite descriptions in corpora, *in* S. Botley & T. McEnery, eds, 'Corpus-based and Computational Approaches to Anaphora', UCL Press, pp. 189–212.

Webber, B. (1979), *A Formal Approach to Discourse Anaphora*, Garland Publishing Inc., New York & London.

Appendix A

Uses of the word *one*: Taxonomy

Section 2.1 reviewed other authors' descriptions of the various uses of *one* and reviewed criteria other authors have used to distinguish *one*-anaphora from other use of *one*. This chapter uses these descriptions to develop a more detailed and complete account of non-anaphoric uses of *one*. In addition, it gives a detailed account of the semantic function of *one*-anaphora.

The purpose of this account is ultimately to divide *one*-anaphors from non-anaphoric uses of *one*. This appendix is divided into two sections: the first of which discusses the various non-anaphoric uses of the word *one*, the second of which discusses in detail the function of *one*-anaphora.

A.1 Non-anaphoric uses of *one*

This section describes uses of *one* that are not *one*-anaphoric. There are four distinct non-anaphoric uses of *one*: the numeric, generic, objective and partitive uses, together with a couple of idiomatic uses. This section uses the following properties of each use of *one* to differentiate them:

word class the part of speech of this use of *one*;

syntactic properties the syntactic relationship between a use of *one* and the rest of the sentence;

semantic properties the semantic relationship between a use of *one* and the properties of whatever it is describing;

contrasting words other words that have the syntactic properties of this use of *one* and which could replace this use of *one* in otherwise identical sentences;

plural form the plural form of *one* that has the same use; and

referentiality whether or not the use of *one* is used by the discourse to refer to an object in the world.

A.1.1 Numeric *one*

A numeric *one* indicates that exactly one of an object referred to by the noun phrase in which the numeral *one* occurs. The numeric use of *one* indicates only a single property—the type of the referent is given by the head noun.

Example A.1 shows a numeric use of *one*:

(A.1) There is *one* ball.

The function of *one* in the noun phrase *one ball* is to indicate that there is exactly one ball. However, the type of the noun phrase's referent is given by the head noun, *ball*.

word class adjective/numeral.

syntactic properties modifies count nouns at the head of noun phrases.

semantic properties indicates that one (1) entity of the type given by the head noun is being referred to.

contrasts with singular determiners: *the*, *a*, *this* etc.

referential the word itself is not referential, as it isn't a noun, but it may occur in referential noun phrases.

plurals other cardinals (*two*, *three* etc.) and the indefinite equivalents (*some*, *many* and *several*) have this function in plural noun phrases.

Examples from the literature:

Author	Example	Author's category
Halliday & Hasan	He made *one* very good point.	Cardinal numeral
Halliday & Hasan	The *one* friend who never let her down was Enid.	Cardinal numeral
Halliday & Hasan	You've already got *one* red one.	Cardinal numeral
Halliday & Hasan	I'd like *some* coffee.—Then make some.	Indefinite article
Luperfoy	It took *one* man to lift the piano.	Determiner
Luperfoy	It took { *four* / *several* / *many* } men to lift the piano.	Determiner
Dahl	I have *one* dog.	Cardinal number

A.1.2 Generic *one*

The generic use of *one* is a pronominal use that need not refer to an individual. Speakers or authors use this type of *one* to state something about herself, a particular person or about people in general

without having to use a proper name or a first, second or third person pronoun. In some cases, *one* may refer to an idealised person, but it may also refer to an individual in a context where the discourse does not allow a conventional reference to be made—for example, it is sometimes used by the author of a work to refer to herself.

Example A.2 illustrates a generic use of *one*:

(A.2) *One* doesn't listen to the people.

word class noun/pronoun.

syntactic properties head of noun phrase.

semantic properties indicates a person, possibly a generic person.

referential sometimes.

contrasts with singular pronouns (*I*, *you*, *he*, *she*), other references to individuals, for example, by proper name.

plural *we* or *you*.

Examples from the literature:

Author	Example	Author's category
Halliday & Hasan	They couldn't do a thing like that to *one*.	Generalised person
Halliday & Hasan	*One* never knows, does *one*?	Generalised person
Halliday & Hasan	It makes *one* think, though.	Generalised person
Halliday & Hasan	*One* can hardly be expected to reveal *one's* innermost secrets to the first casual inquirer, can *one*?	Generalised person
Halliday & Hasan	*One* can hardly be expected to reveal *his* innermost secrets to the first casual inquirer, can *he*?	Generalised person
Dahl	*One* never knows what will happen.	Impersonal

A.1.3 Object *one*

The objective category is motivated by Luperfoy's examples listed in Section 2.1.3, in which *one* and *one dollar bill* are synonymous, and *one* refers to a one dollar note, and does not depend on any previous reference in order to be understood. This use of *one* is rare.

word class noun.

syntactic properties head of noun phrase.

semantic properties not anaphoric; where an anaphoric *one* stands in for a head property, objective *one* refers to an entity of type one—one *is* the head property.

referential yes

contrasts with *two*, *three*..., which might be, for example, used to refer to a two dollar bill or a three dollar bill

plural *ones*

Observe that the plural of *a two*—referring to a two dollar bill—would be *twos*, as in example A.3:

(A.3) Can you give me ten dollars in *twos*?

This plural is not used for the *one*-anaphoric use of *two*. Unfortunately, the object use of *one* does not have a similarly unique plural: *ones* is the plural of some anaphoric uses of *one*, as well as being the plural of the objective *one*.

Examples from the literature:

Author	Example	Author's category
Luperfoy	I'm getting low on small bills. I only have one *one*.	Numeral
Luperfoy	I can trade two fives and ten *ones* for your twenty.	Numeral
Halliday & Hasan	If such a *one* be fit to govern, speak.	"Pro-noun"
Halliday & Hasan	Now, my dearest *ones*; gather round.	"Pro-noun"

A.1.4 Partitive *one*

This use of *one* selects an individual from a set whose members have a certain property. In example A.4, an individual cat is selected, in A.5 an individual apple.

(A.4) *One of the cats* scratched me.

(A.5) The apple picking went well, but *one of them* was rotten.

This usage of *one* is very similar to an anaphoric use of *one* in which the *one*-anaphor selects an individual from a set, as in examples A.6 and A.7, and in fact *one of them* in example A.5 is an anaphoric noun phrase, but the anaphor is the head of the noun phrase—the pronoun *them*.

(A.6) There were **some cats** and *one* scratched me.

(A.7) The apple picking went well, but *one* was rotten.

word class noun.

syntactic properties head of noun phrase, modified by a prepositional phrase beginning with *of*.

semantic properties not anaphoric; refers to an entity of the type specified in the dependant propositional phrase.

referential yes

plurals *two*, *three*..., *some*, *several*, *many*

Observe that the plurals of partitive *one* are like those of the numeric *one*. The partitive use of *one* is something like a numeric *one* elevated to the head of the noun phrase, but observe the difference between *cat* in example A.8 and *the cats* in example A.9:

(A.8) I have *one* cat.

(A.9) I have *one* of the cats.

The cats in example A.9 is referential—it refers to a particular set of cats from which a particular cat is selected. *One cat* in example A.8 simply selects a single individual that has the property of being a cat—no larger set of cats is referred to. This distinction is similar to distinctions made later between different forms of *one*-anaphora.

A.1.5 Idiomatic uses of *one*

The preceding taxonomy does not account for certain idiomatic uses of *one*, of which there are several. One idiomatic use indicates that separate things are united or the same, as illustrated by example A.10. Another, more common, idiom is to refer to a person as *a loved one* or *a little one* as in example A.11. A third idiom uses *for one* to indicate a person who is an exception to a rule, as in example A.12. A fourth idiom is that of using *one* as part of the proper name of an entity as in example A.13: *round one*, *Air Force One*.

(A.10) a. The king and the land are *one*.

b. We are *one*, but we are many.

(A.11) a. Taking me little *one* down to the country to see her Nan, she heard Gloria say, then the click-clack of the ticket-clipper before the man closed the door and they were alone again. (BNC file AC5 sentence 799)

b. The Evil *One* saw his chance and robbed them of their destiny. (BNC file ABV sentence 1470)

c. The Transport Minister, Mr Robert Atkins, said: Christmas is a family time of year, but for some families Christmas is going to be unbearable, because a loved *one* has been killed by a drink-driver. (BNC file A8X sentence 949)

(A.12) Lord Home for *one* disapproved, recalling the only difference I ever had with Mr Macmillan. (BNC file B0H sentence 485)

(A.13) UK hotel companies are anxiously awaiting the outcome of round *one* in the bidding for the former East German hotel group Interhotel. (BNC file A7C sentence 889)

A.2 Anaphoric uses of *one*

This section discusses the semantic function of *one*-anaphora. It describes *one*-anaphora in terms of a central function, which is relating a set of properties to the set of properties mentioned by the antecedent.

The relationship between a *one*-anaphor and its antecedent is considerably more complex than the relationship between, for example, the pronoun and its antecedent in example A.14:

(A.14) **The king** directs the generals and *he* brooks no disobedience.

The king and *he* are co-referential—they refer to the same entity in the world described by example A.14. The relationship between *one*-anaphor and their antecedents differ. The relationship is not always co-reference, indeed, in some cases the antecedent or the *one*-anaphor do not refer to anything in the world described by the discourse.

This section discusses three types of *one*-anaphors: those whose antecedent is a kind (having no reference, only sense); those whose antecedent is a set of entities; and those whose antecedent refers to a single instance.

A.2.1 Antecedent is a type or class

In this variety of *one*-anaphora, the antecedent of the *one*-anaphor is not referential, and refers to a type or class of object, rather than to any object or set of objects. Consider the antecedent *green cars* in example A.15, which does not refer to any particular group of green cars.

(A.15) Do you like **green cars**? Yes I own *one*. It has 70 000km on the clock.

(A.16) For **suppliers and contractors of equal standing**, *the one giving greater consideration to environmental issues* will be preferred. (BNC file HB8 sentence 95)

(A.17) I like **dogs** and I have *one*.

(A.18) We do not distinguish experiences from non-experiences as we might distinguish **oranges** from **apples**, viz. by indicating certain characteristics that might enable anyone, including those who have never tasted either **fruit**, to tell *one* from the other. (BNC file FTV sentence 1392)

Example A.19 shows that the *one*-anaphoric noun phrase itself may add properties to the set of properties it copied from the antecedent:

(A.19) Do you like **green cars**? Yes I own *one with 70 000km on the clock*.

Example A.20 shows that the properties may not all be inherited:

(A.20) I despise **green cars** but I like *my red one*.

A.2.2 Antecedent is a set

In example A.21, a set of two factors is referred to, and then each of the items in the set is referred to, each distinguished by a property.

(A.21) **Two factors**, *one personal*, *one educational*, were to conspire to undermine the credibility of the educational solution in Nizan's mind. (BNC file FTW, sentence 230)

In example A.22, a set of two teachers is referred to, and every member of the set is assigned a property. Each member of the set is referred to, and each is assigned an action.

(A.22) I was enrolled at once in the village school which had **two teachers who were sisters**, *one* taking all the younger classes and the other the older children. (BNC file CDC sentence 20)

(A.23) If, unusually, there is no time pressure, **the targets** can be approached *one* at a time in order of their relative attractiveness. (BNC file HJ5 sentence 2498)

(A.24) If, unusually, there is no time pressure, **the targets** can be approached *one by one* in order of their relative attractiveness.

In example A.25, a set of four purchasers is introduced, and two items are selected from it by virtue of properties that will be assigned to them in the future.

(A.25) With **the four prospective purchasers** which we have at this stage it will be necessary to go through an elimination process, selecting *one* to proceed with and perhaps *one* to keep on hold. (BNC file HJ5 sentence 6107) The A six six five eight if you look at the the bottom of **the two roads**, it's *the one on the left which is going on down off the plan.* (BNC file J9U sentence 790)

In example A.26 a set whose elements have a common property is referred to. A subset (*the first two*) of the videos having an additional property (that of being very successful) is then referred to. Finally, one of the subset, having a number of additional properties, is referred to.

(A.26) We've launched the series of **six Beatrix Potter videos**, the first two have been very successful, *the first one* sold three hundred and fifty thousand copies and is er I think it's number three in the best seller list and we'll give you one of these to take home and er and, and watch to your leisure. (BNC file HUP, sentence 52)

Finally, the *one*-anaphor may not refer to a member of the set, it may contrast with the members of the set:

(A.27) **The two elder children were badly behaved**, I preferred *the youngest one*.

A.2.3 Antecedent is an instance

When the antecedent is an instance, the *one*-anaphor may have identical properties:

(A.28) Would you like *this biscuit*? Yes, I would like *that one*.

(A.29) We will replace your car, with **a new car of the same make and model**, if *one* is available. (BNC file HB5 sentence 791)

Alternatively, the *one*-anaphor may have properties additional to those of its antecedent:

(A.30) And in a sense, the book is very one dimensional in that it follows just **this one aspect of Woodrow Wilson 's character**, *the critical one*. (BNC file HUN sentence 556)

(A.31) **He**, *the one she had assumed to be Forest*, was descending carefully, stealthily, carrying some heavy object. (BNC file CCD sentence 2551)

(A.32) In brief, **a District-wide CMHT** has been established, *one which was originally based on a team of hospital social workers and nurses supervised by a consultant psychiatrist*. (BNC file FTY sentence 821)

(A.33) I mean **that outfit**'s not *a bad one* for her really is it? (BNC file KPU sentence 1750)

Finally, it may contrast with its antecedent, only inheriting a subset of its properties:

(A.34) Needless to say, **this new world** always bore an uncanny resemblance to *the one* I had so recently abandoned. (BNC file FR3 sentence 2011)

(A.35) The theory thus involves a long phase of stillstand with **a sea level** considerably lower than *the present one* in the latter part of the Tertiary period immediately prior to the Ice Age. (BNC file GV0 sentence 709)

(A.36) Then you would need two of the, **two of these sheets** if you were doing two conversations, but only *one* if you 've done one . (BNC file KPV sentence 295)

(A.37) They won **all stages** but *one* in a seven month tournament making them overall leaders by 135 points. (BNC file HBE sentence 398)

A.3 Conclusions

This chapter has described a taxonomy of non-anaphoric uses of *one* and of *one*-anaphors.

The taxonomy of non-anaphoric uses of *one* is quite close to that described in Section 2.1, and contains categories based on those described there. The taxonomy of anaphoric uses of *one* is quite different from that described in Section 2.2, being based around the properties of the antecedent, rather than the functions of the *one*-anaphor, but is the same as the "type of antecedent" property discussed in Section 4.3.1.

Appendix B

Uses of *one* in English: Empirical evidence

This chapter discusses occurences of *one* in the British National Corpus. The word *one* is used in the BNC over one quarter of a million times, and this investigation aimed to use available information to reduce this data set to a set which would be a more managable size. Since the investigations in chapters 3 and 4 required annotating the data manually, a data set of "manageable size" is one which does not take a prohibitive amount of time to annotate by hand and is about five hundred or one thousand sentences in size.

The purpose of investigating the occurences of *one* a corpus of English is twofold. The first purpose is to inform the taxonomy described in Appendix A. The second is to discover good sources of data with which to develop and test the anaphora resolution tool which is the goal of my project.

The word *one* occurs in the BNC 258 317 times, 0.26% of the total number of words. It occurs in 232 355 sentences, or in 3.73% of all sentences in the BNC. The following sections discuss divisions of this set. The first section discusses the usefulness of using the assigned part of speech tags to divide the occurences of *one* into groups. The second section discusses using the part of speech assigned to the use of *one* and an adjacent word to divide the occurences of *one* into smaller groups.

B.1 Evidence From Part of Speech Tags

Table B.1 illustrates the complete list of part of speech tags assigned to *one* in the BNC. 12 distinct part of speech tags were assigned in total, but only 5 of those comprise more than 10 occurences each. Both Table B.1 and Figure B.1 illustrate that the only five significant part of speech tags assigned to *one* are:

1. cardinal numeral

2. indefinite pronoun
3. indefinite pronoun + cardinal numeral
4. cardinal numeral + indefinite pronoun
5. reflexive pronoun

Examples of each of these part of speech tags follow:

cardinal numeral However, clinically significant improvements were often limited to just *one* spouse... (BNC file ALN sentence 896)

indefinite pronoun He may just as well be a decision-maker, *one* who can foresee what decisions he will have to make ... (BNC file ASY sentence 570)

indefinite pronoun + cardinal numeral With *one* bound, Jack was free . (BNC file B7C sentence 1447)

cardinal numeral + indefinite pronoun Its country folk are very much at *one* with the land. (BNC file C93 sentence 1413)

reflexive pronoun ... they are likely to approach *one* another (BNC file EW8 sentence 350)

Word class	Number of occurences	Percentage of total
cardinal numeral (excl ONE)	181339	70.20%
indefinite pronoun	54744	21.19%
indefinite pronoun+cardinal numeral (excl ONE)	15999	6.19%
cardinal numeral (excl ONE)+indefinite pronoun	3553	1.38%
reflexive pronoun	2659	1.03%
”unclassified” items which are not words of the English lexicon	7	0.00%
singular noun+base form of lexical verb (except the infinitive)	5	0.00%
base form of lexical verb (except the infinitive)+singular noun	3	0.00%
base form of lexical verb (except the infinitive)	2	0.00%
singular noun	2	0.00%
infinitive of lexical verb	2	0.00%
singular noun+adjective (unmarked)	1	0.00%

Table B.1: Part of speech annotations for *one* in the BNC

Examples of each of these parts of speech show that the parts of speech do not line up neatly with the categories described in Appendix A.

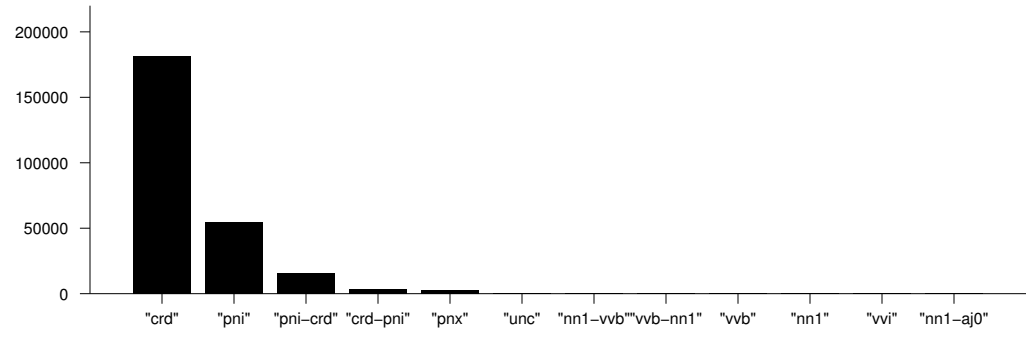


Figure B.1: Graph illustrating the part of speech label of *one* in the BNC

It follows from the fact that there are far fewer part of speech tags assigned to *one* than there are uses of *one* described in Appendx A that different uses of *one* must, at least sometimes have the same part of speech tag. Examples B.1 and B.2 illustrate one case where this occurs. Example B.1 is sentence 612, and example B.2 is sentence 931, in BNC file `ALU`.

(B.1) Their Phyllanthus epiphyllanthus from the West Indies had been lost but he had seen *one* in the physic garden at Amsterdam where, with proper management, it was in great vigour.

(B.2) The almost incessant labour which that art requires leaves so little time for study that *one* can hardly find any person of sufficient experience capable of writing.

At the same time, different tags are sometimes used in the BNC to mark extremely similar uses of *one*. A clear inconsistency in the tagging is illustrated by examples B.3 and B.4, which are sentences 139 and 145 in BNC file `A00` respectively.

(B.3) Those involved with ACET are now helping provide home care for *one* in four of all those dying with AIDS in the UK and up to 4,000 school pupils a month are now receiving education on the subject.

(B.4) Official figures suggest that ACET provided care at home for up to *one* in four of all those who died of AIDS in the UK last year.

one in example B.3 is tagged as **crd**, whereas *one* in example B.4 is tagged as **prn** despite the two occurences appearing in near identical contexts, both lexically and syntactically.

B.2 Evidence from Part of Speech Bigrams

Given that the part of speech tags themselves are both insufficiently specific and inaccurate, I have used bigrams and trigrams to divide the first categorisation based on part of speech further. The first such subdivision was to categorise occurences of *one* both by the part of speech tag on the use of *one*, and the part of speech tag on the preceding and succeeding word. This produces two sets of bigrams: one for *one* and preceding words and one for *one* and succeeding words.

There are 360 sets of succeeding bigrams, and the top two most frequent sets (cardinal *one* followed by a singular noun, and cardinal *one* followed by the preposition *of*) account for more than half—52.59%—of all occurences of *one* in the BNC. Table B.3 and Figure B.3 illustrate the distribution of occurences among the top ten sets of succeeding bigrams. Of the 360 sets of bigrams, 108 contain more than 100 occurences of *one*; 50 contain more than 500 occurences of *one*; and 27 contain more than 1 000 occurences of *one*.

There are slightly more—375—sets of preceding bigrams, and the top two most frequent sets account for far fewer occurences than the top two preceding bigrams sets do. Only 24.64% of occurences fall into the top two preceding bigram sets (the preposition *of* followed by a cardinal number, and sentence initial occurences of *one*). Table B.2 and Figure B.2 illustrate the distribution of occurences among the top ten sets of preceding bigrams. Of the 375 sets of bigrams, 137 contain more than 100 occurences of *one*; 66 contain more than 500 occurences of *one* and 43 contain more than 1 000 occurences of *one*

Word class	No. of occurences	% of total
preposition (except for OF), cardinal numeral (excl ONE)	40162	15.55%
Sentence initial, cardinal numeral (excl ONE)	23493	9.09%
adverb (unmarked), cardinal numeral (excl ONE)	18077	7.00%
-s form of the verb ”BE”, cardinal numeral (excl ONE)	9251	3.58%
adjective (unmarked), indefinite pronoun	8840	3.42%
singular noun, cardinal numeral (excl ONE)	8542	3.31%
general determiner, indefinite pronoun	6649	2.57%
article, indefinite pronoun	6624	2.56%
coordinating conjunction, cardinal numeral (excl ONE)	6568	2.54%
the preposition OF, cardinal numeral (excl ONE)	6012	2.33%

Table B.2: The 10 most frequent POS bigrams for *one* and a preceding word

The information from the part of speech bigram sets does not provide enough information to decide which categories of words are interesting (in the sense of “likely to contain *one*-anaphora” or even “certainly does not contain *one*-anaphora”). It does, however, divide the data into sets which are both large enough to provide a potentially interesting set of examples and small enough

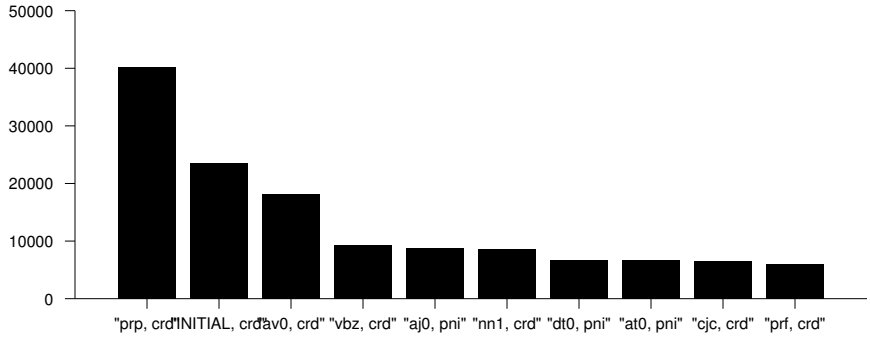


Figure B.2: Graph of the 10 most frequent POS bigrams for *one* and a preceding word

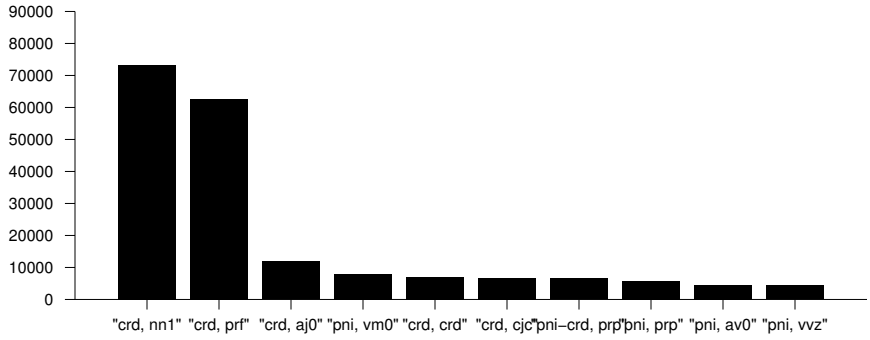


Figure B.3: Graph of the 10 most frequent POS bigrams for *one* and a preceding word

(excepting the first two or three highest frequency sets of bigrams) to be able to be annotated in a reasonable amount of time.

In order to distinguish between potentially interesting, and probably uninteresting, categories, I discuss further analyses of this data in the sections following.

B.3 Evidence From Lexical Bigrams and Trigrams

I computed lexical bigrams subsets of each part of speech bigram set. The lexical bigrams represent the frequency of distinct word pairs marked with the same two parts of speech.

For example, recall that the most frequent pair of parts of speech bigram for *one* is a preceding word are a preposition followed by the word *one* marked as a cardinal numeral. The top five word pairs occuring in this set are:

Word class	No. of occurences	% of total
cardinal numeral (excl ONE), singular noun	73249	28.36%
cardinal numeral (excl ONE), the preposition OF	62602	24.23%
cardinal numeral (excl ONE), adjective (unmarked)	11847	4.59%
indefinite pronoun, modal auxiliary verb	8034	3.11%
cardinal numeral (excl ONE), cardinal numeral (excl ONE)	7072	2.74%
cardinal numeral (excl ONE), coordinating conjunction	6786	2.63%
indefinite pronoun+cardinal numeral (excl ONE), preposition (except for OF)	6535	2.53%
indefinite pronoun, preposition (except for OF)	5866	2.27%
indefinite pronoun, adverb (unmarked)	4580	1.77%
indefinite pronoun, -s form of lexical verb	4356	1.69%

Table B.3: The 10 most frequent POS bigrams for *one* and a succeeding word

1. *in one*, which occurs 7 630 times with these part of speech tags
2. *at one*, which occurs 4 981 times with these part of speech tags
3. *to one*, which occurs 4 057 times with these part of speech tags
4. *from one*, which occurs 3 482 times with these part of speech tags
5. *on one*, which occurs 3 394 times with these part of speech tags

Simply by providing examples of common contexts in which *one* occurs, lexical bigrams point to interesting data sets. The examples above show that bigrams are often insufficient for demonstrating the use of *one*—the context is not obvious from the bigram.

Nevertheless, it is possible to select some data sets from the part of speech sets using lexical bigrams.

From the part of speech bigram data sets, concentrating on sets with more than 1 000 occurences, I chose certain sets as potentially rich sources of *one*-anaphora. Potentially interesting subsets of the preceding part of speech bigram data are listed in Table B.4. Potentially interesting sets of data from the succeeding part of speech bigram sets are listed in Table B.5. It is important to remember that the preceding and succeeding bigrams are not discrete: an occurence of *one* that occurs in one of my chosen preceding bigram sets may also occur in a chosen succeeding bigram set.

It is also possible to make some preliminary judgements about which data sets are unlikely, in the main, to be rich in *one*-anaphora. For example, the most frequent succeding bigram set—cardinal numeral *one* followed by a singular noun—has the following five most frequent lexical bigrams:

Size of set	Parts of speech	Five most frequent word pairs
6 649	general determiner + indefinite pronoun *one*	*that one*, *this one*, *each one*, *another one*, , *one*.
6 624	article + indefinite pronoun *one*	*the one*, *every one*, *no one*, *a one*, , *one*
1 585	ordinal + indefinite pronoun *one*	*first one*, *last one*, *next one*, *second one*, *third one*
1 365	wh-determiner + indefinite pronoun *one*	*which one*, *last one*, *next one*, *second one*, *third one*

Table B.4: Potential sources of *one*-anaphora from the preceding part of speech bigrams based on frequent lexical bigrams

Size of set	Parts of speech	Five most frequent word pairs
8 034	indefinite pronoun *one* + modal auxiliary verb	*one can*, *one would*, *one might*, *one could*, *one must*
2 233	indefinite pronoun *one* + wh-determiner	*one which*, *one whose*, *one what*, *one whatever*, *one whichever*
2 153	indefinite pronoun *one* + wh-pronoun	*one who*, *one whom*

Table B.5: Potential sources of *one*-anaphora from the suceeding part of speech bigrams based on frequent lexical bigrams

1. *one day*, which occurs 4 197 times with these part of speech tags
2. *one thing*, which occurs 3 388 times with these part of speech tags
3. *one hand*, which occures 2 534 times with these part of speech tags
4. *one side*, which occures 2 501 times with these part of speech tags
5. *one way*, which occurs 2 001 times with these part of speech tags.

These five lexical bigrams strongly suggest that many of the occurences of *one* in this set are uninteresting—that is, they are not head of a noun phrase and therefore not *one*-anaphoric.

Inspecting the frequency of lexical trigrams (triplets of words) within the part of speech bigrams yield further "interesting" categories. These are presented in Tables B.6 and B.7

B.4 Conclusion

Part of speech bigrams appear to be a good first-pass way of extracting a managable set of *one* occurences from the 258 317 occurences in the BNC. Associated lexical bigrams and trigrams allow "interesting" sets—in the sense of "these words could be part of a *one*-anaphora"—to be extracted from the hundreds of part of speech bigrams.

Size of set	Parts of speech	Five most frequent word triplets
8 840	adjective (unmarked) + indefinite pronoun *one*	*the only one*, *the other one*, *a new one*, *a good one*, *the right one*
5 738	past form of the verb "BE" + cardinal numeral *one*	*it was one*, *there was one*, *he was one*, *, was one*, *that was one*
2 263	past participle form of lex. verb + cardinal numeral *one*	*'ve got one*, *has become one*, *only got one*, *'s got one*, *n't got one*

Table B.6: Potential sources of *one*-anaphora from the preceding part of speech bigrams based on frequent lexical trigrams

Size of set	Parts of speech	Five most frequent word pairs
4 580	indefinite pronoun *one* + adverb (unmarked)	*one there .*, *one here .*, *one too .*, *one there ,*, *one then ?*
2 913	indefinite pronoun *one* + -s form of the verb "BE"	*one is the*, *one is that*, *one is a*, *one is to*, *one is not*
2 836	indefinite pronoun *one* + the conjunction THAT	*one that 's*, *one that i*, *one that has*, *one that you*, *one that was*
1 381	indefinite pronoun *one* + infinitive marker TO	*one to be*, *one to have*, *one to go*, *one to do*, *one to make*

Table B.7: Potential sources of *one*-anaphora from the suceeding part of speech bigrams based on frequent lexical trigrams

Appendix C

Code documentation

This chapter discusses the code developed in the course of this project, and also discusses the data files produced by this project.

The code is available from `http://www.ics.mq.edu.au/~gardiner/project/gardiner-code.zip`. Check the README file in the zip file for instructions on running any of the code.

Prerequisites for running this code are:

- a version of Python greater than or equal to 2.2;
- an installed version of the Twisted networking framework version 1.0.6 or 1.0.7[1];
- an installed version of pyGoogle[2]
- access to the Connexor parser over TCP/IP (Connexor is accessible on port 5000 of `pompeii.ics.mq.edu.au`); and
- access to the British National Corpus's text files. Much of the annotation data assumes that the BNC files are located in `/home/mary/bnc/Texts`, although it would obviously be possible to change this with search and replace.

In order to run the code from the project, the base directory of the code will need to be in your `PYTHONPATH` environment variable. Otherwise Python will fail with import errors.

C.1 Corpus analysis tools

Before doing any annotation, the code in the `Scripts` directory needs to run. The order the scripts need to run in is:

[1] Available from `http://twisted.sf.net/`, documentation at `http://twistedmatrix.com/`
[2] Available from `http://diveintomark.org/projects/pygoogle/`

bnc-extract.py Extracts all sentences containing *one* in the BNC. Run as `python bnc-extract.py <files>` where `<files>` is a list of all files in the BNC. The results will be printed to standard output, redirect this

bnc-data.py Extracts all the part of speech bigrams from the BNC. Run as `python bnc-data.py <directory> <file>` where `<directory>` is the output directory for the three output files, and `<file>` is the file created in the previous steps

C.2 Corpus annotation tools

There are three sets of corpus annotation tools, which present uses of *one* and their context for hand annotation. These are contained in the `Annotation` directory.

There are three subdirectories: `Antecedent`, containing the antecedent annotating code; `Relationship`, containing the antecedent-*one*-anaphor relationship annotation code; and `Type` containing the use of *one* annotation code.

Each directory contains two files: `annotation.py` which you run when you want to annotate data, and `reporter.py` which you run when you want the results of the annotation summarised.

C.3 Anaphora resolution system

The `oneanaphora` package contains the anaphora resolution system.

There are four subpackages:

resources containing general resources used by the entire project: `bncparser` which processed BNC SGML files; `general` which contains non-project specific resources like a routine to produce all the permutations of a list and `conexor` which contains scripts to parse the output of the Connexor parser, and convert it into a set of Python objects (`ConexorWord` objects) that represent each word of the sentence and their grammatical relationships;

system which contains the three scripts that run various parts of the system: `antecedent.py` (the antecedent finder), `identification.py` (the *one*-anaphor finder), and `properties.py`, the property scorer;

identification which contains the code used to identify uses of *one* – see `categorisation.py` in particular;

resolution which contains the resolution code, see `resolver.py` in particular, which contains classes representing noun phrases, *one*-anaphors, and candidate antecedents.

C.3.1 Hand annotation results

The hand annotation results are located in the `Results/Annotation` subdirectory.

There are several files:

annotation-aj0-pni.txt, annotation-cjc-crd.txt and annotation-cjc-pni.txt contain annotations of over 700 uses of *one* in the BNC, marking which use of *one* they are. The data fields are: filename, sentence number, word number, use of *one*, comma separated. The numeric codes for the uses of *one* are defined in the `oneanaphora.identification.classification` module

annotation-aj0-pni-sentences.txt, annotation-cjc-crd-sentences.txt and annotation-cjc-pni-sentences.txt give the full text of sentences that each use of *one* in the aforementioned set of 700 occurs in—this saves having to parse the BNC SGML files to process them. The words in these files are separated by four pipe symbols (||||)

annotation-aj0-pni-antecedent.txt contains annotations of the antecedents of 495 examples of *one*-anaphors in the BNC. The data fields are: BNC filename, sentence number, word number, sentence number of the antecedent, word number of the beginning of the antecedent, word number of the end of the antecedent, and the text of the antecedent, separated by a pipe (|) symbol.

annotation-aj0-pni-relationship.txt contains annotations of the relationship between *one*-anaphors and their antecedents. The fields are BNC filename, sentence number, word number, type of antecedent, relationship with antecedent. The numeric codes for type of antecedent and relationship with antecedent are contained in the `types` and `properties` lists in the `Annotation.Relationship.annotation` in the annotation code.