Definite Description Processing in Unrestricted Text

Renata Vieira

PhD
University of Edinburgh
April, 1998

Declaration

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

Edinburgh, 2 April 1998.

Renata Vieira

Abstract

Noun phrases with the definite article *the*, that we call DEFINITE DESCRIPTIONS, following (Russell, 1905), are one of the most common constructs in English, and have been extensively studied by linguists, philosophers, psychologists, and computational linguists.

In this dissertation we present an implemented model of definite description processing that is based on extensive empirical studies of definite description use and whose performance can be quantitatively measured.

In almost all approaches to discourse processing and discourse representation, definite descriptions have been regarded as anaphoric[1]; and the models of definite description processing proposed in the literature tend to emphasise the role of common-sense inference mechanisms.

Recent work on discourse interpretation (Carletta, 1996; Carletta et al., 1997; Walker and Moore, 1997) has claimed that the judgements on which a theory is based should be shared by more than one subject. On the basis of previous linguistics and corpus linguistics work, we developed several annotation schemes and ran two experiments in which subjects were asked to annotate the uses of definite descriptions in newspaper articles. We compared their annotations and used them to develop our system and to evaluate its performance.

Quantitative evaluation has become an issue in other language engineering tasks such as parsing, and has shown its usefulness also for theoretical developments. Recently, evaluation techniques have been introduced for semantic interpretation as well, as is the case for the Sixth Message Understanding Conference (MUC-6) (Sundheim, 1995). However, in this case, the emphasis was on the engineering aspects rather than on a careful study of the phenomena. Our goal has been to develop methods whose performance could be evaluated, but that were based on a careful study of linguistic evidence.

The empirical studies we present are evidence that definite descriptions are not primarily anaphoric; they are often used to introduce a new entity in the discourse. Therefore, in the model of definite description processing that we propose, recognising discourse new descriptions plays a role as important as identifying the antecedent of those used anaphorically.

Unlike most previous models, our system does not make use of specific hand coded knowledge or common-sense reasoning techniques; the only lexical source we use is WordNet (Miller et al., 1993). As a consequence, our system can process definite descriptions in any domain; a drawback is that our coverage is limited. Nevertheless, our studies serve to reveal the kind of knowledge that is needed for resolving definite descriptions, especially the bridging cases. The system resulting from this work can be useful in applications such as semi-automatic coreference annotation in unrestricted domains.

[1] Anaphoric expressions are those linguistic expressions used to signal, evoke or refer to previously mentioned entities.

Acknowledgements

First of all I would like to thank my supervisor Massimo Poesio who has been always very helpful and interested in our work. I am also very grateful to my second supervisor Jean Carletta. I've always had from them all the support I needed. My examiners Chris Mellish and Robert Gaizauskas have also contributed to make this work better. Thanks to the Brazilian agency, CNPq, which has funded this work. Thanks also to the support of the Centre for Cognitive Science and the Human Communication Research Centre during these years, also all the secretarial staff and computer support team. Special thanks to Elisabet Engdahl in my first year at the department. I would like to thank also from Edinburgh, the help I have had from Robin Cooper, Simone Teufel, Chris Brew, Janet Hitzeman and Ellen Bard. The people I met at DAARC96 gave me feedback and help at the beginning of this work, in special Ketjil Strand and Kari Fraurud; also thanks to Simon Botley and Tony McEnery. Another source of valuable help were the anonymous reviewers of Computational Linguistics whose comments have contributed to this work.

I am grateful to my family, specially my daughter Camila, for her love, patience and comprehension, sometimes waiting for hours in the lab for me until I spotted a bug. Hours which I said wouldn't take more than 10 minutes; and thanks to the Internet she had some fun during this waiting hours. I want to thank my relatives and friends who have helped me looking after her, specially at the writing up time when we were half world apart. Ponciano, Edi, Delia, Cleonice, Adri, Marcos, Claudia, Mari, Edelson, Gabriel, to all of you a huge thanks. Alvaro Moreira, Rafael Bordini, Debora Abdalla, it was just great to have you around, I love you. Rafael thanks for liking my thesis so much. Alvaro you are great. Debora, I am so glad you were here. Ana Tereza Martins and Ulisses Ferreira, thanks for showing me your sensibility and spontaneity. Lucia and John Falconer I'll miss you and the kids, thanks for all the fun. Norma and Alastair Martin, I'll always remember you. Daniele Zardo and Martin Escardo, you came just at the right time to look after me in the absence of Camila, thanks also for the lovely meals that gave me the energy for the writing up. Jonhantan Whale and Elaine Mowat, my native friends, I'll miss you. Patrick McGivern, Rodger Kibble, Zelal Gungordu, Saturnino Luz, Possi Gontijo, my in-house friends, made me feel better at work. Thanks David Tugwell, my English tutor, no, it is not his fault. Thanks Stephen Eglen, last minute helping. To the Scottish Highlands I am grateful for moments of beauty, peace and inspiration.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Noun phrases with the definite article *the* such as *the car*, that we will call DEFINITE DESCRIPTIONS[1] following (Russell, 1905), are one of the most common constructs in English. The word *the* is by far the most common word in the Brown corpus (Francis and Kucera, 1982), the LOB corpus (Johansson and Hofland, 1989), and the TRAINS corpus (Heeman and Allen, 1995). Definite descriptions have been extensively studied by linguists, philosophers, psychologists, and computational linguists[2].

In this dissertation we present an implemented model of definite description processing that is based on extensive empirical studies of definite description use and whose performance can be quantitatively measured.

Recent work on discourse interpretation (Carletta, 1996; Carletta et al., 1997; Walker and Moore, 1997) has claimed that the judgements on which a theory is based should be shared by more than one subject. On the basis of previous linguistics and corpus linguistics work, we developed several annotation schemes and ran two experiments in which subjects were asked to annotate the uses of definite descriptions in newspaper articles. We compared their annotations and used them to develop our system and to evaluate its performance.

Quantitative evaluation has become an issue in other language engineering tasks such as parsing, and has shown its usefulness also for theoretical developments. Recently, evaluation techniques have been introduced for semantic interpretation as well, as is the case for the Sixth Message Understanding Conference (MUC-6) (Sundheim, 1995). However, in this case, the emphasis was on the engineering aspects rather than on a careful study of the phenomena. Our goal has been to develop methods whose performance could be evaluated, but that were based on a careful study of linguistic evidence.

[1] We are not concerned with other cases of definite noun phrases such as pronouns, or possessive descriptions; hence the term definite description rather than the more general term definite NP. We sometimes use the shorter term description for definite description.

[2] See, e.g., (Russell, 1905; Christopherson, 1939; Strawson, 1950; Donnellan, 1972; Clark, 1977; Grosz, 1977; Cohen, 1978; Hawkins, 1978; Sidner, 1979; Webber, 1979; Clark and Marshall, 1981; Prince, 1981; Heim, 1982; Appelt, 1985; Löbner, 1985; Kadmon, 1987; Carter, 1987; Bosch and Geurts, 1989; Neale, 1990; Kronfeld, 1990; Fraurud, 1990; Barker, 1991; Dale, 1992; Cooper, 1993; Kamp and Reyle, 1993; Poesio, 1993).

In almost all approaches to discourse processing and discourse representation, definite descriptions have been regarded as anaphoric. Anaphoric expressions are those linguistic expressions used to signal, evoke or refer to previously mentioned entities. (See (Hirst, 1981) for a review of anaphora in natural language understanding.) Most models of definite description processing proposed in the literature tend to emphasise the role of common-sense inference mechanisms.

Traditional computational work on the problem has usually made use of world knowledge representation and inference mechanisms in order to deal with resolution of the so-called full definite NPs. Examples are: (Sidner, 1979), her algorithms are based on the availability of a knowledge network; Carter (1987) proposed a shallow processing anaphor resolver in which reasoning is intended to be minimally considered, but it does make use of specific hand coded knowledge, in special for the resolution of definite descriptions; the Core Language Engine (CLE) (Alshawi, 1990), although claimed to be a domain independent system, relies on a core lexicon and world knowledge reasoning, the required world knowledge has to be added by hand for each application. More recently, however, robust systems dealing with coreference have been proposed, for instance (Appelt et al., 1995; Gaizauskas et al., 1995). Following this line, in this work we propose techniques which avoid encoding specific knowledge for the testing of the system, and we do not make use of common sense reasoning techniques; the only lexical source we use is WordNet (Miller et al., 1993). As a consequence, our system can process definite descriptions efficiently in any domain; a drawback is that our coverage is limited. Nevertheless, our studies serve to reveal the kind of knowledge that is needed for resolving definite descriptions, especially the bridging cases.

On the empirical side of this work, we present evidence that definite descriptions are not primarily anaphoric; they are often used to introduce a new entity in the discourse. Therefore, in the model of definite description processing that we propose, recognising discourse new descriptions plays a role as important as identifying the antecedent of those used anaphorically.

The system resulting from this work can be useful in applications such as semi-automatic coreference annotation in unrestricted domains.

1.1 Organisation of the thesis

In Chapter 2 we present a review of linguistic research on definite descriptions. We first look at the variety of uses of the definite article; we then discuss in more detail some of these uses, paying special attention to the different types of relations that a definite description may establish with its antecedent. In this chapter we also present criticism of the view of definite descriptions solely as anaphoric expressions.

In Chapter 3 we present an annotation scheme for the classification of definite description use that we developed on the basis of the literature reviewed in Chapter 2. We then present two experiments on corpus analysis of definite descriptions and their reliability measures. Finally, we present a corpus study on bridging references, which is one of the classes in our annotation scheme.

In Chapter 4 we present a set of heuristics for processing definite descriptions exploiting the results of our corpus studies. We propose alternative treatments of each different type of description and we describe the integration of the heuristics into a working system.

Our system's evaluation according to the human annotation of the texts is presented in Chapter 5. First the heuristics are evaluated separately, then the system as a whole. We use statistical measures to compare the agreement between humans and the system. In Chapter 6 we review other systems which perform similar tasks.

Part of this material appeared before or will appear in the near future: Section §3.2 and Section §3.3 will appear in (Poesio and Vieira, 1997); Section §3.4 appeared in (Vieira and Teufel, 1997); parts of Chapter 4 appeared in (Vieira and Poesio, 1997; Poesio, Vieira and Teufel, 1997).

Chapter 2

Research on Definite Descriptions

The theories presented in this chapter help us to understand the problem of interpreting definite descriptions and inspired both the design of the schemes for corpus annotation and empirical analysis presented in Chapter 3 and the development of the computational experiments presented in Chapter 4.

Hawkins' descriptive analysis (Hawkins, 1978) is reviewed first in Section §2.1. Hawkins presents and exemplifies thoroughly the different uses of the article. After this descriptive introduction we discuss in Section §2.2 some representative works of the two dominant views on the meaning of definite descriptions: uniqueness (Russell, 1905; Russell, 1919) and familiarity (Kamp, 1981; Heim, 1982; Kamp and Reyle, 1993). Further discussion on the familiarity issue is presented on the light of Prince's work (Prince, 1981; Prince, 1992).

Section §2.3 contains some criticism of the dominant view in Natural Language Processing research that the anaphoric use is the standard type of description usage: Löbner (1985) claims that definite descriptions are functional concepts, and Fraurud's corpus analysis reveals that a large number of definite NPs are used as initial mention (Fraurud, 1990).

In Section §2.4 we look at research studying the various ways in which definite descriptions relate to their antecedents (Clark, 1977; Sidner, 1979; Strand, 1997). Several types of co-referent relations (or co-specification[1]) and associative relations (the most complex form of definite description use) are examined in detail.

In Section §2.5 we compare the terms referring to the various types of uses of definite descriptions introduced in the previous sections. We present four tables which relate examples of definite description use to different terminology, according to different authors. Finally, we present our conclusions in Section §2.6.

[1]We use the term co-reference in the sense of Sidner's terminology CO-SPECIFICATION (Sidner, 1979): a definite description co-specifies with its antecedent in a text, when such an antecedent exists, if the definite description and its antecedent denote the same entity. This is probably the most precise way of referring to the relation between an anaphoric expression and its antecedent; note that two discourse entities can co-specify without referring to any object in the world–e.g., in *The (current) king of France is bald. He has a double chin, as well.* In the same way we use the term *refer* in the general sense of specify or evoke.

2.1 Hawkins' descriptive list of the uses of the definite article

The wide range of uses of definite descriptions was already highlighted in (Christopherson, 1939). In the third chapter of his book, Hawkins (1978) further develops and extends Christopherson's descriptive analysis. According to Hawkins, the definite article may be used on the basis of a discourse antecedent (ANAPHORIC and ASSOCIATIVE ANAPHORIC USES) as well as independently from the previous discourse (SITUATIONAL, UNFAMILIAR WITH EXPLANATORY MODIFIERS and UNEXPLANATORY MODIFIER USES). We present below Hawkins' taxonomy. The examples are often repeated from or similar to those in (Hawkins, 1978).

Anaphoric Use

These are definite descriptions that refer back to an antecedent in the discourse (both description and antecedent evoke the same entity).

(2.1)
a. Fred was discussing *an interesting book* in his class. I went to discuss *the book* with him afterwards.
b. Fred was wearing *trousers*. *The pants* had a big patch on them.
c. Bill was working at *a lathe* the other day. All of a sudden *the machine* stopped turning.
d. Mary *travelled* to Paris. *The journey* lasted six hours.
e. *A man and a woman* entered restaurant. *The couple* was received by a waiter.

As seen in the examples, a definite description may use the same descriptive predicate as its antecedent, or any other capable of indicating the same antecedent (e.g., a synonym, a hyponym, a nominalization, summation, etc.).

Associative Anaphoric Use

Speaker and hearer may have (shared) knowledge of the relations between certain objects evoked by the discourse (the TRIGGERS) and their components or attributes (the ASSOCIATES): associative anaphoric uses of definite descriptions exploit this knowledge.

(2.2)
a. Bill drove past our house in *a car*. *The exhaust fumes* were terrible.
b. Bill bought *a new car* to please Mary but she didn't like *the colour*.
c. Fred was discussing *an interesting book* in his class. He knows *the author*.
d. I went to *a wedding* last weekend. *The bride* was a friend of mine. She baked *the cake* herself.

Immediate Situation Use

The next two uses of definite descriptions identified by Hawkins are used to refer to an object in the situation of utterance. The referent may be visible, or its presence may be inferred.

VISIBLE SITUATION USE This type of use occurs when the object referred to is visible to both speaker and hearer, as in the following examples:

(2.3) a. Please, pass me *the salt*.
b. Put it on *the table*.

IMMEDIATE SITUATION USE These are definite descriptions whose referent is a constituent of the immediate situation in which the use of the definite description is located, without necessarily being visible. This use is commonly found in notices such as:

(2.4) a. Beware of *the dog*.
b. Don't feed *the pony*.

At the same time the hearer is informed of the existence of these objects, he is also being instructed to use the immediate situation of utterance to determine which dog or pony is meant.

Larger Situation Uses

Hawkins lists two classes of definite descriptions that are used in situations in which the speaker appeals to the hearer's knowledge of entities existing in the non-immediate or larger situation of utterance—knowledge they share by being members of the same community, for instance. Whereas in associative anaphoric uses the trigger is an NP introduced in the discourse, in larger situation uses the trigger is the situation itself.

SPECIFIC KNOWLEDGE IN THE LARGER SITUATION This is the case in which both the speaker and the hearer know about the existence of the referent, as in the example below, in which it is assumed that speaker and hearer are both inhabitants of Halifax, a town which has a gibbet at the top of Gibbet Street:

(2.5) *The Gibbet* no longer stands.

GENERAL KNOWLEDGE IN THE LARGER SITUATION USE Specific knowledge is not a necessary part of the meaning of definite descriptions in larger situation use. While some hearers may have specific knowledge about the actual individuals referred to by a definite description, others may not. General knowledge about the existence of certain types of objects in certain types of situations is sufficient. An example is the following utterance in the context of a wedding (as first utterance between two people):

(2.6) Have you seen *the bridesmaids*?

Such a first mention of *the bridesmaids* is possible on the basis of the knowledge that weddings typically have bridesmaids. In the same way, a first mention of *the bride*, *the church service*, or *the best man* would be possible.

Note, however, that background knowledge may be different from individual to individual: one hearer might rely on his specific knowledge of a particular referent to interpret a description, whereas the other relies on his general knowledge to interpret the same description.

Unfamiliar Uses with Explanatory Modifiers

Hawkins classifies as unfamiliar those definite descriptions which are not anaphoric, do not rely on information about the situation of utterance, and are not associates of some trigger in the previous discourse. Hawkins groups these definite descriptions in classes according to their syntactic and lexical properties, as follows.

NP COMPLEMENTS One form of unfamiliar definite descriptions is characterised by the presence of a complement to the head noun.

(2.7) a. Bill is amazed by *the fact that there is so much life on Earth*.
b. The philosophical aphasic came to *the conclusion that language did not exist*.
c. Fleet Street has been buzzing with *the rumour that the Prime Minister is going to resign*.
d. I remember *the time when I was a little girl*.

NOMINAL MODIFIERS The presence of a nominal modifier is, according to Hawkins, the distinguishing feature of these phrases.[2]

(2.8) a. I don't like *the colour red*.

[2]The examples in (2.8) could perhaps be regarded as larger situation use (or else proper names), if one considers common knowledge of certain abstract concepts.

b. *The number seven* is my lucky number.

REFERENT ESTABLISHING RELATIVE CLAUSES Relative clauses may establish a referent for the hearer without a previous mention, when the relative clause refers to something mutually known.

(2.9) a. What's wrong with **Bill**? Oh, *the woman* **he** *went out with last night* was nasty to him. (But: ?? Oh, *the woman* was nasty to him.)
b. ...*the box (that is)* **over there**.

ASSOCIATIVE CLAUSES Associative clauses incorporate both the trigger and the associate of an associative anaphoric sequence. The modifiers of the head noun specify the referent with which the definite description is associated.

(2.10) a. I remember *the beginning of the war* very well.
b. There was a funny story on *the front page of the Guardian* this morning.
c. ... *the bottom of the sea*.
d. ... *the fight during the war*.

The syntactic structure of a definite description does not guarantee that it is unfamiliar. The definite description in (2.11), for example, does not refer to a discourse new entity, even though it has an NP complement.

(2.11) Frank told Sheriff Smith that Ringo would arrive on Thursday. *The news that Ringo would be in town* filled the Sheriff with worry.

Unexplanatory Modifiers Use

Finally, Hawkins lists a small number of modifiers (that he calls unexplanatory) which require the use of the definite article:

(2.12) a. My wife and I share *the same secrets*.
b. *The first person to sail to America* was an Icelander.
c. *The fastest person to sail to America* ...

There is nothing in the modifier which inform the hearer what is being referred to, Hawkins says: in the first of the examples above, the definite points merely to an identity between two sets of secrets.

Hawkins' detailed descriptive analysis is helpful to frame the ideas developed in the rest of the chapter. His taxonomy presents, however, some problems: sometimes descriptions may fit more than one class, or else they may vary according to the hearer. These problems will be discussed in Section §2.3.2.

2.2 Uniqueness and familiarity

The research concerned with the meaning of the definite article in English is divided into two main traditions: theories based on uniqueness and theories based on familiarity.

Russell's influential work (Russell, 1905; Russell, 1919) is the best known work in the uniqueness perspective (see Section §2.2.1). Examples of work following the familiarity approach to definite NPs are Discourse Representation Theory (DRT) (Kamp, 1981; Kamp and Reyle, 1993) and File Change Semantics (FCS) (Heim, 1982). They extend the semantic representation from the sentence level (as considered by Russell) to the discourse level. The goal of these authors is to account for the interpretation of new utterances with respect to a given context, and the integration of the utterance information into that context, dealing with referential processes. In Section §2.2.2 we discuss how DRT deals with definite descriptions, and we review Heim's FCS with respect to the same problem.

According to these two approaches the referent of a definite description is required to be either uniquely identifiable or familiar to the hearer. A great number of uses of the definite article can be accounted for using either familiarity or uniqueness but neither approach alone can account for all felicitous uses. In (Birner and Ward, 1994) we find a clear discussion of this problem.

In Section §2.2.3, we look at a different perspective on the notion of familiarity: Prince's theory introduces the important distinction between hearer and discourse familiarity.

2.2.1 Russell's theory of descriptions

In Russell's analysis, descriptions do not belong to the class of referring terms (or constants) like proper names, but to the class of denoting phrases like quantifiers. According to Russell, the meaning of a proposition expressed by an utterance of the form:

- *the F is G*

is represented by a quantifier phrase consisting in

1. an existential condition (there is at least one F),
2. a uniqueness condition (there is at most one F), and

3. a proposition predication (everything that is F is G).

This leads to the formal expression below.

- $(\exists x)(Fx\&(\forall y)(Fy \rightarrow y = x)\&Gx)$

More recent approaches to natural language semantics still follow Russells analysis of definite descriptions; an example is Montague semantics (Montague, 1974; Gamut, 1991), a very influential work in the field. However, while Russells analysis works well for functional concepts (descriptions such as *the father of Russell*, or *the center of the solar system*), the uniqueness condition is too strong for natural language description in general. Birner and Ward (1994) give (2.13) as an example of non-unique referent referred to by a definite description.

(2.13) [In a room with three equally salient windows.] It's hot in here. Could you please open *the window*?

Russell's analysis has been revised by several authors who have addressed the problem of making uniqueness relative to the relevant situation (Kadmon, 1987; Neale, 1990; Cooper, 1993).

2.2.2 Discourse Representation Theory and File Change Semantics

Both DRT and FCS propose the treatment of indefinite NPs as reference establishing terms (as opposed to a truth conditional quantifier analysis) and definite NPs are treated as anaphoric expressions. The ideal purpose is to provide a unified account for all indefinites and for all definites. The general idea is exemplified in (2.14): the indefinite NP, *a man*, introduces a new discourse referent ($d1$). The definite NP *he* in the second sentence is then identified to one already introduced referent satisfying that NP, which is ($d1$).

(2.14) [A man]$_{d1}$ walks. [He]$_{d2}$ sings.
$d1 = d2$

The most distinguished phenomenon treated by such approaches is the anaphoric linkage between sentences. However, these authors have found problems when integrating definite descriptions in their frameworks. Birner and Ward's example of an unfamiliar entity referred to by a non-anaphoric definite description is (2.15).

(2.15) In her talk, the lecturer introduced *the notion that syntactic structure is derivable from pragmatic principles*.

In (Kamp and Reyle, 1993), definite descriptions are presented under "Loose ends". The main problem recognised by these authors is the different ways in which descriptions can be used:

> We consider it a non-trivial task to identify and describe all the different purposes to which singular "the"-phrases can be put. And we see it as even more difficult to develop workable criteria that determine for each individual occurrence of a definite description which type of use it instantiates (page 253).

They propose that definite descriptions be treated by a "stopgap", leaving the difficulties to be dealt with later. This stopgap would simply introduce a condition of the form $\beta(x)$, interpreted as "x represents the individual denoted by β". A proper processing rule would tell how this should be reduced further, in a way which varies with each particular use of definite description.

In Heim's framework, indefinite NPs introduce new referents to the discourse. When a definite NPs is uttered it accesses a previously introduced referent. Heim explains it metaphorically by saying that a hearer when trying to understand the utterances in (2.16), constructs and update a file, which is empty at the beginning of the utterance. After (2.16.a) is uttered, the hearer takes two new cards. On card 1, the hearer writes *is a woman* and *was bitten by 2*. On card 2, the hearer writes *is a dog* and *bit 1*. After (2.16.b), the hearer updates the file by adding a new card 3 and writes on it *is a paddle* and *was used by 1 to hit 2*. The hearer also updates card 1 with *hit 2 with 3* and card 2 with *was hit by 1 with 3*.

(2.16) a. A woman was bitten by a dog.
b. She hit him with a paddle.

Heim shows with this example that the hearer seems to treat indefinites and definites with the following rule: for every indefinite, start a new card; for every definite, update a suitable old card.

Heim's theory accounts well for the anaphoric uses and the immediate situation uses (deictic descriptions) of definite descriptions. To account for these latter uses one has to consider that objects in the non-linguistic context are also represented by cards.[3] Heim acknowledges that some definite descriptions fail to satisfy the condition that a card should exist previous to its utterance in a discourse:

> Every use of a definite NP requires that there already be an appropriate card in the file... But in fact there are many uses of definites, in particular definite descriptions, which do not fit this theory (page 370).

[3]This assumption, however, faces the problem of how to delimit the non-linguistic information that is needed for the semantic analysis.

To accommodate the uses of descriptions which do not satisfy the stated conditions she proposes an adjustment of the discourse file by adding the information needed: a new file card is introduced not on the normal conditions but under accommodation. She says, for associative descriptions, that a card introduced through accommodation has to be linked by cross-references to some already present file card:

> ... the cross-references form a bridge that connects the new discourse referent to the network of discourse referents that is already established (page 373).

For immediate or larger situation uses the new card may be linked to the discourse situation.

> There seems to be no exception to this requirement. Mere addition of a card without cross-references (as happens with indefinites) is never acceptable in accommodation (page 374).

She also observes that some definites fulfil the requirement for a cross-reference automatically, her example is:

(2.17) John read a book$_i$ and wrote to [the woman who had written it$_i$]$_j$.

The descriptive content of NP$_j$ explicitly contains a cross-reference to card i. Therefore, she says, in these cases there is no need for an additional requirement for cross-references, since such a requirement is generally in force.

The question here is whether this mechanism is only a repair move or whether it represents the real character of the definite description.

2.2.3 Prince's assumed familiarity

Prince studied in detail the connection between the speaker/writer's and hearer/reader's assumptions about each other and the linguistic realization of noun phrases (Prince, 1981; Prince, 1992). Although she studies noun phrases in general, the taxonomy she proposes has proved equally useful for our analysis of definite descriptions in particular. What is original and especially interesting in Prince's work is the important distinction between two kinds of familiarity, a distinction not explicitly observed in Hawkins' theory or in Heim's.

Prince criticises the traditional binary distinction between 'given' and 'new' discourse entities as too simplistic, and proposes a much more detailed taxonomy of 'givenness'—or, as she calls it, ASSUMED FAMILIARITY—meant to address this problem. She distinguishes between discourse and hearer familiarity, as seen below.

Hearer new / Hearer old

One factor affecting the choice of a noun phrase, according to Prince, is whether a discourse entity is old or new with respect to the hearer's knowledge. Typically, a speaker will use a proper name or a definite description when he or she assumes that the addressee already knows the entity that the speaker is referring to, as in (2.18).

(2.18) I'm waiting for it to be noon so I can call *Sandy Thompson*.

On the other hand, if the speaker believes that the addressee does not know of Sandy Thompson, in general, an indefinite will be used:

(2.19) I'm waiting for it to be noon so I can call *someone in California*.

Discourse entities can also be new or old with respect to the discourse model.

Discourse new / Discourse old

According to Prince, an NP may refer to an entity that has already been 'evoked' in the current discourse (TEXTUALLY EVOKED), or it may evoke an entity which has not been previously mentioned (SITUATIONALLY EVOKED, UNUSED, INFERRABLE, CONTAINING INFERRABLE, BRAND NEW). 'Discourse novelty' is distinct from 'hearer novelty': both Sandy Thompson in (2.18) and the *someone in California* mentioned in (2.19) may well be discourse new even if only the second one will be hearer new. On the other hand, for an entity being discourse old entails it being hearer old.

Assumed Familiarity

BRAND NEW An NP may introduce an entity which is both discourse and hearer new. Brand new entities are usually introduced by indefinites, such as *someone in California* in example (2.19).

BRAND NEW ANCHORED A new entity is anchored, according to Prince, if it is linked to another discourse entity, this link is contained in the NP representing the entity and this link is not itself new. An example is given in (2.20).

(2.20) *A guy I work with...*

Prince seems to be considering only indefinites in this class, but a definite such as *the guys I work with* could perhaps be regarded as brand new anchored in the same sense. [4]

EVOKED NPs may invoke situationally evoked or textually evoked entities. Only textually evoked entities are discourse old. Situationally evoked entities correspond to Hawkins' visible/immediate situation use.

UNUSED NPs may evoke hearer old but discourse new entities. Unused NPs describe entities that are known to the speaker/hearer but which havent been mentioned (used) previously in the discourse. These are like those cases called by Hawkins larger situation/specific knowledge.

INFERRABLE Some discourse entities are not discourse old or even hearer old, but they are not entirely new, either. Hawkins called such uses of definite descriptions associative anaphoric: *a book...the author*. Prince called such entities inferrables. Prince did not introduce a class for those entities which are inferrable from the situation (Hawkins' larger situation/general knowledge); they will be referred to in Section §2.5 as SITUATIONALLY INFERRABLE.

CONTAINING INFERRABLE Prince proposes a category for entities which are like inferrables, but whose connection with previous hearer's knowledge is specified as part of the NP itself. Her example is *the door of the Bastille* in (2.21).

(2.21) *The door of the Bastille* was painted purple.

At least three of the unfamiliar uses of Hawkins–NP complements, referent-establishing relative clauses, and associative clauses–fall in this category. As pointed out by Prince (and Fraurud, see Section §2.3.2), the distinction between containing inferrable and unused is sometimes ambiguous: what is unused for one reader may be containing inferrable for another, depending on their individual knowledge background.

2.3 Critiques to the familiarity/anaphoric approach to definite NPs

Although most authors acknowledge the great variety of description use, each of the two traditions discussed above tries to identify a basic meaning. The familiarity approach to definite NPs regards anaphoric definite descriptions as the prototypical use, especially in discourse research. This attitude has been criticised by some authors. In this section we present some of these critiques.

[4] There are also some definite descriptions that describe new entities and are linked to entities that are new, as in *the footsteps of a yeti*. Their place in Princes framework is not clear.

2.3.1 Löbner's theory

Löbner (Löbner, 1985) criticises the assumption made in DRT-like (Kamp and Reyle, 1993) theories that anaphoric use is the prototypical use of definite descriptions. He observes that the interpretation of descriptions may depend on arguments and attributes given in the referring act itself or by the immediate situation, and not only on textual antecedents. He also criticises the idea that uniqueness is a property of definite descriptions, and that they behave like quantifiers; he takes descriptions to be terms like proper names.

Löbner adopts Christopherson's (1939) view according to which the fundamental property of definite NPs is that they refer unambiguously. Löbner claims that the definite article indicates that the noun is to be taken as a FUNCTIONAL CONCEPT (FC). This idea is based on the distinction between sortal and relational nouns: sortal nouns identify a class (*woman*), while relational nouns describe objects as standing in a certain relation to others (*wife*). Functional nouns are a subclass of relational nouns. Functions relate objects unambiguously (one to one) to others: they assign values to arguments. Functional concepts identify a referent when the situation and proper arguments are given.

Löbner's classificatory scheme, is based on the distinction between SEMANTIC and PRAGMATIC DEFINITES. Semantic definites are those cases in which the interpretation is independent of the utterance's previous discourse or immediate context of utterance; the general situation, however, is always an argument[5]. The semantic definites Löbner lists correspond to Hawkins' larger situation and unfamiliar uses. Pragmatic definites, on the other hand, are essentially dependent on the particular context of utterance for their non-ambiguous interpretation.

Semantic Definites

Löbner defines a semantic definite as an NP denoting a functional concept. According to the number of arguments definites take, they are classified into FC1s, FC2s and FC3s. All of them involve the general situation as one of their arguments, often implicitly.

SEMANTIC FC1S These semantic definites are concepts for objects which play a unique role in a given situation. This class includes:
a) proper names;
b) sortal nouns followed by a proper name of some sort;
c) cases in which a subordinate clause specifies an abstract sortal head as FC1s[6];
d) combinations of certain adjectival attributes (superlatives, ordinals, as well as *next, last, only*, etc.) with sortal or relational nouns forming a complex FC1; and
e) those cases called by Löbner SIMPLE FC1s which are dependent on temporal and spatial location[7].
Examples of each type are shown in (2.22):

[5]This argument relates the description to the location, time and circumstances of the utterance.

[6]The subordinate clauses in these cases are like disambiguating attributes (see pragmatic endophoric definites below, example (2.27)); however, they specify the referent of the sortal noun instead of just relating it to some other entity, e.g. *rumour* in (2.22 c).

[7]These (and proper names) are the only cases based exclusively on the speaker's common general knowledge . The other cases present an appositive-like structure or clausal adjectival complements.

(2.22) a. *the Empire State Building, the London Symphony Orchestra*;
b. *the year 1984, the word "the", the opera Rigoletto*;
c. *the rumour that Reagan is going to resign, the dream to become rich*;
d. *the next/last/third president of the association*;
e. *the weather, the time, the air, the moon*.

All these definite descriptions yield functional concepts. They always take one argument relative to the given situation.[8] They name something unambiguously which may not have been mentioned before; hearers do not need to find this named entity in the immediate context. These descriptions correspond to some of Hawkins' larger situation and unfamiliar uses.

SEMANTIC FC2S WITH EXPLICIT ARGUMENTS Generally an FC2 is connected to its second argument by a possessive relation (in the sense that something or someone 'has' something). These cases syntactically consist of a definite article which precedes a complex expression containing the FC2 noun and a PP of the form *of NP*, as the examples in (2.23). A FC2 results in a FC1 when complemented with its argument[9].

(2.23) a. *the president of the U.S.*;
b. *the meaning of the definite article*.

These descriptions correspond to Hawkins' unfamiliar uses (associative clauses). Löbner says that FC2s have the property of inheriting their argument status; he calls this property TRANSPARENCY. If the argument is anaphoric (for instance, *the book*) the FC2 *the author of the book* is also anaphoric. If the argument is indefinite the whole NP is indefinite (or weak definite): for instance, *the footsteps of a yeti*. For further discussion on weak definites see also (Poesio, 1994).

SEMANTIC FC2S WITH IMPLICIT ARGUMENTS These descriptions depend on the immediate physical environment, which functions as an IMPLICIT DEICTIC ARGUMENT. Hawkins' introductory situational uses fall in this category.

(2.24) This is *the clutch*.

[8] These concepts assign a functional value to situations. Descriptions such as *the sun, the moon, the Earth* assign the same value to a wide range of locations and time. For other descriptions the referents or values are more locally determined: *the weather, the atmosphere*. Proper names usually apply to a certain referent relatively to a domain of situations. A name like *Paul* is dependent on the social circumstances for its unambiguous interpretation. In many languages personal names are used with the definite article.

[9] Löbner notes that the number of arguments referring to the situation may in fact vary: compare, for instance, descriptions such as *the price of an apartment, the price of an apartment in Korea, the price of an apartment in Korea in the eighties*.

In the example above, the argument is a car in the immediate physical environment. Another example of implicit deictic FC2 is Hawkins' larger situation use based on general knowledge:

(2.25) *The Prime Minister* has resigned.

The location of the utterance is included in the territory of a state to which the description refers (indirectly). Löbner also includes in this class those expressions which refer indirectly to referents previously introduced in the discourse, such as *a book... the author*, referring to them as FC2s with IMPLICIT ANAPHORIC ARGUMENT. (These cases correspond to Hawkins' associative anaphoric uses.) Löbner states that the crucial condition under which FC2s with implicit arguments are possible is that the head noun in these uses provides a two-place functional concept for which there is an appropriate argument in the immediate context (physical or linguistic).

In (Löbner, 1996) it is claimed that the semantic/thematic roles of verbs are also FC2s. For every reading event, he says, there is the role of the reader and the role of the read; underlying theses roles are the functional concepts *the reader of this reading event*, and *what is read in this reading event*. Further roles may be connected to a reading event, such as medium, time, location, speed and others. In this later paper Löbner adopts (as Sidner (1979), see Section §2.4.2) a frame-like semantic network to explain FC2s with implicit anaphoric argument.

SEMANTIC FC3S In these cases the definite article precedes a noun that is complemented with two arguments.

(2.26) *the distance between A and B*

Pragmatic Definites

Pragmatic definites have non-functional head nouns[10] and thus depend on the particular situation or immediate context for unambiguous reference. They are divided in anaphoric, endophoric and deictic uses.

ANAPHORIC These descriptions are resolved to a previously introduced referent (as in *a book... the book*). Hawkins' anaphoric uses fall in this category.

ENDOPHORIC /CATAPHORIC These definites have relational or sortal head nouns with disambiguating attributes, as in example (2.27). This use is classified by Hawkins' as unfamiliar with referent establishing relative clauses.

(2.27) *the woman Bill went out with last night*.

[10]Notice that it is the use, not the noun itself, that is relational or sortal.

DEICTIC These uses refer to the immediate context, and correspond to Hawkins' immediate situation uses.

Finally, Löbner builds on the idea that endophoric, cataphoric, anaphoric and deictic uses are all like functional concepts. The idea is that these descriptions denote functional concepts on pragmatic grounds. Löbner's analysis makes it clear that descriptions, besides being resolved (disambiguated) with discourse antecedents, may represent functional concepts, being disambiguated by arguments and attributes. Thus, the list of descriptions that may be discourse new, according to Löbner, consists of:

- semantic FC1s, or functional concepts (which are relative to the situation, their one argument),
- semantic FC2s with explicit arguments (FC2s result in FC1s when complemented with their arguments),
- FC2s with implicit argument, when the argument is given by the non linguistic context[11] (what he calls deictic argument),
- endophoric pragmatic definites disambiguated through their attributes.

DRT and File Change Semantics account for anaphoric and deictic pragmatic definites only. This restricted coverage is the target of Löbner's criticisms.

2.3.2 Fraurud's study of first mention and subsequent uses

Fraurud (1990) presents a corpus-based study of definite NPs use in Swedish texts[12], based on a binary classification scheme:

- SUBSEQUENT MENTION (corresponding to Hawkins' anaphoric definite descriptions and Prince's discourse old), and
- FIRST MENTION (including all other definite descriptions).

Fraurud's notion of subsequent mention is defined in terms of co-referentiality (NPs referring to the same entity). She notes that an NP which is co-referent with another NP is not necessarily anaphoric. The interpretation of an anaphor is crucially dependent on the identification of a discourse referent introduced by an antecedent (as is usually the case for pronouns); whereas co-referentiality only implies that a discourse referent previously mentioned in the discourse is evoked by an NP, but the NP's interpretation need not to be essentially dependent on this previous mention (as for subsequent mention of proper names).[13]

[11] Semantic FC2s whose argument is anaphoric (i.e., introduced in the discourse, as in Hawkins' associative anaphoric uses) are not discourse new.

[12] Frauruds corpus is distributed between the following text types: brochures, newspapers, textbooks and debate books (all professional, non-fiction prose).

[13] Löbner's semantic FC1s, FC2s with explicit argument, and pragmatic endophoric are not anaphoric (in the strict sense), but can be used in subsequent mentions.

Fraurud's simplified taxonomy is due to the fact that she was primarily interested in verifying the empirical basis for the claim that indefinite NPs trigger the establishment of a new discourse referent in a discourse model while definite NPs trigger the search for or the retrieval of a prior discourse referent. She recognises that the existence of first mention definite NPs is acknowledged in the literature, but criticises the fact that they tend to be treated as secondary relative to the anaphoric use of definite NPs, giving as example Heim's File Change Semantics (discussed in Section §2.2.2).

In her study Fraurud observes that only 34.8% of initial mention NPs were actually indefinites, and of all indefinites, only 9.4% were referred back to. She points to the problem for NP processing of having a vast number of entities made available for anaphoric reference and just a small portion being referred to. But perhaps the most interesting result is the large proportion of definite NPs in first mention uses found in her corpus: 60.9%.

Also interesting is Fraurud's observation about the syntactic complexity of first mention definite NPs. She claims that genitive/possessive constructions of the form *the X's Y* or *the Y of X*, postposed prepositional phrases, and restrictive adjectival modifiers make the NP "self-contained'. These NPs, as Löbner's FC2s, explicitly sign their relation to other referents; and therefore, one would expect that complex definites are more often used as first than subsequent mention. And in fact, 75% of the complex definite NPs in her corpus were first mention. She also notes that a functional sub-classification such as Hawkins' would give a better picture of the frequencies of different types of first mention definite NPs.[14]

A critique of Hawkins' taxonomy

Fraurud criticises Hawkins' taxonomy, claiming that it imposes methodological problems for a classification. According to Fraurud, the interpretation of some first mention definite NPs (Hawkins' associative anaphora) often involves more than one ANCHOR (Fraurud's term for an associated antecedent), or different sources of the same anchor may be available. For example, she notes that the reader of a Swedish newspaper can equally well interpret the definite description *the king* in an article about Sweden by reference to the common knowledge (situational anchor) or by reference to the content of the article (discourse anchor): an ambiguity between larger situation and associative uses. In the examples below, given by Fraurud, different anchors from different sources are used at the same time in the interpretation of the definite description *the next train*. Consider (2.28), uttered in a ticket office of the central station of Stockholm: the interpretation of *the next train* can rely both on a situational anchor (Stockholm) as well as an anchor given by the previous discourse (Gothenburg).

(2.28) I am going to Gothenburg. When does *the next train* leave?

[14] This has been carried out in this thesis and is reported in Section §5.3.2.

In a text about the European Union, she found (2.29), in which the description *the link* referred to is the link between EU (previous mention) and NATO (post-modifier). It would be difficult to place cases like this in one class or another.

(2.29) Through De Gaulle *the link to NATO* was broken.

Fraurud criticises Hawkins' class "unfamiliar uses with explanatory modifiers", saying that unfamiliarity is not necessarily a property of NPs with explanatory modifiers. In an utterance like *the engine of my car*, the receiver may or may not have previous knowledge of the referent (an ambiguity between unfamiliar and larger situation/specific knowledge). She considers, however, that some instances may "fit well" with Hawkins' types, whereas others would need a more flexible description of the information and processes involved. (The corpus analyses presented later in this dissertation give an idea of the proportion of cases that "fit" such a taxonomy.)

2.4 How descriptions relate to antecedents

In this section we look in more detail at the anaphoric and associative anaphoric role of definite descriptions by discussing the various ways in which a definite description may relate to its antecedent. The main distinction is between co-referential and associative relations: in the first case, the description refers to an entity introduced by an antecedent; in the second case it refers to a different entity which is associated to a previously mentioned entity. Each of these two classes may be further divided into subcategories. As we will see below, several authors have proposed different subclassifications.

2.4.1 Clark's bridging references

Clark's paper "Bridging" (1977) is concerned with the construction of implicatures as part of the process of comprehension (understood as the computation of an antecedent). He identifies the possible semantic relations between the referring expression and its antecedent. Clark is only concerned with implicatures derived from textual relations, which correspond to Hawkins' anaphoric and associative anaphoric uses. The distinctions he made are reviewed here for the specific case of definite descriptions.

Direct reference

Clark notes that a description often makes direct reference to previously mentioned objects, events or states.

IDENTITY Examples given for this class are:

(2.30) a. I met *a man. The man* told me a story.
b. I *ran two miles. The run* did me good.

He also gives as an example of direct reference (identity) the following:

(2.31) Her house was large. *The size* surprised me.

In the case of (2.31), the term "direct" refers to the fact that *the size* (of the house) has already been mentioned (when describing it as being large). This notion of "identity" does not seem to conform to a notion of co-referentiality. In other approaches (see Strand, Section §2.4.3, for instance) the reference *the size* is seen as associated to the noun *house* rather than to the adjective *large*.

PRONOMINALIZATIONS These are cases in which the description uses only a subset of the properties that characterise a previously mentioned entity. We have a continuum: *an elderly gentleman, the elderly gentleman, the elderly man, the gentleman, the man, the oldster, the adult, the person, he*. The semantic relations of synonymy and hypernymy belong to this class together with the use of pronouns.

(2.32) I met *an elderly gentleman. The man* told me a story.

EPITHETS This class contains those cases in which the bridging reference adds new information to the entity referred to.

(2.33) I met *a man. The bastard* stole my money.

In (2.33) the antecedent for *the bastard* is the entity referred to by *a man*—that entity is also *a bastard*, but this information is new. The extra information is concerned with the speaker's opinion of the facts rather than the facts themselves.

SET MEMBERSHIP In this class are those cases in which the description picks out an element from a previously mentioned set.

(2.34) a. I met *two people. The woman* told me a story.
b. *I swung three times. The first swing* missed by a mile.

Indirect reference by association

Clark, like the other authors we have discussed, notes that the description may not have a directly mentioned antecedent but one which is closely related to it. He notes that the associated information varies in its predictability from absolutely necessary to quite unnecessary, distinguishing three levels:

NECESSARY PARTS

(2.35) a. I entered *the room. The ceiling* was high.
b. I entered *the room. The size* was overwhelming.

In the first example above, *the ceiling* can be definite, with the following implicature: the room mentioned has a ceiling; that ceiling is the antecedent of *the ceiling*.

PROBABLE PARTS

(2.36) a. I entered *the room. The windows* looked out to the bay.
b. *I went shopping. The walk* did me good.

In cases like these, there is no guarantee, for instance, that the room has windows or that going shopping means walking, but these are likely.

INDUCIBLE PARTS In these cases the implicature cannot be assumed directly nor probably, only through induced inference.

(2.37) I entered *the room. The chandeliers* sparkled brightly.

Indirect reference by characterisation

A description may characterise a role played in an event or circumstance mentioned earlier. Clark presents a variety of such cases:

NECESSARY ROLES

(2.38) a. *John was murdered. The murderer* got away.
b. *I went shopping. The time I started* was 3 p.m.

The implicature in (2.38.a) is such that some person performed John's murder; that person is the antecedent for *the murderer*.

OPTIONAL ROLES

(2.39) a. *John died. The murderer* got away.
b. *John was murdered. The knife* lay nearby.

Clark observes that often noun phrases contain as part of their specification the information of how they relate to other events as in *the person who murdered John, the knife with which it was done*. Adjectives can carry out a characterising function too, as in *the guilty party got away*. He says that what adjectives, relative clauses and derived nouns (such as *murderer*) do is to pick out the role the intended antecedent plays in the previously mentioned events. Clark comments that sometimes the distinctions between parts and roles may be impossible to maintain.

Relations of reasons, causes and consequences

As we have already seen, the antecedent of a bridging description is often an event and not an object and may give the reason for, cause of, or consequence of other events or states. Clark's examples for this class do not include the use of definite descriptions. We present instead an example from our corpus (discussed in Chapter 3).

(2.40) *An earthquake... The suffering people are going through...*

2.4.2 Sidner's co-specification and specification rules

Sidner (1979) lists several ways (rules) in which a full definite NP may derive its co-specification or specification from the FOCUS (a list of the most salient elements in the discourse, i.e., what the discourse is about). The focus for definite description interpretation includes:

- the CURRENT FOCUS, the most salient element in the last sentence according to a set of rules proposed by Sidner;
- the POTENTIAL FOCUS, elements in the last sentence other than the current focus;
- the STACKED FOCUS, the set of current foci previous to the last sentence[15].

Sidner presents several algorithms that work together to resolve anaphoric NPs and to keep track of the discourse focus. Her algorithms rely on a semantic network that encodes elements and their associations[16], provides links expressing their general class, and provides for inheritance of associations.

The rules listed by Sidner are the following:

[15] It is not clear if the Actor Focus Stack should be also considered for definite description interpretation.
[16] Sidner relies on work by (Hendrix, 1975; Roberts and Goldstein, 1977; Bobrow and Winograd, 1977) on semantic networks and frame systems.

Explicit Backwards Co-specification

CO-SPECIFICATION 1 Definite description and focus have the same head and no new information is introduced by the definite.

(2.41) *A small office... The office*

She mentions the difficulty imposed by definites with new information since it is not clear whether they co-specify with the focus or refer to a new discourse element. Cases in which co-specifying definite NPs introduce new information are not very common but they do occur in natural language texts. Examples from our corpus are: *the campaign... the Dinkins campaign, the dollar... the U.S. dollar*. [17] Clark has also observed that a definite description may specify or add new information to the antecedent (epithet).

CO-SPECIFICATION 2 The definite's head noun lexically generalizes that of the focus and has no restrictive postmodifiers.

(2.42) *A ferret... The animal...*

She claims that generalizations accompanied by restrictive relative post-nominal modifiers fail to co-specify with the focus. This class is similar to Clark's pronominalization.

Implicit Backwards Specification

Here the definite does not co-specify with the focus. It is said, instead, to specify an element closely related to the focus by association.[18] She proposes the following restriction on the elements available for the computation of specifications: NPs in the stacked foci are not considered as focus for these cases. Sidner says that stacked foci do not seem to be used in this way perhaps because the additional processing time would not

[17] It is difficult to judge on new or old information. Often the information is not new, but only indirectly connected to the antecedent. In the sequence: a) *rule changes proposed last summer*... b) *the rules*... c) *the proposed rules*, strictly speaking, *proposed* in (c) is no new information, but this known information was not introduced explicitly with the antecedent *rules* in (b), but given indirectly in (a).

[18] This notion of association is not very precise. In (Sidner, 1978), it is said that "any entity closely associated with the entity which represents the focus can be mentioned using a simple definite NP". Sidner then gives examples of acceptable and unacceptable cases:

(2.43)
a. I went to *a new restaurant* with Sam.
b. *The waitress* was nasty.
c. *The food* was great.
d. *The soup* was salty, but the wine was good.
e. * *The rug* was ugly.

make it possible to extend the judgements to the focus stack. This means that a definite description can only specify an element in the previous sentence.

ASSOCIATED SPECIFICATION The definite names an entity associated with the focus directly or by inheritance on the network hierarchy. The inferences made in the association involve common sense knowledge about the world.

(2.44) *A meeting... The participants...*

INFERRED SPECIFICATION As above but the inferences involve hearers' suppositions which are not necessarily true.

(2.45) *The dead heiress... The murderer*

This class may include a broad range of relations.

SET-ELEMENT SPECIFICATION The focus is a set, the description is singular and has the same head as the focus and additional modifiers whose role is to determine which member of the set is being discussed.

(2.46) There were *clowns* performing in the square.
The clown with the unicycle did a fantastic stunt.

Sidner comments that these cases are easier to distinguish than other specifications, because the head noun is the singular form of the noun phrase represented in the focus. There are, however, set-element sequences such as *a couple...the woman* which would involve knowledge of set-element relations as well as generalization and/or associations. It is not clear whether cases like this would be handled by the associated specification or inferred specification rules.

COMPUTED SPECIFICATION The specification of the description may be computed from that of the focus. The description has an ordinal modifier, the same head as the focus and no relative clause modifiers. Sidner observes that descriptions containing full relatives (such as *the first person to sail to America*) use the relative clause and not the focus to compute its specification.

(2.47) *A meeting... The last meeting but two*

With the restrictions she imposes, Sidner misses cases like *A conference ... the first talk*, or Clark's example *I swung three times. The first swing...* (as seen in Section §2.4.1). Again, it is not clear if these would be cases treated by the rules of associated or inferred specification.

When no relation can be established, Sidner says that definite NPs with no modifiers are odd uses, but in (Sidner, 1978) she comments that descriptions such as *the moon, the sun, the Earth* have default referents in initial sentences. For those descriptions which have modifiers she says that they specify outside the discourse context.

2.4.3 Strand's taxonomy of linking relations

Strand's approach (Strand, 1996) is also mainly concerned with those cases of definite description use in which an explicit contextual relation (LINK) holds between the description and an antecedent (ANCHOR). Strand, as Sidner, assumes the availability of a semantic representation of the text (in this case, DRT, (Kamp and Reyle, 1993)) and inference mechanisms. He proposes a taxonomy of linking relations in which five main classes are distinguished along with fifteen subclasses. They are as follows:

Co-referentiality

The antecedent and definite refer to the same entity through identical or different description.

IDENTICAL HEAD The anchor and the definite description share the same head noun

(2.48) *A yellow car... The car...*

GENERALIZATION The definite description is more general than or is a synonymous of the antecedent.

(2.49) *A car... The vehicle...*

SPECIFICATION The description is more specific than the antecedent.

(2.50) *A car... The sedan...*

REDESCRIPTION The definite description is a fully alternative description of the antecedent which neither entails nor is entailed by any conditions on (properties of) the antecedent.

(2.51) *A car... The notorious wreck...*

Strand's co-referential class differs from Clark's direct reference. Whereas Clark classifies *The house was large... The size surprised me* as direct reference, Strand does not. Other differences in their taxonomies are discussed in Section §2.5.

Narrowing

The definite is part/member or an argument/role of the antecedent.

SET-MEMBER The description is a member or a subset of the set indicated by the antecedent.

(2.52) *A school class... The girls...*

WHOLE-PART The description constitutes a part of its antecedent.

(2.53) *A car... The engine...*

EVENT-ARGUMENT The description is an argument of an antecedent event.

(2.54) *John was murdered. The murderer...*

Widening

These are cases which expand on familiar sets.

MEMBER-SET The description is a set of which the antecedent is a member or a subset.

(2.55) *John and his nephew... The family...*

PART-WHOLE The description has the antecedent as its part.

(2.56) *A wall... The building...*

Adjoining

PART-PART The antecedent and description are members of the same state or parts of the same whole.

(2.57) *Last Wednesday... The next day...*

POSSESSOR-THING The antecedent possesses the description.

(2.58) *A professor... The car...*

Delimitation

In these cases the anchor may be seen as an argument to the description.

ARGUMENT-EVENT The description is an event in which the antecedent is an argument delimiting its denotation.

(2.59) *Israel and Egypt... The peace agreement...*

SUBCATEGORIZATION The description subcategorizes for something of the antecedent's type. This applies to so called 'relational nouns' like *father, weight, price, owner, driver*, etc.

(2.60) *A bicycle... The price...*

TIME-ANCHORED The time region indicated by the antecedent gives a more delimited or unambiguous reading to the description.

(2.61) *Last Wednesday... The news...*

SPACE-ANCHORED The space region indicated by the antecedent delimits the description.

(2.62) *A Greek village... The taxi drivers...*

Strand also mentions the existence of implicit or inferred anchors: for instance, when someone is telling about a visit to a Greek village, the (implicit) time of visit may be an anchor to the referents of the descriptions.

Strand acknowledges the problem of multiple anchors/links being available for a description resolution. He says that one should give preference to the most informative link and that identity should be preferred whenever possible. However, besides the problem of deciding between identity and non-identity, it seems hard to find a way of identifying a 'most informative' link. Strand mentions that an opposite approach is one like Sidner's, where a saliency order is followed.

2.5 Comparison of terminology

In this section we compare the classifications presented in the previous sections. We present four tables which relate examples of definite description use with the classes identified by the authors we have discussed. The first two tables (Table 2.1 and Table 2.2) describe the anaphoric and associative uses respectively. Seven different schemes are listed (Hawkins, 1978; Prince, 1992; Fraurud, 1990; Löbner, 1985; Clark, 1977; Sidner, 1979; Strand, 1997). The last two tables (Tables 2.3 and 2.4) consider only four of the authors, since not all authors refer to the phenomena presented there (situational and unfamiliar uses).[19] The terms that appear in the tables in italics are our guesses for the examples not explicitly discussed by the author of the corresponding scheme. Question marks were placed where the authors were generally silent about the case, and it was not clear whether their classification would apply or not to the example.

We can see that those authors who present a more comprehensive characterisation of uses of definite descriptions (Hawkins, Prince, Fraurud and Löbner) do not discriminate anaphoric and associative descriptions in as much detail as the others (Clark, Sidner and Strand) do. On the other hand the first authors pay special attention to situational and unfamiliar uses. Also note that there is no absolute consensus about the sub-classifications of the various uses.

Table 2.1 lists the anaphoric uses. Hawkins and Prince does not make any distinction among them. It is not clear whether Prince would consider the definite description in a sequence like *he travelled... the journey* as textually evoked, nor if Fraurud would consider that as subsequent mention. Löbner only refers explicitly to "direct anaphora", those cases based on an antecedent with identical head. But what he calls pragmatic anaphoric seems to apply well for all examples in the table.[20] Clark, on the other hand, distinguishes among four different ways in which a co-reference relation may be realized, but he is silent about the cases that Strand calls specification and widening. Sidner

[19] Although Sidner notices that definite descriptions may specify outside the linguistic context, she does not explain in which different ways. Strand briefly mentions the existence of implicit and inferred anchors. Clark is only concerned with discourse relations.

[20] Löbner says that the construction of a universe of discourse is comparable to the braiding of a complex multi-dimensional network, with object and event nodes; every node is a potential discourse referent, and anaphoric descriptions are used to refer to nodes in the net, usually providing only sortal information for the retrieval of their referent.

considers only two types of co-reference: identical head and generalization. Both Fraurud and Strand observe for a sequence like *a man, a woman ... the couple* a difference in the entities represented by the description and antecedent. Strand (1997) explains that events in his framework are represented by discourse referents and a link for cases like *he travelled... the journey* would be of the coreferentiality class. For named entities he explains that usually the relation is coreference, and the subclass specification or re-description. Clark and Strand give the most comprehensive account for the anaphoric use.

Table 2.2 summarises the classifications of associative uses and is the most complex of all. It is difficult to complete the table for each different author, since usually they are not explicit about all the possible associations capable of linking bridging descriptions with their anchors. Hawkins reckons the difficulty in providing the defining parameters for the set of possible associates; he then comments on the more general defining characteristics of these associations. He says that speaker and hearer share general knowledge of relationships between triggers and associates, usually part-of relations and attributes. It is not clear if he would consider event roles[21] or cases involving hearers' supposition as associative anaphora. Prince is not specific, either, about which are the possible associations between bridging descriptions and their anchors. She calls them all inferrables and says that "when a speaker evokes some entity in the discourse, it is often the case that s/he assumes that the hearer can infer the (discourse) existence of certain other entities, based on the speaker's beliefs about the hearers' beliefs and reasoning ability" (Prince, 1992) (page 304). Based on a general idea of "reasoning ability of speakers" I inferred that she would classify as inferrable all the examples in the table.

Fraurud's first mention class seems to apply in general; exceptions are, perhaps, those uses classified by Clark as set-membership. Löbner says that associative anaphora are semantic FC2s with anaphoric arguments. He considers, in particular, those descriptions which have a relational noun (use), and whose argument is specified by an antecedent. Some of the associations exemplified in the table seem to be based on other grounds, however; such cases were indicated by question marks. Clark's account points to several distinct relations. His set-membership relation is classified as direct reference; all other relations listed under Clark in Table 2.2 are classified as indirect reference. Sidner's cases of associated specification descriptions correspond clearly to those explicitly referred by Hawkins and Prince as associative anaphoric and inferrable. It is not clear how broad her class of associated specification was meant to be, but we considered it to be very general; her inferred specification rule applies for those cases based on hearers' suppositions. Clark's and Strand's classifications are the most detailed. Strand (1977) suggests that a causation link (a third subclass in the adjoining class) might be applied for the cases in which there is a relation of reason, cause and consequence. Strand does not classify optional roles (cases which involve hearers' suppositions) which are observed by Clark.

The situational uses are presented in Table 2.3. Hawkins and Löbner agree that some cases refer directly to the physical context (pragmatic deictic, visible and immediate situation) whereas others (semantic FC2s with deictic argument and larger situation)

[21] Hawkins does not explicitly refer to a description as being associated to a previous VP, although this is considered in the examples of anaphoric uses.

only relate to the context, in the sense that their interpretation involves context identification and reasoning. Some uses may rely either on specific or general knowledge: in the example of a wedding situation, the interpretation of a descriptions such as *the bride* may involve either specific knowledge of the referent or the general knowledge that weddings have brides. The same ambiguity is expressed in terms of situationally inferrable[22] or unused in Prince's taxonomy. Fraurud also reckons that first mention uses may require situational anchors or referents, although she does not name different classes for them.

In Table 2.4 the description *the colour red* is given as unfamiliar by Hawkins but would probably fit better in Prince's unused. All but one unfamiliar uses are semantic definites in Löbner's scheme. Löbner discriminates *the woman Bill went out with* from uses like *the fact that...* probably because in the first case the referent is just associated to another known entity, whereas in the latter the conceptual referent is determined by the complement (although both are based on complements which disambiguate a sortal head noun)[23].

Note that the type of descriptions that illustrates situational and unfamiliar uses are potentially discourse new. However, nothing prevents them from being used in subsequent mention. Hawkins and Löbner give the most complete classifications for discourse new descriptions.

2.6 Summary

Together, the theories we have discussed in this chapter account for both the anaphoric and non-anaphoric uses of definite descriptions.

For the anaphoric uses, we need to understand the ways in which a definite description may relate to its antecedent. We have presented studies which consider various kinds of relations between a description and its antecedent. One main distinction of the different types of relations is between co-referential and associated relations, and each of them may be realized in several distinct ways.

Anaphoric (co-referential)[24] relations may be direct (description and antecedent having the same head noun) or they may be expressed by equivalent nouns (synonyms), through generalization (hypernyms), and sometimes through specialisation (hyponyms). Also, a proper name may introduce an entity which is afterwards referred to by a description of the entity type. Some authors also consider that the antecedent for a definite description may be introduced by a VP. The first type (direct anaphora) is the easiest to be treated systematically; the other co-referential relations are based on common sense knowledge, a requirement which is also essential for the interpretation of the associative uses. One extra difficulty in dealing with descriptions interpreted through associated relations is that the discourse might provide various anchors/links (different but

[22] Prince does not use this term herself but we explained it in Section §2.2.3.

[23] Later in (Löbner, 1996) he presents *the book I gave you yesterday* as an FC1 but he does not say whether he is considering it semantic or pragmatic.

[24] There is a difference to be noted between anaphora and co-referentiality: an NP which is co-referent to another NP is not necessarily anaphoric. This is discussed in Section §2.3.2.

Anaphoric uses	Hawkins	Prince	Fraurud	Löbner	Clark	Sidner	Strand
a book/ the book	ANAPHORIC	TEXTUALLY EVOKED	SUBSEQUENT MENTION	PRAGMATIC (anaphoric)	IDENTITY	CO-SPECIFIC 1	COREF (ident. head)
a lathe/ the machine	ANAPHORIC	TEXTUALLY EVOKED	SUBSEQUENT MENTION	*pragmatic anaphoric*	PRONOMINA-LIZATION	CO-SPECIFIC 2 (generalizing)	COREF (generalization)
a car/ the sedan	ANAPHORIC	TEXTUALLY EVOKED	SUBSEQUENT MENTION	*pragmatic anaphoric*	?	?	COREF (specification)
a man/ the bastard	ANAPHORIC	TEXTUALLY EVOKED	SUBSEQUENT MENTION	*pragmatic anaphoric*	EPITHET	?	COREF (redescription)
he travelled/ the journey	ANAPHORIC	?	?	*pragmatic anaphoric*	IDENTITY	?	COREF (?)
a man a woman/ the couple	ANAPHORIC	TEXTUALLY EVOKED	? (summation)	*pragmatic anaphoric*	?	?	WIDENING (members-set)
Pinkerton Inc./ the company	ANAPHORIC	TEXTUALLY EVOKED	SUBSEQUENT MENTION	*pragmatic anaphoric*	*pronomina-lization*	*co-specific 2 generalizing*	COREF (redescription)

Table 2.1: Classifications of definite descriptions: anaphoric uses

Associative uses	Hawkins	Prince	Fraurud	Löbner	Clark	Sidner	Strand
ANTECEDENT	TRIGGER	TRIGGER	ANCHOR	ARGUMENT	ANTECEDENT	FOCUS	ANCHOR
a book/ the author	ASSOCIATIVE	INFERRABLE	FIRST MENTION	SEMANTIC FC2 (anaphoric arg.)	?	ASSOCIATED SPECIFIC.	DELIMIT. (subcateg.)
the room/ the ceiling	ASSOCIATIVE	INFERRABLE	FIRST MENTION	SEMANTIC FC2 (anaphoric arg.)	NECESSARY PARTS	ASSOCIATED SPECIFIC.	NARROWING (whole-part)
the wall/ the building	*associative*	*inferrable*	FIRST MENTION	?	?	*associated specific.*	WIDENING (part-whole)
the room/ the window	ASSOCIATIVE	INFERRABLE	FIRST MENTION	SEMANTIC FC2 (anaphoric arg.)	PROBABLE PARTS	ASSOCIATED SPECIFIC.	NARROWING (whole-part)
the room/ the chandelier	*associative*	*inferrable*	FIRST MENTION	?	INDUCIBLE PARTS	INFERRED SPECIFIC.	DELIMIT. (space anch.)
a couple/ the woman	*associative*	*inferrable*	?	?	SET-MEMBERSHIP	*associated specific.*	NARROWING (set-member)
clowns/ the clown with the unicycle	*associative*	*inferrable*	?	?	SET-MEMBERSHIP	SET-ELEM. SPECIFIC.	NARROWING (set-member)
she was killed/ the murderer	?	*inferrable*	FIRST MENTION	SEMANTIC FC2 (event role)	NECESSARY ROLES	*associated specific.*	NARROWING (event-arg.)
she died/ the murderer	?	*inferrable*	FIRST MENTION	?	OPTIONAL ROLES	INFERRED SPECIFIC.	?
a professor/ the car	?	*inferrable*	FIRST MENTION	?	?	INFERRED SPECIFIC.	ADJOINING (poss.-thing)
an earthquake/ the suffering of people	*associative*	*inferrable*	FIRST MENTION	?	REASON/CAUSE /CONSEQ.	INFERRED SPECIFIC.	ADJOINING (causation)
Israel and Egypt/ the peace agreement	*associative*	*inferrable*	FIRST MENTION	SEMANTIC FC3 (anaphoric arg.)	?	*associated specific.*	DELIMIT. (arg.-event)
last Wednesday / the news	*associative*	*inferrable*	FIRST MENTION	?	?	*associated specific.*	DELIMIT. (time anch.)
the first... the next... the last ...[a]	?	*inferrable*	?	?	SET-MEMBERSHIP	COMPUTED SPECIFIC.	ADJOINING (part-part)

Table 2.2: Classifications of definite descriptions: associative uses

[a] These descriptions are not considered to be complemented with full relatives.

Situational uses	Hawkins	Prince	Fraurud	Löbner
pass me the salt	VISIBLE SITUATION	SITUATIONALLY EVOKED	FIRST MENTION	PRAGMATIC (deictic)
beware of the dog	IMMEDIATE SITUATION	SITUATIONALLY EVOKED	FIRST MENTION	PRAGMATIC (deictic)
(at a wedding) the bride	LARGER SIT. (gen./sp. kn.)	*situationally infer./unused*	FIRST MENTION	SEMANTIC FC2 (deictic arg.)
the Prime Minister	LARGER SIT. (gen./sp. kn.)	*situationally infer./unused*	FIRST MENTION	SEMANTIC FC2 (deictic arg.)
the weather	LARGER SIT. (general kn.)	?	FIRST MENTION	SEMANTIC FC1 (simple NP)
the Gibbet	LARGER SIT. (specific kn.)	UNUSED	FIRST MENTION	SEMANTIC FC1 (proper name)

Table 2.3: Classifications of definite descriptions: situational uses

Unfamiliar uses	Hawkins	Prince	Fraurud	Löbner
the fact that ...	UNFAMILIAR (np compl.)	*containing inferrable*	FIRST MENTION	SEMANTIC FC1
the colour red	UNFAMILIAR (nom. modif.)	*unused*	FIRST MENTION	SEMANTIC FC1
the woman Bill went out with	UNFAMILIAR (rel. clause)	*containing inferrable*	FIRST MENTION	PRAG. ENDOPH. (with attribute)
the bottom of the sea	UNFAMILIAR (ass. clause)	CONTAINING INFERRABLE	FIRST MENTION	SEMANTIC FC2 (explicit arg.)
the same secrets	UNEXPLAN. MODIFIERS	?	FIRST MENTION	SEMANTIC FC1 (complex NP)
the first person to sail to ...	UNEXPLAN. MODIFIERS	*containing inferrable*	FIRST MENTION	SEMANTIC FC1 (complex NP)

Table 2.4: Classifications of definite descriptions: unfamiliar uses

equally suitable entities) for their interpretation. When associative relations need to be established we might have to face a difficult decision among several options. Strand's idea of deciding on a more informative relation is plausible but still difficult to implement, or even to define.

For other uses of definite descriptions the interpretation is not based on an antecedent given by the linguistic context of utterance (or discourse): a description may refer to an entity in the physical environment, or something of the speaker's common knowledge. Also, the complexity of the description's syntactic structure may provide complementary information to the interpretation of a definite description (within the description itself).

These theories serve as the background for the work discussed in the next chapters, where we present two related experiments involving:

- an empirical analysis of the uses of definite descriptions, aimed at further evaluating the relative importance of the different uses of descriptions; and
- the development and testing of a set of heuristics to process definite descriptions in written discourse.

In these experiments we adopted a simplification of these classifications, roughly according to their subdivision in the four tables just presented. The principal distinguishing factors we adopted to define the classes are:

- the existence of a co-referential antecedent (previous mention of an entity, as in Table 2.1);
- the presence of an associated antecedent (previous mention of an associated entity, as in Table 2.2);
- independence from previous discourse elements for the description interpretation (sometimes based on reader's previous knowledge of an entity or knowledge of the situation, as in Table 2.3; sometimes based on the reader's ability to infer an entity through the inherent complexity of a definite NP, as in Table 2.4).

Our choice of a simple classification scheme was ruled by our goals of accounting for definite descriptions in unrestricted texts, and we wanted to make the annotation uncomplicated for the subjects employed in the empirical analysis. Previous attempts to annotate detailed co-referential relations had resulted in very low agreement levels[25].

[25] In other co-reference annotation experiments, such as the ones in the MUC-6 (Sundheim, 1995), relations other than identity were dropped due to difficulties in annotating them.

Chapter 3

Corpus Study

Studies on the anaphoric (non-anaphoric) role of definite descriptions have usually been included in an analysis of definite noun phrases in general. Previous empirical analyses of definite description use (Prince, 1981; Fraurud, 1990; Prince, 1992) have been based purely on the author's interpretation. Other works on corpus annotation for anaphoric relations , such as (Chinchor and Sundheim, 1995) and (Mcenery, Tanaka and Botley, 1997), have also considered anaphoric relations in general with no specific attention to the case of definite descriptions.

Our work differs from those: we devote special attention to the use of definite descriptions in particular; and we present an analysis which is based not only on the author's interpretation, but also on other people's judgements. The reason for going towards an inter-subjective analysis is that replicability of judgements has become an issue in dialogue and discourse research in the areas of Computational Linguistics and Cognitive Science (Carletta, 1996):

> Now, researchers are beginning to require evidence that people besides the authors themselves can understand and make the judgements underlying the research reliably (page 249).

Replicability of human judgements (corpus annotation) has also been a requirement for recently proposed methods for computer systems evaluation—e.g., as done for the coreference task of the Sixth Message Understanding Conference[1] (MUC-6) (Sundheim, 1995).

We have undertaken two experiments testing agreement on anaphoric structure of definite descriptions. In the first one, the subjects were students of Linguistics, our instructions were explicit about the correlation between syntax and usage types of definite descriptions, and the annotation was only classificatory. The second experiment was a variation of the first one. As well as revising some features of the first experiment,

[1] In the MUC-6 competition, the replicability of the corpus annotation for the coreference task was measured simply in terms of recall and precision by taking one of two different annotations as the key. The measures we adopted here also take into account the chance agreement.

such as avoiding correlating syntax and types of use, adopting a semantic annotation and employing naive subjects, we also tested further aspects of the theories of definite description use and interpretation by adopting a more detailed classification scheme. One use of our experiments was to provide an empirical verification of the various theories of uses of definite descriptions presented in Chapter 2; secondly, it has produced the test data used in the computational side of the work discussed in Chapter 4. In this chapter we describe in detail our two experiments and present their reliability analysis regarding the inter-coder agreement. We also describe a corpus study specifically centred on bridging references.

The chapter is organized as follows. In Section §3.1, we describe the corpus, the annotation scheme we developed and the reliability metrics we used. The first experiment is presented in Section §3.2, and the second experiment in Section §3.3. A study dedicated specifically to bridging references is described in Section §3.4. In Section §3.5 we examine the text annotation results, and discuss their theoretical and methodological implications for the processing of definite descriptions.

3.1 Preliminaries

In this section we describe the corpus we studied, our classification scheme, and the reliability metrics adopted in our research.

3.1.1 Description of the corpus

The corpus used in our studies consists of 33 randomly[2] chosen articles from the Wall Street Journal contained in the subset of the Penn Treebank I Corpus included in the ACL/DCI CD-ROM.

The texts included in the first annotation exercise are: wsj_0203, wsj_0207, wsj_0209, wsj_0301, wsj_0305, wsj_0725, wsj_0760, wsj_0761, wsj_0765, wsj_0766, wsj_0767, wsj_0800, wsj_0803, wsj_0804 wsj_0808, wsj_0820, wsj_1108, wsj_1122, wsj_1124, and wsj_1137. The texts of the first experiment contain 1040 definite descriptions in total. The texts used in the second experiment are wsj_0766 (repeated from the first corpus), wsj_0003, wsj_0013, wsj_0015, wsj_0018, wsj_0020, wsj_0021, wsj_0022, wsj_0024, wsj_0026, wsj_0029, wsj_0034, wsj_0037, and wsj_0039, containing 464 definite descriptions.[3]

The study of bridging descriptions in Section §3.4 used the same texts used in the first experiment.

[2]The texts which were very short (less than 5 sentences) or which contained many numeric figures were not included in the selection.

[3]With respect to the size of our corpus, we were advised by Jean Carletta (personal communication) that in a classification experiment each main class should be represented by at least 20/30 items. This gave us an indication that, for our purposes, the size of our corpus was adequate. Our smallest class in the second experiment had 29 instances according to one of the annotators, more than that with the others. In the first experiment our main classes were all larger than that. Usually just a portion of the annotation exercise is used to measure reliability, in our studies we have considered the entire corpus.

3.1.2 Annotation schemes

Our corpus annotation involved a classification of uses of descriptions. Our choice of a classification scheme was ruled by our goals of accounting for definite descriptions in unrestricted texts, and we also wanted to make the annotation uncomplicated for the subjects. We used slightly different schemes in each experiment, that will be discussed separately.

3.1.3 Reliability metrics

In order to evaluate the results of a multi-coder experiment, it is necessary to have a way to measure the agreement among coders. The techniques we employed to measure the coders' agreement are presented next.

The Kappa statistic

The KAPPA STATISTIC (Siegel and Castellan, 1988), recently proposed by Carletta as a measure of agreement for discourse analysis (Carletta, 1996), is a test suitable for the cases when several subjects have to assign items to one of a set of classes.[4] The computation of the coefficient K of agreement among coders takes into account the possibility of chance agreement. K is dependent on the number of coders, the number of items being classified, and the number of choices of classes to be ascribed to items.

The K coefficient of agreement between annotators is defined as

$$K = \frac{P(A) - P(E)}{1 - P(E)} \tag{3.1}$$

where $P(A)$ is the proportion of times the annotators agree and $P(E)$ is the proportion of times that the annotators are expected to agree by chance. When there is complete agreement among the raters, $K = 1$; if there is no agreement other than that expected by chance, $K = 0$.

According to (Krippendorff, 1980), working in the field of content analysis where reliability has long been an issue, when the correlation between two variables is measured if the coefficient of agreement is less than 0.8 on one of the variables then strong relationships are likely to be missed even if they do exist.[5] For purposes like this, $K > 0.8$ is generally taken to indicate good reliability, whereas $0.68 \leq K < 0.8$ allows tentative

[4] It is the classification aspect of the annotation task which makes the use of the Kappa statistics appropriate to measure inter-coder agreement.

[5] For instance, a correlation between fish colour and length of life, requires each separate experiment to result in $K > 0.8$ in order to establish safely a relationship between the two, such as blue fishes live longer than yellow ones.

conclusions to be drawn.[6] As we are not correlating two variables in this theses, we cannot use these standards to interpret our results. As we believe the interpretation of K figures to be an open question, we interpret the figures resulting from our tests in a comparative way (by comparing better and worse agreements).

The method for computing K is illustrated here by an hypothetical annotation of definite description instances by three different coders.

Definite description	**C1**	**C2**	**C3**	**S**
1. the N	0	0	3	1
2. the O	0	2	1	0.33
3. the P	0	3	0	1
4. the Q	0	2	1	0.33
5. the R	3	0	0	1
6. the S	1	1	1	0
7. the T	0	0	3	1
8. the U	0	0	3	1
9. the V	0	2	1	0.33
10. the W	3	0	0	1
11. the X	3	0	0	1
12. the Y	3	0	0	1
13. the Z	3	0	0	1
$N = 13$	$C1 =$ 16	$C2 =$ 10	$C3 =$ 13	$Z =$ 10

Table 3.1: Computation of the K coefficient

The first column in Table 3.1 (**Definite description**) refers to the descriptions to be classified. The columns **C1**, **C2**, and **C3** stand for the classification options presented to the subjects. The numbers in these columns ($n_{i,j}$) indicate the number of classifiers that assigned the description in row i to the class in column j. The last row in the table shows the total number of descriptions (N), the total number of descriptions assigned to each class ($C1, C2, C3$) and, finally, the total agreement for all descriptions (Z). The numbers in the final column (labelled **S**) represent the percentage agreement for each definite description; Table 3.2 shows how this percentage agreement is calculated. The equations for computing $P(E)$, $P(A)$, and K are shown in Table 3.3.

In these formulas,

- S_i is the agreement for description i (S_1 and S_2 are shown as examples),
- m is the number of classes,
- C is the number of coders,
- N is the number of items being classified,

[6] Carletta (1997) finds it intriguing that other areas such as medical research have agreed on much lower levels (Landis and Koch, 1977).

$$S_i = \frac{1}{C(C-1)} * \sum_{j=1}^{m} n_{ij}(n_{ij} - 1)$$

$$S_1 = \frac{1}{3(2)} * [0 + 0 + 3(2)] = \frac{1}{6} * 6 = 1$$

$$S_2 = \frac{1}{6} * [0 + 2(1) + 1(0)] = \frac{1}{6} * 2 = 0.33$$

Table 3.2: Agreement on each item i (S_i)

- NC is the total number of assignments ($N * C$),
- $P(E)$ is the agreement expected by chance,
- $P(A)$ is the total agreement, and
- K is the coefficient of agreement.

$$Z = \sum_{i=1}^{N} S_i$$

$$P(A) = \frac{Z}{N} = \frac{10}{13} = 0.77$$

$$NC = 39$$

$$P(E) = (\frac{C1}{NC})^2 + (\frac{C2}{NC})^2 + (\frac{C3}{NC})^2 = (\frac{16}{39})^2 + (\frac{10}{39})^2 + (\frac{13}{39})^2$$

$$= 0.17 + 0.07 + 0.11 = 0.35$$

$$K = \frac{P(A)-P(E)}{1-P(E)} = \frac{(0.77-0.35)}{(1-0.35)} = \frac{0.42}{0.65}$$

$$= 0.65$$

Table 3.3: Computing the K coefficient of agreement

Confusion matrix

The confusion matrix is another method for comparing the results of multiple coders. The example of confusion matrix in Table 3.4 shows the agreement on classes between two different coders, A and B. Each matrix entry $n_{i,j}$ indicates the number of definite descriptions assigned to class i by one subject and to class j by another. For instance, entry $n_{1,1}$ shows that 5 items were assigned as **C1** by both coders. The confusion matrix

B A	C1	C2	C3	Total B
C1	**5**	1	0	6
C2	1	**3**	1	5
C3	0	0	**2**	2
Total A	6	4	3	13

Table 3.4: Confusion matrix

Class	Assignments	Comparisons	Agreem	Disag	% Agreem
C1	16	32	30	2	94%
C2	10	20	12	8	60%
C3	13	26	18	8	69%

Table 3.5: Per-class agreement

specifies the agreed distribution: whereas both coders have assigned 6 items to class **C1**, they have done it only 5 times for the same items. The table shows, for instance, that one item assigned by coder A as **C1** is assigned by B as **C2** (entry $n_{1,2}$).

Per-class agreement

The K coefficient gives a global measure of agreement. It is sometimes interesting to measure the agreement per class, i.e., to understand where annotators agreed the most and where they disagreed the most. The confusion matrix does this to some extent, but only works for two annotators.

The 'per-class agreement' is computed for each class separately by taking the proportion of pairwise agreements relative to the number of pairwise comparisons, as follows: whenever all three subjects ascribe a description to the same class, there are three assignments, 6 pairwise comparisons and 6 pairwise agreements for that class—100% agreement. If two subjects ascribe a description to **C1** and the other subject to **C2**, there are two assignments, four comparisons and two agreements for **C1**, which is 50% agreement; and one assignment, two comparisons and no agreement for **C2** (0% agreement).[7]

Table 3.5 shows the per-class agreement computed according to Table 3.1. There are, according to that table, 16 assignments, 32 pairwise comparisons and 30 agreements for **C1**, resulting in a percentage agreement of 94% for that class. The resulting agreements for each class in this measure are also regarded as comparable figures rather than an absolute measure of agreement. In this example, the class which presents least agreement among the annotators is **C2**.

[7] An equivalent technique for doing this is proposed by Krippendorff (1980): taking each class and eliminating items classified as such by any coder, then see which of the classes when eliminated causes K to increase most. This class is the one which introduces more disagreements. We found our method simpler to present.

Agreement on antecedents

The K coefficient and per-class agreement just presented only evaluate the agreement on classification of the uses of definite descriptions. In the second experiment, a way of assessing agreement on the identification of a discourse antecedent (for those classes in which the definite description interpretation was based on previous discourse) was also needed; to do this we considered the rate of agreement on the antecedents over agreement on the proper classes. Suppose, for instance that we have 100 cases classified as anaphoric by all three coders and that they had identified the same antecedent for only 90 of these 100 cases, then we have 90% agreement on the antecedent. In the next sections, where we describe the experiments, examples will illustrate situations in which the coders mark the same class but different antecedents.

3.2 First experiment

The goals of our first experiment in annotating definite description uses were:

- to evaluate the classification schemes discussed in Chapter 2;
- to observe the distribution of the different uses;
- to estimate the degree of difficulty involved in the processing of definite descriptions, by
  - figuring out the relative importance of anaphoric definite descriptions that are resolved with same head indefinite antecedents,
  - learning what else is necessary to process definite descriptions in written texts;
- to produce annotated texts to be used in our computational experiments.

3.2.1 Annotation Scheme

The classes we adopted are described in detail and exemplified in the following. The examples were extracted from the corpus.

I. Direct Anaphora This class includes the uses of definite descriptions which refer back to an antecedent introduced in the discourse. The descriptions in this class have the same descriptive content as their antecedents.[8]

(3.2) a. *a rig - the rig:* Grace Energy just two weeks ago hauled *a rig* here 500 miles from Caspar, Wyo., to drill the Bilbrey well, a 15,000-foot, $ 1-million-plus natural gas well. *The rig* was built around 1980, but has drilled only two wells, the last in 1982.

b. *the U.S. - the U.S.* Only 14,505 wells, including 4,900 dry holes, were drilled for oil and natural gas in *the U.S.* in the first nine months of the year, down 22.4% from the like 1988 period. But that was off less than at midyear, when completions lagged by 27.1%. And the number of rigs active in *the U.S.* is inching up.

This class does not include other cases of co-referential descriptions in which the association is based on more complex forms of lexical or common-sense knowledge, such as synonyms, hypernyms, information about events, etc. It differs from Hawkins' 'anaphoric use' or Prince's 'textually evoked' classes because it only includes definite-antecedent pairs with the same head noun. Quirk et al. (1985) and Löbner (1985) also refer to this class as direct anaphora.

II. Bridging This class contains both:

- definite descriptions that stand in an anaphoric (co-referent) relation with an antecedent explicitly mentioned in the text, but are not identified by the same predicate as their antecedent, and
- definite descriptions in an associative relation with an antecedent explicitly mentioned in the text, such as Hawkins' associative anaphoric descriptions and Prince's inferrables.

Examples are:

(3.3) a. *a stately Victorian home - the house:* Toni Johnson pulls a tape measure across the front of what was once *a stately Victorian home*. A deep trench now runs along its north wall, exposed when *the house* lurched two feet off its foundation during last week's earthquake.

[8] We use the term direct anaphora, as Fraurud (Section §2.3.2), for subsequent uses (with same head noun) in general. Subsequent uses of proper names, for instance *the U.S., the U.S.*, belong to this class, although they are not strictly anaphoric.

b. *Kadane Oil Co. - the company: Kadane Oil Co.*, a small Texas independent, is currently drilling two wells itself and putting money into three others. One of its wells, in southwestern Oklahoma, is a "rank wildcat", a risky well where oil previously hasn't been found. "At this price, $ 18 plus or minus, and with costs being significantly less than they were several years ago, the economics are pretty good", says George Kadane, head of *the company*.

c. *the 80-year-old house - the kitchen:* Once inside, she spends nearly four hours measuring and diagramming each room in *the 80-year-old house*, gathering enough information to estimate what it would cost to rebuild it. While she works inside, a tenant returns with several friends to collect furniture and clothing. One of the friends sweeps broken dishes and shattered glass from a countertop and starts to pack what can be salvaged from *the kitchen*.

d. *something has changed - the change:* With all this, even the most wary oil men agree *something has changed*. "It doesn't appear to be getting worse". "That in itself has got to cause people to feel a little more optimistic", says Glenn Cox, the president of Phillips Petroleum Co. Though modest, *the change* reaches beyond the oil patch, too.

Recognizing the antecedent of these definite descriptions involves at least knowledge of lexical associations, and possibly general common-sense knowledge.[9] We grouped them together in order to observe how frequent is the need for complex lexical inferences when resolving anaphoric definite descriptions, as opposed to simple head matching.

III. Discourse new These definite descriptions introduce a novel discourse referent not associated to some previously established object in the text—i.e., they are discourse new in Prince's sense. This class includes both definite descriptions that exploit situational information (Hawkins' larger situation uses, Prince's unused) and discourse new definite descriptions introduced together with their links or referents (Hawkins' unfamiliar uses). They were grouped in the same class because of claims by Prince and Fraurud that distinguishing the two classes is generally difficult.

(3.4) a. *the Securities and Exchange Commission:* Investors are appealing to *the Securities and Exchange Commission* not to limit their access to information about stock purchases and sales by corporate insiders.

b. *the third quarter:* Norton Co. said net income for *the third quarter* fell 6 % to $ 20.6 million, or 98 cents a share, from $ 22 million, or $ 1.03 a share.

c. *the government:* Also, as former Reagan antitrust chief Charles Rule has noted, this would "establish the precedent that *the government* may charge parties for the privilege of being sued regardless of whether the government prevails".

[9] See (Löbner, 1985; Barker, 1991; Poesio, 1994) for discussions of lexical conditions on bridging references.

d. *the Iran-Iraq war:* About the same time, *the Iran-Iraq war*, which was roiling oil markets, ended.

e. *the economic know-how to steer the city through a possible fiscal crisis:* They wonder whether he has *the economic know-how to steer the city through a possible fiscal crisis*, and they wonder who will be advising him.

f. *The appetite for oil-service stocks: The appetite for oil-service stocks* has been especially strong , although some got hit yesterday when Shearson Lehman Hutton cut its short-term investment ratings on them.

g. *the fact that few local non-Jewish politicians have been as vocal for Jewish causes in the past 20 years as Mr. Dinkins has:* Mr. Dinkins also has failed to allay Jewish voters' fears about his association with the Rev. Jesse Jackson, despite *the fact that few local non-Jewish politicians have been as vocal for Jewish causes in the past 20 years as Mr. Dinkins has*.

h. *the first raise he can remember in eight years:* Mr. Ramirez, who arrived late at the Sharpshooter with his crew because he had started early in the morning setting up tanks at another site, just got *the first raise he can remember in eight years*, to $ 8.50 an hour from $ 8.

i. *Rudolph Giuliani, the former crime buster:*[10] After his decisive primary victory over Mayor Edward I. Koch in September, Mr. Dinkins coasted, until recently, on a quite comfortable lead over his Republican opponent, Rudolph Giuliani, *the former crime buster* who has proved a something of a bust as a candidate.

j. *The man most likely to gain custody of all this: The man most likely to gain custody of all this* is a career politician named David Dinkins.

IV. Idiom This class includes idiomatic expressions and metaphorical uses.

(3.5) *the soup:* A recession or new OPEC blowup could put oil markets right back in *the soup*.

V. Doubt The subjects could also express 'doubt' about the classification of the definite description.

We did not have a class for immediate situation uses (deictic descriptions), since it was assumed that they would be rare in written text.[11]

[10]Cases of appositive constructions like this one were considered as complex expressions containing a definite description, and the whole expression was then considered as discourse new. In the Treebank they are represented as an NP consisting of two NPs. Alternatively, definite descriptions occurring in such a structure could be regarded as coreferent to the proper name that appears first and therefore considered as discourse old.

[11]This was indeed the case. However, a few instances of an interesting kind of immediate situation use were observed. In these cases, the text is describing the immediate situation in which the writer is, and the writer apparently expects the reader to reconstruct this situation:

3.2.2 Methods

Subjects

The corpus was analysed by the author and two other subjects. The two subjects were English native speakers, graduate students in Linguistics. Henceforth, annotators A and B.

Materials

A collection of 20 articles containing 1040 definite descriptions was first classified by the author. Next, two subjects were asked to perform the same task. They had to assign each definite description to one class. The classes are as described in Section §3.2.1, but the terminology presented there differs from the nomenclature first adopted in the instructions. Discourse new descriptions were referred to in the instructions as larger situation and unfamiliar uses; bridging as associative; and direct anaphora as same head anaphora. The subjects could also express 'doubt' about the classification of the definite description. Some of the classes were introduced to the subjects by making explicit reference to their probable syntactic structure, following Hawkins' style (Hawkins, 1978). The instructions followed by the subjects are given in Appendix A.

Since the classification of definite descriptions is sometimes ambiguous, the subjects were instructed to resolve conflicts according to a preference ranking, i.e., to choose a class with higher preference when two classes seemed equally applicable. The ranking was (from most preferred to least preferred): 1) direct anaphora, 2) discourse new, and 3) bridging. The coders used a computer interface in which they indicated a class for each description. The annotators were given one text to familiarise themselves with the task before starting with the annotation properly. They took on average 12 hours to complete the whole task.

3.2.3 Results

The results of the author's analysis are summarized in Table 3.6. The results of annotators A and B are shown in Table 3.7. As the tables indicates, all annotators assign approximately the same percentage of definite descriptions to each of the five classes; however, the classes do not always include the same elements. This can be gathered by the confusion matrix in Table 3.8, where an entry $n_{i,j}$ indicates the number of definite descriptions assigned to class i by subject A and to class j by subject B. Considering the only the cases for which there are agreement between A and B, the distribution of the main classes corresponds to the following: 26% of the cases were agreed to be direct anaphora; 9% of the cases were agreed as being bridging descriptions; 45% of the cases were agreed to be discourse new (with a total of 20% disagreement).

(3.6) "And you didn't want me to buy earthquake insurance", says Mrs. Hammack, reaching across *the table* and gently tapping his hand.

(3.7) "I will sit down and talk some of the problems out, but take on the political system ? Uh-uh", he says with a shake of *the head*.

Class	# (author)	% (author)
I. Direct anaphora	304	29.23%
II. Bridging	193	18.55%
III. Discourse new	503	48.37%
IV. Idiom	26	2.50%
V. Doubt	14	1.35%
Total	1040	100

Table 3.6: Author's classification of definite descriptions in Experiment 1

Class	# (A)	% (A)	# (B)	% (B)
I. Direct anaphora	294	28.27%	332	31.92%
II. Bridging	160	15.38%	150	14.42%
III. Discourse new	546	52%	549	52.78%
IV. Idiom	39	3.75%	2	0.19%
V. Doubt	1	0.09%	7	0.67%
Total	1040	100%	1040	100%

Table 3.7: Coders' classification of definite descriptions in Experiment 1

The Kappa statistic was used to measure the agreement in a more precise way. The overall coefficient of agreement between the three annotations (A's, B's and the author's) on classes I-IV is $K = 0.69$ (for 1032 descriptions)[12]; $K = 0.72$ on classes I-III (992 descriptions), that is, when those descriptions marked as idioms are ignored.

The per-class agreement measure was also computed. The rates of agreement for each class thus obtained are presented in Table 3.9.

[12]Doubts were not considered in the computation of agreement.

B **A**	**I.**	**II.**	**III.**	**IV.**	**V.**	**Total B**
I. Direct anaphora	**274**	26	32	0	0	332
II. Bridging	9	**97**	44	0	0	150
III. Discourse new	8	37	**465**	38	1	549
IV. Idiom	0	0	1	**1**	0	2
V. Doubt	3	0	4	0	**0**	7
Total A	294	160	546	39	1	1040

Table 3.8: Confusion matrix of coders' classification

Class	Total	Comparisons	Ag	Disag	% Ag
I. Direct anaphora	930	1860	1646	214	88%
II. Bridging	503	1006	596	410	59%
III. Discourse new	1598	3196	2684	512	84%
IV. Idiom	67	134	42	92	31%
V. Doubt	22	44	2	42	4%

Table 3.9: Per-class agreement in Experiment 1

3.2.4 Discussion

Distribution of uses

One of the most interesting results of this first experiment is that a large proportion of the definite descriptions in the corpus are not related to an antecedent previously introduced in the text.[13] Surprising as it may seem, this finding is in fact just a confirmation of the results of other researchers. Fraurud (1990) reports that 60.9% of definite descriptions in her corpus of 11 Swedish texts are 'first-mention', i.e., do not co-specify with an entity already evoked in the text[14]; Gallaway (1996) found a distribution similar to this in (English) spoken child language. These findings give support to Löbners claim that familiarity is not the basis for definiteness (Section §2.3.1).

Agreement among annotators

The second notable result was the relatively low agreement among annotators. The reason for this disagreement was not so much annotators' errors as the fact, already mentioned, that the classes are not mutually exclusive. The confusion matrix in Table 3.8 indicates that the major classes of disagreements were definite descriptions classified by annotator A as discourse new and by annotator B as bridging, and vice versa. One such example is *the country* in (3.8); this definite description could be classified as discourse new (larger situation) because it refers to the country of the newspaper's publication; but it could also be classified as being bridging on *the U.S.* [15]

[13] I recall here Prince's hypothesis that containing inferrables (probably most of the discourse new descriptions) are suitable for multi-receiver discourse, in particular, formal written prose, where the sender either is not sure of the receivers' knowledge or where s/he believes that there are relevant differences among the receivers.

[14] Note that there is a difference between our classification scheme and Fraurud's study: she does not distinguish among first mention definites those whose interpretation is based on associated antecedents (Hawkins' associative anaphora, Prince's inferrables) from those whose interpretation is independent from previous discourse (Hawkins' larger situation and unfamiliar uses, Prince's unused, containing inferrables or brand new anchored).

[15] As discussed above, this problem with Hawkins' classification scheme (ambiguity between larger situation and associative uses, as well as between Prince's unused and inferrable) had already been noted by Fraurud—e.g., (Fraurud, 1990), page 416.

(3.8) The missing watch is emblematic of the problems Mr. Wathen encountered in building his closely held California Plant Protection Security Service into the largest detective and security agency in *the U.S.* through acquisitions.
... (other 5 sentences) ...
Over the next 20 years, California Plant Protection opened 125 offices around *the country*.

The figures in Table 3.9 indicate a better agreement on direct anaphora and discourse new definite descriptions, much worse agreement on the other classes. (The percentages for idioms and doubts are very low; but these classes are also too small to allow us to draw any conclusions.)

Problems with the first experiment

The data we collected in this experiment was used to develop our prototypes, and we needed to collect some test data. Instead of repeating the same experiment we revised it. As a result, the classification scheme and the annotation instructions were changed. Some problematic aspects we observed in this first experiment were:

- the subjects had background knowledge of Linguistics;
- we explicitly referred to the correlation between syntactic structure and types of uses of descriptions in the instructions, so that we could have modelled too strictly the subjects' response;
- there was a preference ranking for ambiguous cases, which could have influenced the relative importance of each class;
- the annotation lacked the semantic side of the interpretation, that is, the subjects did not have to identify the antecedent for the anaphoric and bridging cases.

3.3 Second experiment

The design of the second experiment included several changes whose goals were:

- to understand whether the classification disagreements in the first experiment reflected disagreements on the identification of antecedents;
- to verify whether the distribution and agreement of the first annotation exercise was a result of the preference ranking among classes: to test this, the subjects were not given an explicit preference ranking in the second experiment, just a set of questions in the format of decision tree was offered to help the coder in the performance of the task[16];

[16] By analysing the coders' responses I noted that they did not follow the instructions strictly.

- to study more carefully the distribution of types of definite descriptions, in particular,
  - assess the relative important of co-referent descriptions with different descriptive content from their antecedents, and
  - assess the distribution of definite descriptions in the discourse new class into two distinct classes (larger situation and unfamiliar);
- and finally, to ask non-linguistically trained subjects to perform the classification task.

3.3.1 Revised annotation scheme

The direct anaphora class of the first experiment was replaced with a broader CO-REFERENT class including all cases in which a definite description is co-referent with its antecedent, whether or not the head noun is the same. The subjects were asked to classify as co-referent a definite like *the house* referring back to an antecedent introduced as *a Victorian home*, which would not have counted as direct anaphora in the first experiment. This resulted in a taxonomy which was at the same time more semantically oriented and closer to Hawkins' and Prince's classification schemes: this broadened co-referent class coincides with Hawkins' 'anaphoric' and Prince's 'textually evoked' classes, whereas the resulting, narrower bridging class (called now associative) coincides with Hawkins' 'associative anaphoric' and Prince's inferrables.[17] The intention was to see whether the distinctions proposed by Hawkins and Prince led to a better agreement among annotators than the taxonomy used in the first experiment, i.e., whether the subjects would be more in agreement about the semantic relation between a definite description and its antecedent than they were about the relation between the head noun of the definite description and the head noun of its antecedent. By doing this, we could also observe the distribution of co-referent descriptions based on antecedents with different descriptive content. In order to get an idea of the extent of agreement among annotators about the semantic interpretation of definite descriptions, the subjects were asked to indicate the antecedent in the text for the definite descriptions they classified as co-referent or associative.

Another change in the taxonomy was to split the discourse new class in the first experiment in two classes, as in Hawkins' and Prince's schemes. This was done to see whether indeed these two classes were difficult to distinguish (as suggested by Fraurud); and also to get a clearer idea of the relative importance of the two kinds of definites grouped together in the first annotation. The two classes were called LARGER-SITUATION (based on common knowledge) and UNFAMILIAR (based on the internal structure of the description). Idiom and doubt were not given as an optional category; instead, the coders were instructed to write down a comment if they had any difficulty in classifying a description.

The modified taxonomy, in summary, is as follows:

[17] These distinct relations are also referred to as REITERATION and COLLOCATION by (Halliday and Hasan, 1976).

I. **Co-referent**—includes all descriptions which co-refer by any means with previous discourse.

II. **Associative**—descriptions which stand in an associated relation (i.e.,non co-referent) with previous discourse.

III. **Larger situation**—there is no textual anchor participating in the interpretation of the description, the receiver is likely to have the required knowledge for its interpretation.

IV. **Unfamiliar**—the description is new to the receiver, generally speaking, the description's interpretation is anchored on information contained in the description itself.

3.3.2 Methods

Subjects

Three subjects were used for Experiment 2. The subjects were English native speakers, graduate students of Mathematics, Geography and Mechanical Engineering at the University of Edinburgh. They will be referred as C, D, and E below.

Materials

The subjects were asked to annotate 14 randomly selected Wall Street Journal articles, all but one of them different from those used in the first experiment, and containing 464 definite descriptions in total.
Unlike the first experiment, there was no suggestion of a relation between the classes and the syntactic form of the definite descriptions in the instructions[18]. The subjects were asked to indicate whether the entity referred to by a definite description

1. had been mentioned previously in the text,
2. was new but related to an entity already mentioned in the text,
3. was new but presumably known to the average reader, or
4. was new in the text and presumably new to the average reader.

When the description was indicated as old (1) or related to some other entity (2), the subjects were asked to locate the previous mention of the related entity in the text. Unlike the first experiment, the subjects did not have the option to classify a definite description as 'Idiom'; they were instructed to make a choice whenever possible and write down their doubts. Also, "doubt" was not given as an option; we did this to avoid an overestimation of the number of doubts in the presence of minor difficulties (which we thought could happen with naive coders). The written instructions and the script given to the subjects can be found in Appendix B. As in the first experiment, the subjects were given one text to practice before starting with the analysis of the corpus. They took in average 8 hours each to complete the whole task plus half an hour of training.

[18]The reader is invited to compare the instructions for the two experiments given in Appendix A and Appendix B.

Class	#(C)	%(C)	#(D)	%(D)	#(E)	%(E)
I. Co-referent	205	44%	211	45%	201	43%
II. Associative	40	8.5%	29	6%	49	11%
III. Larger situation	119	25.5%	115	25%	93	20%
IV. Unfamiliar	92	20%	82	18%	121	26%
V. Doubt	8	2%	27	6%	0	0%
Total	464	100%	464	100%	464	100%

Table 3.10: Coders' classification of definite descriptions in Experiment 2

Class	Total	Comparisons	Ag	Disag	% Ag
I. Co-referential	617	1234	1066	168	86%
II. Associative	118	236	74	162	31%
III. Larger situation	327	654	466	188	71%
IV. Unfamiliar	295	590	380	210	64%
Doubt	35	70	2	68	3%

Table 3.11: Per-class agreement in Experiment 2.

3.3.3 Results

The distribution of definite descriptions in the four classes (and indication of doubt or ambiguity) according to the three coders are shown in Table 3.10.

There were 283 cases of complete agreement among annotators on the classification (61%): 164 cases of complete agreement on co-referential definite descriptions, 7 cases of complete agreement on associative ones, 65 cases of complete agreement on larger situation cases, and 47 cases of complete agreement on the unfamiliar class.

As in the first experiment, the coefficient of agreement among annotators, K, was calculated; the result for annotators C, D and E, 430 descriptions (the 34 cases which were marked at least once as doubt were left out), and the four classes I-IV is $K = 0.63$.

The extent of agreement among subjects on the antecedents for co-referential and associative definite descriptions was also measured. A total of 164 descriptions were classified as co-referent by all three coders; of these, 155 (95%) were taken by all coders to refer to the same entity (although not necessarily to the same mention of that entity). Counting, instead, the cases in which at least two annotators assigned a definite description to the co-referential class, and not just the cases in which all three agreed, the result is 510 agreements out of a total of 537 cases, again, a percentage of 95%.

There were only 7 definite descriptions classified by all three annotators as associative; in 5 of these cases (71%) the three annotators also agreed on a textual anchor (i.e., on the discourse entity to which the associative reference was related to).

The rates of agreement for each class are presented in Table 3.11.

3.3.4 Discussion

Distribution of uses

The distribution of definite descriptions among discourse new, on the one side, and co-referential with associative references, on the other, was roughly the same in the second experiment as in the first experiment, and roughly the same among the annotators. The average percentage of discourse new descriptions (larger situation and unfamiliar together) was 46%, against an average of 50% in the first experiment. Having split this class in two in this experiment, an indication of the relative importance of each class may be inferred. It is approximately 50% for each class (note that the first two annotators classified the majority of these as larger situation, whereas the last annotator classified the majority as unfamiliar).

As expected, the broader definition of the co-referent class resulted in a larger percentage of definite descriptions being included in this class compared to the class direct anaphora, an average of 44% against previous 30%, and a smaller being included in the associative class compared to the bridging class of the first experiment (9% against 16%).[19]

Agreement among annotators

The agreement among annotators in the second experiment ($K = 0.63$) was worse than the one obtained in the first one ($K = 0.69$, for classes I-IV or $K = 0.72$ for three classes, excluding idioms). We hypothesised that the reason for the worse agreement could be because we had split one class into two (discourse new into larger situation and unfamiliar); indeed by merging back these classes into one, as in the first experiment, the result went up from $K = 0.63$ to $K = 0.68$. This distance gives an idea of the difficulty in distinguishing between larger situation and unfamiliar definite descriptions.

It could also be the case that the texts used in the second experiment were more 'difficult'[20] than those used in first experiment: the results for one text included in both corpora were $K = 0.64$ (or $K = 0.73$ excluding idioms) in the first experiment and $K = 0.64$ for the second experiment, but $K = 0.75$ when merging larger situation and unfamiliar cases into a single class. Also, the coefficient of agreement changes dramatically from text to text: in this second experiment, it varies from $K = 0.42$ to $K = 0.92$ depending on the text, and dismissing the worse 3 texts from the corpus in the second experiment, the measure is $K = 0.73$ (for 3 categories). Looking at the coders' annotation it was noted that cases of premodification were frequent in cases of disagreements. We then calculated agreement only on those definite descriptions with no premodification, and the result was $K = 0.74$ for a total of 243 descriptions (for 3 categories).

We reanalysed the results grouping definite descriptions into fewer classes, i.e. just two, to see if we could get a better agreement. First, the binary division suggested by

[19] Hawkins (1978) refers to associative anaphoric uses as the most frequent use of the definite article. This claim, however is not supported by our studies. I would say that they represent instead the most complex use of the definite article.

[20] Considering that some instances fit the distinctions better than others, as pointed out by Fraurud.

Fraurud was tried: all co-referent definite descriptions on one side (subsequent mention), and all other definite descriptions in the other (Fraurud's first mention). Splitting things this way did result in an agreement of $K = 0.76$, i.e., within the 'tentative' margins of agreement ($0.68 \leq x < 0.8$), although not quite a strong agreement. The alternative of putting in one class all 'discourse-related' definite descriptions—co-referent and associative—and putting larger situation and unfamiliar definite descriptions in a second class resulted in an agreement of $K = 0.73$. In contrast, drawing a distinction between associative definites, on the one hand, and all other definite descriptions, on the other, resulted in a very low agreement: $K = 0.24$.

This indicates that although the subjects did not do very well at distinguishing first mention from subsequent mention entities, they were much better at that than they were at drawing more complex distinctions. Even clearer is that the worst case was distinguishing associative references from other definite descriptions.

Table 3.12 summarises the figures just presented.

Classes	K
I/II/III/IV (430 dds)	0.63
I/II/(III-IV) (430 dds)	0.68
I/(II-III-IV) (430 dds)	0.76
(I-II)/(III-IV) (430 dds)	0.73
II/(I-III-IV) (430 dds)	0.24
I/II/(III-IV) (243 dds) (with no premodifiers)	0.74

Table 3.12: Summary of the Kappa tests

Similar results were obtained by computing the 'per-class' percentage of agreement. The rates of agreement for each class thus obtained are presented in Table 3.11. There is a better agreement on co-referential definite descriptions, worse on associative references; the percentage agreement on the classes larger situation and unfamiliar taken individually is much lower than the agreement on the class discourse new taken as a whole (84%) in the first experiment. Considering Fraurud's observation that some definite descriptions "fit well" with Hawkins' types, the percentage of agreement for each class uses gives us an idea of the magnitude of the ambiguity problem in the taxonomy.

The results in Table 3.11 confirm the indications obtained by computing agreement for a smaller number of classes: our subjects agree much better on co-referent definite descriptions than on associative ones. The cases of disagreement are discussed in more detail next.

Yet another possible source of disagreement is the fact that whereas a 'syntactic' (same head) notion of what counts as co-referential was used in the first experiment, in the second experiment we used a 'semantic' one. Going from a 'syntactic' to a 'semantic' definition of anaphoric definite description resulted in worse agreement both for co-referential and for associative references: looking at the per-class figures, it was noticed that the per-class agreement on direct anaphoric and co-referential definite descriptions

went down from 88% in the first experiment to 86% in the second one, while the agreement for bridging and associative definite descriptions went down rather dramatically from 59% to 31%. Another difference is that in the first experiment our coders were Linguistics students. However, there were very few examples of true mistakes in the annotation, as discussed below; therefore, the choice of naive annotators does not seem to have contributed for a worse agreement.

Because the experiments were so long (12 and 8 hours of work distributed over 2 weeks), we looked at the agreement on judgement made earlier and later, dividing the corpus in two halves, to see if there was an effect. The agreement on classes I-III in the first experiment for the first half of the corpus is $K = 0.74$, for the second half $K = 0.68$. One possible interpretation for this difference is that the subjects have found it difficult to remember the distinctions, so soon after they read them they did better and later began to drift towards idiosyncratic interpretations of the categories. Alternatively, they may have paid less attention to the task as they got tired.

The agreement on three classes in the second experiment for the first half of the corpus is $K = 0.68$, and $K = 0.68$ for the second half. There is no drift in experiment 2. This suggests that naive coders could get a good handle on the task quickly and keep their understanding as they worked.

Analysis of classification disagreements

There are two basic kinds of disagreements among annotators: about classification, and about the identification of an antecedent.

There were 29 cases of complete disagreement among annotators with respect to the classification, i.e., cases in which no two annotators classified a definite description in the same way, and 144 cases of partial disagreement. All four of the possible combinations of total disagreement were observed, but the two most common combinations were associative/co-referential/unfamiliar and associative/larger situation/unfamiliar; all six combinations of partial disagreements were also observed. We will just discuss the cases most interesting from the perspective of designing a corpus annotation scheme.

There were very few true mistakes. On the whole, most of the disagreements were due to genuine problems in assigning a unique classification to definite descriptions.

Often the mistakes were of the form exemplified by (3.9). In this case, all three annotators indicate the same antecedent (*the potential payoff*) for the definite description *the rewards*, but whereas two of them classify *the rewards* as co-referential, one of them classifies it as associative. What seems to be happening here and in similar cases is that even though the subjects were asked to classify co-referentiality they ended up using a notion of bridging as defined in the first experiment (co-referent or associated with different descriptive content). There were 10 such cases of partial disagreement between associative and co-referential in which all three subjects indicated the same antecedent for the definite description.

(3.9) New England Electric System bowed out of the bidding for Public Service Co. of New Hampshire, saying that the risks were too high and *the potential payoff* too far in the future to justify a higher offer.

. . .

"When we evaluated raising our bid, the risks seemed substantial and persistent over the next five years, and *the rewards* seemed a long way out".

A particularly interesting version of this problem appears in the following example, when two annotators took the verb *to refund* as antecedent of the definite description *the refund*, but one of them interpreted the definite as co-referential with the eventuality, the other as associative.

(3.10) Commonwealth Edison Co. was ordered *to refund* about $250 million to its current and former ratepayers for illegal rates collected for cost overruns on a nuclear power plant.

The refund was about $55 million more than previously ordered by the Illinois Commerce Commission and trade groups said it may be the largest ever required of a state or local utility.

It is interesting to note that the theories presented in Chapter 2 are not uniform with respect to events as antecedents. Hawkins and Clark are the only ones to explicitly consider event as antecedents for anaphoric descriptions; Sidner considers events as antecedents for associated specification but not for co-specification. The remaining authors do not refer directly to the problem.

As could be expected by the discussion of K measures above, the most common disagreements (35 cases of partial disagreement out of 144, 24%) were between the classes larger situation and unfamiliar. One typical source of disagreement was the 'introductory' use of definite descriptions, common in newspapers: thus, for example, some of the annotators would classify *the Illinois Commerce Commission* as larger situation, other as unfamiliar. In many cases in which this form of ambiguity was encountered, the definite description worked effectively as a proper name: *the world-wide supercomputer law*, *the new US trade law*, or *the face of personal computing*. Certain proper names, specially those used with the definite article and including a common noun in them, seem to have this ability to refer to an entity even if it is not already known to the reader, as long as its existence is easily inferred and can be added to the reader's model of the discourse.

Rather surprisingly, from a semantic perspective, the second most common form of disagreement was between the co-referential and associative classes. In this case, the problem typically was that different subjects would choose different antecedents for a certain definite description. In example (3.10), the third annotator indicated *$250 million* as the antecedent for *the refund*, and classified the definite description as co-referential. An example of complete disagreement is the following:

(3.11) Mr. Rapanelli recently has said *the government of President Carlos Menem, who took office July 8,* feels a significant reduction of principal and interest is the only way the debt problem may be solved.

In this case, all three interpretations are acceptable: we may take the definite description *the government of President Carlos Menem, who took office July 8,* either as a case of associative reference on the previously mentioned *Argentina*, or as a case of larger situation, or as a case of unfamiliar definite description, especially if we assume that this latter class coincides with Prince's containing inferrables.

It seems in the end that the classification disagreements among annotators depend to a large extent on the task they are asked to do, rather than reflecting true differences in semantic intuitions.

Analysis of antecedent disagreements

There were also cases of disagreement about the antecedent of a definite description. The most common cases of antecedent disagreement were, however, those in which a disagreement between co-referential and associative classes also occurred, as seen in example (3.12) for the description *the same neighbourhood*. Coder C marked it as associative on *the collapsed section of double-decker highway Interstate 880* whereas coder E marked it as co-referent with *Oakland*:

(3.12) When Aetna adjuster Bill Schaeffer visited a retired couple in *Oakland* last Thursday, he found them living in a mobile home parked in front of their yard.

The house itself , located about 50 yards from *the collapsed section of double-decker highway Interstate 880*, was pushed about four feet off its foundation and then collapsed into its basement.

The next day , Mr. Schaeffer presented the couple with a check for $ 151,000 to help them build a new home in *the same neighbourhood*.

The problem of multiple anchors, as exemplified here, seems to indicate that a choice towards a "more informative" anchor (as Strand suggests) compete with notions such as saliency, availability, or closeness of the antecedent.

Further comments

Another reason for the low agreement in the experiments might be the fact that the annotators were not intensively trained and the task definition was very broad. It may be possible to achieve better results if well trained annotators are employed and the task definition and instructions are refined.

3.4 A corpus study of bridging references

As seen in the last chapter, Section §2.4, descriptions relate to their antecedent in many different ways. In the annotation exercises described earlier in this chapter, however, the bridging and associative classes were considered with no further subcategorization. We did a study to further examine the bridging class[21]. In this case we did not conduct a detailed multi-coder annotation, marking all different forms of linking relations; instead, we will present a preliminary analysis undertaken by the author and collaborators (Vieira and Teufel, 1997; Walde, 1997). The 204 cases of bridging descriptions identified in a compilation of the three annotations of the first experiment (the standard annotation as described in Section §5.1.2) were thus subclassified.

3.4.1 Types of bridging descriptions

Six classes were identified, adopting a classification which differs from all those discussed in Chapter 2. The classes discussed there are motivated by semantic distinctions. Here we were primarily motivated by differences in processing requirements.

Synonymy/Hyponymy/Meronymy This class (henceforth, Syn/Hyp/Mer) includes those definite descriptions which are in a synonymy, hyponymy or meronymy relation with their anchors, i.e., the kind of semantic relation that is currently encoded in WordNet (Miller et al., 1993), a public available lexicographic database (used as an approximation of a knowledge base by our implementation described in Section §4.5.1). Examples are:

Synonymy

(3.13) a. *new album... the record*;
b. *three bills... the legislation*.

Hyponymy/hypernymy

(3.14) a. *rice... the plant*;
b. *the daily television show... the program*.

[21] We will examine the bridging class as defined for the first experiment, since we are interested in investigating all those relations which are not based on a same head antecedent: associative as well as co-referent relations; their use imposes extra difficulties on the processing of definite descriptions.

Meronymy/holonymy (part of and has parts relations)

(3.15) a. *plants... the pollen*;
b. *house... the chimney*.

Some of these classes are related to Clark's direct reference (synonymy = identity, hypernymy = pronominalization) and Strand's coreferent descriptions (hypernymy = generalization, hyponymy = specification). The others corresponding to meronymy relations are referred by Clark as indirect reference (necessary and probable parts) and Strand's narrowing (whole-part) and widening (part-whole).

Names This class includes definite descriptions that refer back to proper names such as people's and company's names, as in:

(3.16) a. *Bach... the composer*;
b. *Pinkerton's Inc... the company*.

Cases such as these have not been explicitly mentioned in the literature covered in the last chapter; however, they clearly correspond to Fraurud's subsequent mention, Prince's evoked, Hawkins' anaphoric use, Strand's co-referentiality, etc. The automatic recognition of co-reference of such named entities requires different methods from those used for other (indirect) co-referent cases.

Compound Nouns This class includes bridging descriptions whose LINGUISTIC ANCHOR (i.e., the element in the text to which they are related) is a noun occurring as part of a compound noun other than the head. Examples include:

(3.17) a. *stock market crash... the markets*;
b. *discount packages... the discounts*.

This class has not been specifically observed by those authors mentioned previously.

Events These are cases where the linguistic anchors of definite descriptions are not NPs but VPs or sentences. Examples are:

(3.18) a. Individual investors and professional money managers *contend*. They make *the argument ...*;
b. Kadane Oil Co. *is currently drilling two wells and putting money into three others*. *The activity*...

Cases like these have been used as examples by Hawkins, Strand and Clark. Hawkins exemplifies the anaphoric use with sequences like *He travelled... The journey*. Clark categorizes some of these cases as indirect reference by necessary roles and optional roles. Strand accounts for event-argument relations.

Discourse Topic There are some cases of definite descriptions which are related to the (often implicit) discourse topic (in the sense of (Reinhart, 1981)) of a text, rather than to some specific NP or VP. For instance,

(3.19) a. *the industry* (in a text whose discourse topic is oil companies);
b. *the first half* (in a text whose discourse topic is a concert).

These cases are not discussed by the authors we reviewed.

Inference We collect in this class all the cases of bridging descriptions whose relation with their NP anchor is based on more complex inferential relations: for example, cases in which the relation between the anchor and the description is reason, cause, consequence, or set-membership:

(3.20) a. *last week's earthquake... the suffering people are going through*;
b. *Democratics, Republicans... the two sides*.

This class works as a waste basket: everything that involves more complex reasoning and does not fit one of the previous classes is included here. Thus, this class applies to cases that Prince defines as inferrable and Sidner as inferred specification. It also includes Clark's inducible parts, set-membership, epithets, relations of reasons, causes and consequences and Strand's part-part, time-anchoring, argument-event, possessor-thing, set-member, space-anchored and subcategorization.

Class	Total	%
Syn/Hyp/Mer	12/14/12	19%
Names	49	24%
Compound Nouns	25	12%
Events	40	20%
Discourse Topic	15	7%
Inference	37	18%
Total	204	100%

Table 3.13: Distribution of bridging references

The last three classes represent the cases for which a computational treatment cannot avoid being knowledge intensive. All of them require common sense reasoning: one requires VPs to be taken into account in the selection of possible anchors, another requires the discourse topic (aboutness) to be traced.

The relative importance of these classes in our corpus is shown in Table 3.13. This classification is based on what we took to be the main (the most informative) linking relation for each of the 204 bridging descriptions in the corpus[22].

3.4.2 Comparison with other classifications

In Tables 3.14 and 3.15 we compare the sub-classes of the bridging class just discussed with some of the taxonomies revised in Section §2.4. Table 3.14 shows the differences between the direct anaphora (Vieira 1) and co-referent (Vieira 2) classes adopted in experiments 1 and 2, respectively. As our instructions were not explicit about nominalization and summation, the scheme for Vieira 2 has an ambiguity between co-referent and associative for these cases, which might explain some of the disagreements.

3.5 Conclusions

We have now concluded our study of definite description use. We have seen in Chapter 2 and the present chapter that definite descriptions might be related to the discourse in many different ways; i.e.:

1. by direct coreference to a previously mentioned entity with same descriptive content (i.e., same head noun);
2. by (indirect) coreference to a previously mentioned entity with different descriptive content (different head noun);

[22]One problem with bridging references is that they are often related to more than one antecedent in the discourse (Fraurud, 1990; Strand, 1997).

Anaphoric uses	Clark	Strand	Vieira 1	Vieira 2
a book/ the book	IDENTITY	COREF (ident. head)	DIRECT ANAPHORA	CO-REFERENT
a lathe/ the machine	PRONOMINA-LIZATION	COREF (generalization)	BRIDGING (hypernym)	CO-REFERENT
a car/ the sedan	?	COREF (specification)	BRIDGING (hyponym)	CO-REFERENT
a man/ the bastard	EPITHET	COREF (redescription)	BRIDGING (inference)	CO-REFERENT
he travelled/ the journey	IDENTITY	COREF ?	BRIDGING (event)	*co-referent associative*
a man a woman/ the couple	?	WIDENING (members-set)	BRIDGING (inference)	*co-referent* em associative
Pinkerton Inc./ the company	*pronomina-lization*	COREF (redescription)	BRIDGING (names)	CO-REFERENT

Table 3.14: Comparison with other classifications: anaphoric uses

3. by associated reference to a previously mentioned entity—associated relations cover a wide range of distinct phenomena, as seen in Table 2.2 in Section §2.5) and in Section §3.4.1; or

4. a definite description may be discourse new.

Our empirical analysis assessed the familiarity hypothesis, and found that discourse new descriptions were very frequently used in our corpus (an average of 50% of the cases). Definite descriptions based on a same head antecedent were the second most frequent type in the corpus (30%), whereas definite descriptions which involve the most complex forms of lexical knowledge and inference (2 and 3 above) were not as frequent (20%). (See Section §3.2.3 and Section §3.3.3.)

3.5.1 Consequences for processing of definite descriptions

These findings suggest that identifying discourse new descriptions (which relate to general situation, common knowledge, or descriptions' complements) plays a role in processing definite descriptions that is as important as identifying the antecedent of subsequent mentions and bridging descriptions. This is one of the main hypotheses advanced in this dissertation, and has significantly influenced the design of the computer system described in the next chapters. Further support for our hypothesis comes from the literature discussed in the previous chapter:

- Heim observed that some definites fulfil the requirement for a cross-reference automatically, exemplified by (3.21):

Associative uses	Clark	Strand	Vieira 1 and 2
ANTECEDENT	ANTECEDENT	ANCHOR	ANCHOR
a book/ the author	?	DELIMITATION (subcategoriz.)	BRIDGING/ASSOC. (inference)
the room/ the ceiling	NECESSARY PARTS	NARROWING (whole-part)	BRIDGING/ASSOC. (meronymy)
the wall/ the building	?	WIDENING (part-whole)	BRIDGING/ASSOC. (holonymy)
the room/ the window	PROBABLE PARTS	NARROWING (whole-part)	BRIDGING/ASSOC. (meronymy)
the room/ the chandelier	INDUCIBLE PARTS	DELIMITATION (space anch.)	BRIDGING/ASSOC. (inference)
a couple/ the woman	SET-MEMBERSHIP	NARROWING (set-member)	BRIDGING/ASSOC. (inference)
clowns/ the clown with the unicycle	SET-MEMBERSHIP	NARROWING (set-member)	BRIDGING/ASSOC. (inference)
he killed her/ the murderer	NECESSARY ROLES	NARROWING (event-arg.)	BRIDGING/ASSOC. (event)
she died/ the murderer	OPTIONAL ROLES	?	BRIDGING/ASSOC. (event)
a professor/ the car	?	ADJOINING (posses.-thing)	BRIDGING/ASSOC. (inference)
an earthquake/ the suffering of people	REAS./CAUSE /CONSEQ.	ADJOINING (causation)	BRIDGING/ASSOC. (inference)
Israel and Egypt/ the peace agreement	?	DELIMITATION (arg.-event)	BRIDGING/ASSOC. (inference)
last Wednesday / the news	?	DELIMITATION (time anch.)	BRIDGING/ASSOC. (inference)
the first, the next, the last, the second	SET-MEMBERSHIP	ADJOINING (part-part)	BRIDGING/ASSOC. (inference)

Table 3.15: Comparison with other classifications: associative uses

(3.21) John read a book$_i$ and wrote to [the woman who had written it$_i$]$_j$

The descriptive content of NP$_j$ is said to explicitly contain a cross-reference to card i. In these cases, she says, there is no need for an additional requirement for cross-references, since such a requirement is generally in force.

- Löbner proposed that the defining property of definite descriptions, from a semantic point of view, is that they denote a functional concept as, for example, in *the father of Mr. Smith, the first man to sail to America, the fact that there is so much life on earth*. His semantic definites with explicit or situational arguments and his pragmatic endophoric definites are functional concepts which may well be discourse new.

This hypothesis has an important consequence. Hawkins studied in detail the connections between discourse new descriptions and certain syntactic constructions. This suggests that discourse new descriptions may be treated by heuristics which avoid mechanisms which are knowledge intensive and require inference. This would make the goal of automatic annotation of definite description use in unrestricted texts much easier to achieve.

Discourse old descriptions related to a same head noun antecedent do not require much common sense knowledge in order to be treated; however, special care is needed to take into account discourse structure (and its effect on salience) and noun modification. Other cases (such as indirect or associate reference described in 2 and 3 above) cannot be handled without referring to world knowledge. We tried WordNet as an approximation of the required general knowledge to deal with some of these complex cases.

In the next two chapters we describe and evaluate some computational experiments exploiting the ideas listed above.

3.5.2 Some caveats

Our experiments have also shown, however, that the distinctions traced above did not result in a consistent classification of definite description use. Better results were observed when the classes were reduced to two: coreferent and non coreferent. The most problematic class was bridging descriptions. The annotators were not intensively trained and the task definition was very broad. We believe that better results may be achieved by refining the task definition and instructions and if well trained annotators are employed.

Chapter 4

Processing Definite Descriptions in Unrestricted Text

In this chapter we present the prototype of a system for resolving definite descriptions in written text. Our goal was to build a system whose performance on unrestricted text[1] could be measured. Such a system cannot rely on purpose-specific knowledge coding and sophisticated inference mechanisms; instead we developed a shallow system and used WordNet (a publicly available lexicographical database) (Miller et al., 1993) as a source of common sense knowledge to deal with some cases of bridging descriptions.

The architecture of the system was based on the theory of definite descriptions processing first advanced by Fraurud and elaborated in the previous chapters, according to which interpreting definite descriptions in written discourse involves recognizing whether a description is, in Fraurud's terms, subsequent or first mention; or, in our terms, direct anaphora, discourse new or bridging. Different heuristics were devised for each of the classes. The development of the system was driven by our analysis of the corpus of the first experiment, and a set of heuristics was tested over this corpus (our training data). The heuristics we developed are discussed in this chapter; the experiments with the heuristics that led to the optimal system configuration are discussed next in Chapter 5.

When integrating our heuristics, we first determined the order of application of the heuristics by hand; subsequently we experimented with acquiring this order automatically: an alternative version of the prototype performs an analysis of the features of each definite description, and generates a list of these features together with the corresponding classification from the annotated corpus. The result is given as input to an implementation of Quinlan's ID3 learning algorithm (Quinlan, 1993).

This chapter is structured as follows. We first revise Fraurud's proposal for definite NP processing in Section §4.1. After that we present an overview of our system's structure in Section §4.2. Then, we present in detail the techniques employed to resolve each

[1] The pre-eminent aim of the project was to come up with techniques to resolve descriptions in unrestricted texts at large but this work was related, and the main results limited, to only one text genre—newswire material from the Penn Treebank (Wall Street Journal). Furthermore we make use of the parsed version of the texts. When we say unrestricted text we mean more precisely domain independent text.

of the different types of uses of definite descriptions: the heuristics for resolving direct anaphora are presented in Section §4.3; the heuristics for identification of discourse new descriptions are presented in Section §4.4; and in Section §4.5 we present experimental heuristics dealing with some of the bridging cases and use WordNet as a source of common-sense knowledge. The integration of the heuristics into an algorithm is presented in Section §4.6. In Section §4.7 we present an alternative algorithm based on a learned decision tree.

4.1 Fraurud's proposal

In Chapter 2, Section §2.3.2, we discussed Fraurud's criticism of the anaphoric (familiarity) approach to discourse processing of definite NPs. Because the large proportion of first mention definites she found in natural texts, Fraurud (1990) claims that:

> ... a model where the processing of first-mention definites always involves a failing search for an already established discourse referent as a first step seems less attractive. A reverse ordering of the procedures is, quite obviously, no solution to this problem, but a simultaneous processing as proposed by (Bosch and Geurts, 1989) might be (page 421).

Fraurud then suggests (on the basis of Löbner's account, Section §2.3.1) that properties other than definiteness/indefiniteness may also guide the selection of an appropriate interpretation strategy. One such property is the semantic characterization of the head noun: sortal head nouns are anaphoric, and an anaphoric procedure should be applied; relational head nouns are first mention, and a non-anaphoric procedure should be applied[2].

There are, however, several exceptions to the claim that sortal heads make anaphoric definites: for instance, definites such as *the fact that..., the moon, the word "the", the Empire State Building*, etc. She notes herself this claim to be too strong, and suggests either a non-anaphoric procedure for relational nouns, or a non-anaphoric procedure if an anaphoric procedure fails.

She proposes that to interpret a first mention definite NP, one should construct a new discourse referent[3] with links to one or more anchors and/or identify a background referent[4]. Anchors, she reckons, may be established prior to, or else, be contained in the definite. The global context (time, place and circumstances) also provides anchors, whereas background referents are related to the reader's previous knowledge. The procedure suggested by Fraurud is the following:

1. Establish a new discourse referent, D.

[2]Fraurud observed that first mention definites had dictionary definitions based mainly on relations to other concepts, such as the X of a/the Y, where X is a hyponym of the noun and Y described a type of anchor.

[3]Discourse referents are representations in the discourse model of entities explicitly mentioned.

[4]Background referents are entities that have not been mentioned in the discourse

2. Resolve D:
   (a) Identify one or more anchors to which D can be linked (by suitable relations):
      i. Determine the relevant number and types of anchors (arguments),
      ii. Select anchors.
   (b) Identify a background referent to which D can be linked by an identity relation.

Concerning this procedure she remarks:

- there is no temporal order for the procedures in 2,
- the procedures 2(a) and 2(b) are complementary: some of them may apply some may not, depending on the case. She considers the following possibilities:
  - an anchor may be selected by the reader without lexico-encyclopedic knowledge to guide the selection of anchors, just on the basis of the saliency of referents. An example is the interpretation of *the carburettor* as "something in the car" in a sequence such as *the car... the carburettor*, even when the owner of the car has never heard the word *carburettor* before
  - no background referent is identified because of lack of knowledge or non-existence (as in *the product of three and four is twelve*)
  - a background referent is directly identified without the help of anchors (*the Little Mermaid*)
- suitable relations in 2(a) are said to range from part-of and belong-to to spatio-temporal and situational relations.

Fraurud's example is the description *the king* taken as a relational noun, for which a new discourse referent is to be established. Some available lexico-encyclopedic knowledge would provide the information that a king is related to a period and a country; these would constitute the anchors. The selection of the anchors would identify the pertinent period and country, and this would make possible the identification of a referent: say, for the anchors 1989 and Sweden, the referent identified would be Carl Gustav XVI.

It is important to notice, however, that finding the relevant anchors is not a simple task. The easiest cases to be treated are those in which the anchors are contained in the description. Besides, expressions such as *the king*, *the government*, *the president*, etc may as well be used as subsequent mentions, in a sequence like *Carl Gustav XVI... The king*...; and if preference is to be given to more informative relations (as Strand (1996) suggests), then the identification of any proper co-reference relations should be assured and preferred to associative links. Furthermore, co-referential relations, specially when there is head noun identity, are a lot easier to deal with.

Fraurud's theoretical proposal answers for the interpretation of first mention definite NPs (or discourse new and bridging descriptions). Common to all is the establishment of a new discourse referent; then, they may be linked to an antecedent anchor or not, and they may identify a background referent or not. Basically, Fraurud proposes that one should take into account first mention descriptions when processing definite descriptions, and this we have pursued in our implementations, described in the next sections.

4.2 An overview of our system

We implemented a system that classifies definite descriptions according to their use. Our system is based on the same assumptions about definite description processing as Fraurud's proposal, i.e., it does not only try to identify the antecedents of subsequent mention definites, but it also tries to recognise first mention ones.

We have followed Strand's claim and give preference to the identification of co-referential links, so our priority is to find out the maximum number of co-referent descriptions. If the system finds a suitable same head antecedent for a definite description it classifies that description as direct anaphora. The system then tries to identify descriptions which are discourse new using heuristics that we developed on the basis of Hawkins' analysis and of our corpus study. We also implemented a version of the system which tries to find anchors for bridging descriptions, but this latter version is still in a preliminary form.

Whereas Fraurud claims that first mention definite descriptions should introduce a new discourse referent and either be linked to anchors (which may be linguistic or situational), or identify background referents, we adopted a somewhat different approach to the problem:

- we try to identify descriptions which are discourse new—for them there are no anaphoric anchors, there might be anchors which are contained in them, or situational anchors, but these situational anchors or background referents are not necessarily identified;
- we try to find the linguistic (anaphoric) anchors for those definite descriptions we have called associative and bridging descriptions, which lie somewhere between new and old descriptions.

The system is implemented in Sicstus Prolog.

4.2.1 Input

Our system processes newswire texts from the Penn Treebank I, making use of the syntactic information in the parsed version of the texts. The texts in the Penn Treebank corpus consist of a set of parsed sentences represented as Lisp lists. During a pre-processing phase, a representation in Prolog list format is produced for each sentence, and the noun phrases contained in it are extracted. The output of this pre-processing phase is passed to the proper system. For example, the sentence in (4.1) is represented in the Treebank as (4.2) and the input to the system after the pre-processing phase is (4.3)[5]. Note that all nested NPs are extracted, and that embedded NPs such as *the Organization of Petroleum Exporting Countries* are processed before the NPs that embed them (in this case, *the squabbling within the Organization of Petroleum Exporting Countries*).

[5]We will use a different notation for prolog variables to simplify the presentation and avoid quotation of atoms beginning with capital letters. Prolog variables will be indicated by the use of "_" in the beginning and in the end of the variables, as in _X_ for variable X.

(4.1) *Mideast politics* have calmed down and *the squabbling within the Organization of Petroleum Exporting Countries* seems under *control* for now.

(4.2)
```
( (S (S
      (NP Mideast politics)
      have
      (VP calmed
          down))
     and
     (S (NP the squabbling
         (PP within
          (NP the Organization
           (PP of
            (NP Petroleum Exporting Countries)))))
        (VP seems
         (PP under
          (NP control)))
        (PP for
         (NP now))))
  .)
```

(4.3)
```
[NP,Mideast,politics].

[NP,Petroleum,Exporting,Countries].

[NP,the,Organization,
      [PP,of,[NP,Petroleum,Exporting,Countries]]].

[NP,the,squabbling,[PP,within,[NP,the,Organization,
      [PP,of,[NP,Petroleum,Exporting,Countries]]]]].

[NP,control].

[[S,[S,[NP,Mideast,politics],have,[VP,calmed,
 [PP,down]]],and,[S,[NP,the,squabbling,[PP,within,
 [NP,the,Organization,[PP,of,[NP,Petroleum,Exporting,
 Countries]]]]],[VP,seems,[PP,under,[NP,control]],
 [PP,for,now]]]],.].
```

4.2.2 General structure

Figure 4.1 shows the general architecture of the system. The system processes parsed texts from the Treebank. It generates a representation of the discourse which consists of a list of potential antecedents with which definite descriptions may be resolved. The

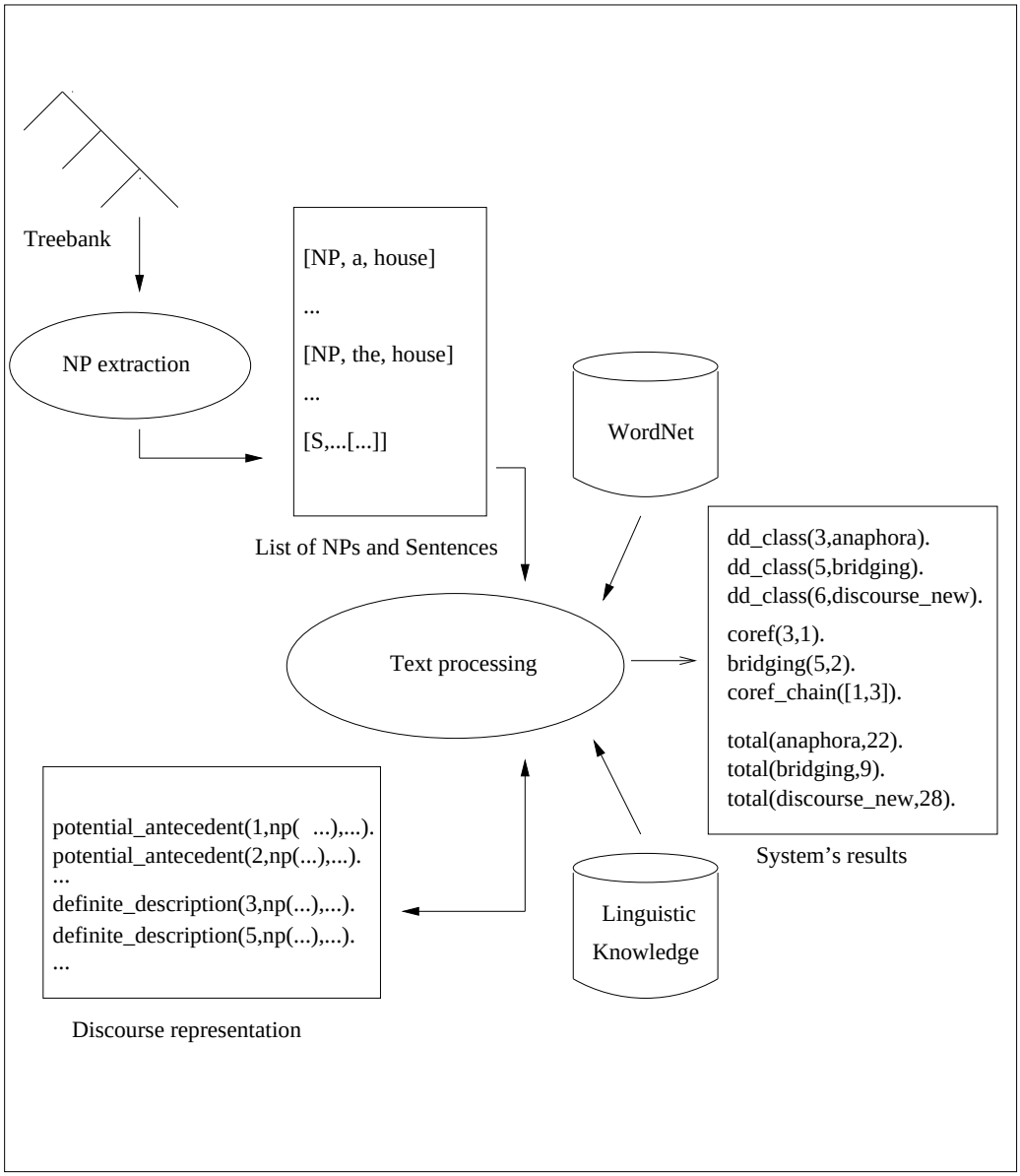

Figure 4.1: System architecture

system makes use of this representation and of linguistic knowledge about apposition structures, copular constructions, postmodifying restrictive clauses, etc. to identify discourse new definite descriptions to resolve them with an antecedent. WordNet is also consulted. The system's output consists of a classification of the instances of definite descriptions in the text, and of the identification of co-referential and bridging links.

In the next sections (Section §4.3, Section §4.4, Section §4.5) we present the main issues of the heuristics we developed to deal with the different types of definite descriptions. After that, in Section §4.6, we present the complete algorithm.

4.3 Direct anaphora

The key problems to be dealt with in order to resolve anaphoric definite descriptions are:

- to identify the potential antecedents, and
- try to match definite descriptions with the available antecedents.

Our basic strategy involves simply matching the head noun of the definite description with the head noun of a potential antecedent. The central information for the resolution of an anaphoric description with its antecedent is then the head noun.

4.3.1 Identifying head nouns

In the parsed texts of the Penn Treebank, the head noun is the atom in the far right position. For example, the nouns *politics* and *squabbling* are the heads of the following NPs:

(4.4) a. `[NP,Mideast,politics];`

b. `[NP,the,squabbling,[PP,within,[NP,the,Organization...]]].`

Headless definites such as the following were not taken into account.

(4.5) a. `[NP,the,[ADJP,fourth,largest]];`

b. `[NP,the,[ADJP,least,[WHPP,of,[WHNP,which]]]];`

c. `[NP,the,[ADJP,highest,[PP,in,[NP,the,[ADJP,southern,,],`
`so-called,Mezzogiorno,region]]]].`

Also there are some cases for which the head noun does not occur as an atom at the determiner level:

(4.6) a. `[NP,The,[NP,[NP,$,20,000],[NP,tax]]];`

b. `[NP,the,[ADJP,West,German],[NP,[NP,Bundesbank],'s,lead]];`
c. `[NP,the,[NP,[NP,$,40,000],[SBAR,0,[S,[NP,they],originally,`
`[VP,needed]]]]].`

A total of 17 cases of headless definite descriptions like those in (4.5) and (4.6), above, were counted in the first corpus; i.e., a not very significant percentage (0.2%). An additional problem was the way coordination is sometimes represented in the Treebank: our algorithm does not recognize that a noun such as *reporters* in (4.7) below is a head noun:

(4.7) `[NP,reporters,and,editors,[PP,of,[NP,The,WSJ]]].`

4.3.2 Potential antecedents

Every NP in the text is given an NP index (an integer), and sentences are ascribed a sentence index in the same manner. The sentence location of each NP is stored. Selected NPs are made available for the resolution of definite descriptions; we call them potential antecedents. They are represented in the system by Prolog assertions, which specify their NP index, the whole NP structure, the NP head noun, the NP type (definite, indefinite, bare plural, possessive), as illustrated by (4.8) below.

(4.8) `potential_antecedent(I,np(NP),head(H),type(T)).`

Depending on the choice of potential antecedents, different recall/precision trade-offs can be achieved. Some experiments were undertaken to identify the best group of potential antecedents. Four different NP subsets were taken into account in these tests:

1. indefinite NPs (those containing the indefinite articles *a, an, some* and bare/cardinal plurals[6]);
2. indefinite NPs and definite descriptions (NPs containing the definite article);
3. indefinite NPs, definite descriptions, and possessive NPs (with a possessive pronoun or possessive mark);
4. all NPs.

[6]Only plural nouns ending in *s* are dealt by the system.

Consider the NPs in (4.10), extracted from (4.9): the subsets 1, 2, 3 and 4 above correspond respectively to

- (4.11.a);
- (4.11.a) plus (4.11.b);
- (4.11.a) plus (4.11.b) and (4.11.c);
- (4.11.a) plus (4.11.b), (4.11.c) and (4.11.d).

They represent the available antecedents of each group

(4.9) In an interview with reporters of The Wall Street Journal, the candidate appears quite confident of victory and of his ability to handle the mayoralty.

(4.10)
```
 [NP,The,Wall,Street,Journal].
[NP,reporters,[PP,of,[NP,The,Wall,Street,Journal]]].
[NP,an,interview,[PP,with,[NP,reporters...l]]].
[NP,the,candidate].
[NP,victory].
[NP,the,mayoralty].
[NP,his,ability,[S,[NP,*],to,[VP,handle,[NP,the,mayoralty]]]].
```

(4.11) a.
```
potential_antecedent(_Index_,np(_NPstructure_),
                     head(reporters),
                     type(indef)).
potential_antecedent(_Index_,np(_NPstructure_),
                     head(interview),
                     type(indef)).
```

b.
```
potential_antecedent(_Index_,np(_NPstructure_),
                     head(Journal),
                     type(def)).
potential_antecedent(_Index_,np(_NPstructure_),
                     head(candidate),
                     type(def)).
potential_antecedent(_Index_,np(_NPstructure_),
                     head(mayoralty),
                     type(def)).
```

c.
```
potential_antecedent(_Index_,np(_NPstructure_),
                     head(ability),
                     type(possessive)).
```

```
d.  potential_antecedent(_Index_,np(_NPstructure_),
                         head(victory),
                         type(other)).
```

The comparative results for each subset are presented in the next chapter, in Section §5.3.1.

4.3.3 Segmentation

Antecedents may have a limited 'life span', i.e., NPs may only serve as antecedents for anaphoric expressions within a subset of the whole text. There may be semantic restrictions on their accessibility: this is the case for indefinite descriptions in the scope of operators such as negation and some modal verbs. Generic expressions may also introduce non-permanent referents. Making them accessible may cause errors. Also, it is not always the case that an indefinite NP introduces a discourse referent, as seen in the example below, (4.12).

(4.12) 42. The secret to being *a good adjuster*$_i$ is counting.
...
75. **The adjuster*$_i$ hadn't completed all the calculations, but says: We're talking policy limits.

Also, texts are divided in segments organized hierarchically, and the antecedents introduced in a segment at a lower level are typically not accessible from a segment at a higher level (Grosz, 1977; Grosz and Sidner, 1986). An example from the corpus is (4.13) where *the house* in sentence 50 does not refer to a house mentioned previously throughout the text in sentences 2 to 19, but it refers to another house implicitly introduced in sentence 49, after that, in sentence 65, the text returns to the previously mentioned house:

(4.13) 2. A deep trench now runs along its north wall, exposed when *the house*$_i$ lurched two feet off its foundation during last week's earthquake.
...
19. Others grab books, records, photo albums, sofas and chairs, working frantically in the fear that an aftershock will jolt *the house*$_i$ again.
20 The owners, William and Margie Hammack, are luckier than many others.
...
49. When Aetna adjuster Bill Schaeffer visited a retired couple in Oakland last Thursday, he found them living in a mobile home parked in front of their yard.
50. *The house*$_j$ itself, located about 50 yards from the collapsed section of double-decker highway Interstate 880, was pushed about four feet off its foundation and then collapsed into its basement.

...
65 As Ms. Johnson stands outside the Hammack house$_i$ after winding up her chores there, the house$_i$ begins to creak and sway.

Recognizing the hierarchical structure of texts is a difficult problem, although some algorithms are beginning to appear (Hearst, 1997; Richmond, 1997). We adopted a very simple segmentation technique, considering only the antecedents within fixed-size windows of previous sentences. In order to alleviate the problems arising from such a rudimentary conception of segmentation, we developed a modified heuristic allowing for some exceptions[7] . A potential antecedent would be considered for the resolution of a definite description when the antecedent's head is identical to the description's head, and

- its distance from the description is within the established window, or else
- the potential antecedent is itself a subsequent mention, or else
- the definite description and the antecedent are identical NPs (including the article).

Segmentation was also tested against and combined with RECENCY, by which we mean the heuristic of considering only the very last occurrence of a head noun as potential antecedent. The comparative results of the alternative heuristics just described are presented in Section §5.3.1.

4.3.4 Noun modifiers

The simplest form of resolution performed by the system involves trying to find a discourse antecedent with the same head as the definite description. Once the head nouns of antecedents and descriptions have been identified, a direct string matching is executed. In this way examples such as (4.14) are handled correctly.

(4.14) Grace Energy hauled *a rig* here... *The rig* was built around 1980.

But we have also to take into account the information provided by the prenominal and the postnominal part of the noun phrase. For example, *a blue car* cannot serve as the antecedent for *the red car*, or *the car of John* for *the car of Jane*. Examples from the corpus of antecedents that would be incorrectly suggested by simply matching heads without regarding premodification are:

[7] These exceptions are plausible for the kind of texts we worked with, which are usually not very long. Longer texts might require more restrictive rules.

(4.15) a. *the business community... the younger, more activist black political community*;
b. *the population... the voting population*;
c. *the East Coast... the West Coast*.[8]

In general, taking care of these modifications would require complex semantic reasoning. Instead, we considered some heuristics such as:

- allow an antecedent to match with a definite description if the premodifiers of the description are a subset of the premodifiers of the antecedent;
- allow a non-premodified antecedent to match with any same head definite.

The first of these heuristics deals with definites which contain less information than the antecedent, such as:

(4.16) a. *an old Victorian house... the house*;
b. *a retired couple in Oakland... the couple*;
c. *the San Francisco earthquake... the earthquake*.

It prevents matches such as:

(4.17) *the business community... the younger, more activist black political community*.

The second heuristic deals with definites that contain additional information. Examples from our corpus of pairs that match thanks to this heuristic are:

(4.18) a. *a check... the lost check*;
b. *the campaign... the Dinkins campaign*.

Cases in which co-referent descriptions present totally different premodification from their antecedents are less common, but some examples were found:

(4.19) a. *the very countercultural chamber group Tashi...the old Tashi*;
b. *the pixie-like clarinetist... the soft-spoken clarinetist*;
c. *a nuisance tax... The $ 20,000 tax*.

[8] In the case of proper names (e.g. 'the East Coast'), it is more correct to identify the head with the entire compound noun 'East Coast', and not the simple noun 'coast', but we have no means of identifying these cases. They are treated like the others.

Usually they indicate non-coreference, as for *the company's abrasive segment* and *the engineering materials segment*.

For cases like *the rules* in (4.20) where the last mention refers to a modified concept (new rules different from the previous ones), the heuristic suggests a wrong antecedent.

(4.20) Currently, *the rules* force executives...
The rule changes would...
The rules will eliminate...

In cases such as *the population... the voting population* where the new information indicates a subset, superset or part of a previous mentioned referent, the heuristic also produces an error.

Wrong resolutions due to postmodification also occur; however, same head antecedents with different postmodification are not as common as those with differences in premodification. Examples are:

(4.21) a. *the end of October... the end of November*;
b. *a chance to accomplish several objectives... the chance to demonstrate an entrepreneur like himself could run Pinkerton's better than an unfocused conglomerate or investment banker*;
c. *the sale of one residence... the sale of a home site*.

The heuristic used to deal with postmodification is to compare the description and antecedent, preventing resolution in those cases where both are postmodified and the modifications are not the same. These ideas could perhaps be developed to accept a resolution in which one postmodification is a subset of the other, resolving for instance *the use of Filipino* with *the use of Filipino language*. However, cases like that were not very frequent. The evaluation of these heuristics is presented in Section §5.3.1.

4.3.5 Co-referential chains

When a definite description (B) is resolved with an antecedent (A) a co-referential link is asserted: this is represented by a Prolog assertion of the following form.

(4.22) `coref(B,A).`

Descriptions resolved with the same antecedent form an equivalence class, we call this class a CO-REFERENTIAL CHAIN. Co-referential chains may be expressed in different ways. Consider three NPs indexed by A,B and C; the alternative markings (4.23.a) and (4.23.b) represent the same class (A,B,C), which is expressed by the predicate in (4.23.c).

(4.23) a. `coref(B,A). coref(C,A).`

b. `coref(B,A). coref(C,B).`

c. `coref_chain([A,B,C]).`

Whenever a description is resolved with an antecedent, the old co-referential chain (when there is one) is retracted and a new one is asserted, as in the example, where (4.24.b) is the resulting state after asserting the co-referential link between C and B to the state represented in (4.24.a).

(4.24) a.
```
coref(B,A).
coref_chain([A,B]).
```

b.
```
coref(B,A).
coref(C,B).
coref_chain([A,B,C]).
```

This idea of co-referential chains helps in the automatic evaluation of the system (presented in the next chapter).

4.4 Discourse new descriptions

Another set of heuristics is used to identify definite descriptions introducing new referents in the discourse (unfamiliar and larger situation uses of the definite article, in Hawkins' terminology). The identification of such descriptions is based on syntactic and lexical features of the noun phrase. The features used to recognise unfamiliar definites suggested by Hawkins[9] include:

- the presence of special predicates:
  - the occurrence of pre-modifiers such as *first* or *best* when accompanied with full relatives, e.g., *the first person to sail to America* (unexplanatory modifiers);

[9]Hawkins relates in detail different types of use with syntactic and grammatical structures. Although these relations do not guarantee that a certain type of use is realized, they have proved to be useful for identifying the uses of definite descriptions systematically.

  - a head noun taking a complement such as *the fact that there is life on Earth* (NP complements);

- the presence of restrictive modification, as in *the inequities of the current land-ownership system* (definite descriptions that contain relative clauses and associative clauses).

In addition, we considered as unfamiliar those definites occurring in[10] :

- appositive structures (e.g.,*Glenn Cox, the president of Phillips Petroleum Co.*);
- copular constructions (e.g.,*the man most likely to gain custody of all this is a career politician named David Dinkins*).

Three classes of larger situation definites can also be recognised on the basis of syntactic and lexical features[11]:

- definites that behave like proper nouns, like *the United States* (recognized by checking upper case);
- definites which have proper nouns in their premodification, such as *the Iran-Iraq war*;
- definites referring to time, such as *the time* or *the morning* (which as unexplanatory modifiers and NP complements are recognised by consulting a list of special predicates).

Although each of the heuristics refers, in principle, to one of the uses (larger situation or unfamiliar) they work better as identifying all together the class of discourse new descriptions. Next we present in detail the heuristics we used.

4.4.1 Special predicates

Some cases of discourse new definite descriptions are identified by comparing the definite NP (head noun or modifiers) with a list of predicates that are either functional or likely to take a complement. The list of predicates that may take NP complements currently includes the nouns *fact, result, conclusion, idea, belief, saying* and *remark*. The system also checks if a complement is present.[12] In these cases, the definite description may be functional on purely semantic grounds, because the relative clause specifies its value. An example is given in (4.25).

[10]Definite descriptions in appositive and copular constructions alternatively may be regarded as coreferent with their complements. In this work we have considered these constructions as a unity that introduces a new referent to the discourse.

[11]Some other larger situation uses could be recognized by having the context of utterance represented by the time and place of the newspaper publication. But we haven't consider this information in our implementations.

[12]These predicates may also appear in copular constructions (*the fact is that ...*).

(4.25) Mr. Dinkins also has failed to allay Jewish voters' fears about his association with the Rev. Jesse Jackson, despite *the fact that few local non-Jewish politicians have been as vocal for Jewish causes in the past 20 years as Mr. Dinkins has*.

A second list of special predicates consulted by the system corresponds to Hawkins' unexplanatory modifiers: *first, last, best, most, maximum, minimum, only*, some comparatives such as *more, closer, greater, bigger* and superlatives in general[13]. The presence of a complement is verified for some of the modifiers (*first, last* and comparatives *greater, larger*, etc.), but not for superlatives. When these features are verified the definite may be classified as discourse new because the modifier makes the description a complex functional. See examples below.

(4.26) a. Mr. Ramirez just got *the first raise he can remember in eight years*, to $ 8.50 an hour from $ 8.

b. Mr. Stolzman offered *the most substantial music of the evening* just after intermission.

c. She jumps at *the slightest noise*.

d. Y.J. Park and her family scrimped for four years to buy a tiny apartment here , but found that *the closer* they got to saving the $ 40,000 they originally needed, *the more* the price rose.

Finally, there is a list of special predicates related to larger situation uses (based on general knowledge) which consists of terms indicating time reference. This list is composed of the words *hour, time, morning, afternoon, night, day, week, month, period, quarter, year* and their respective plurals. Examples of such definites from the corpus are:

(4.27) a. Colleagues today recall with some humour how meetings would crawl into *the early morning hours*.

b. Some legislators think *the time* may be ripe to revise the constitution.

c. The mood is more upbeat trucks rumble along the dusty roads and burly men in hard hats sweat and swear through *the afternoon sun*.

d. We've been putting in long hours, Mr. Ramirez says–six-day weeks and 13-hour days for *the last two months*.

e. Only 14,505 wells were drilled for oil and natural gas in the U.S. in the first nine months of *the year*, down 22.4% from *the like 1988 period*.

[13]This list should be made more comprehensive; so far it includes the cases observed in the corpus analysis and a few other similar modifiers.

These descriptions indicate FC1 concepts. Other kinds of such uses are *the moon*, *the sky*, or *the pope*, *the weather*. They could also be recognised on the basis of lexical information, but coding this sort of information by hand was avoided for the moment.

Although the constructions just presented may all license a discourse new interpretation, nothing prevents these expressions from being used anaphorically. This question is further discussed in the evaluation of the system performance in Chapter 5.

4.4.2 Restrictive modification

Another structural feature of definite noun phrases verified by the system is the presence of restrictive modification. Definite descriptions may include pre and/or postmodifiers. Premodifiers come before the head; they are mostly adjectives and nouns. Premodification is usually non-restrictive. Postmodifiers come after the head, and are mostly interpreted as restrictive (Quirk et al., 1985)[14].

Restrictive and non-restrictive modifiers According to (Quirk et al., 1985), in restrictive modification, the NP's head refers to an entity which can be identified only through the modification that has been supplied. In non-restrictive modification the head refers to an entity independently identified, the modification in this case provides additional information which is not essential for identifying the referent. Non-restrictive modification mostly occurs in pre-head position. When non-restrictive modification occurs in post-head position, it is usually enclosed by commas.

Restrictive postmodification

Restrictive postmodification is the most frequently observed feature of first mention descriptions. In these cases, the restrictive postmodifier provides an anchor for the interpretation of the description. This anchor may provide a link to the rest of the discourse or make the description a functional concept with explicit arguments. The different forms of restrictive postmodification encountered in the corpus are presented and exemplified below.

Relative clauses may be introduced by relative pronouns such as *who, whom, which, where, when, why, that*, or "zero" relative pronoun.

(4.28) a. *The girl who I met...*
b. *The place where he lives...*
c. *The reason why she left...*
d. *The boy that is playing guitar...*
e. *The guy we met...*

[14]Examples in this subsection are taken from, or similar to those in (Quirk et al., 1985), unless otherwise specified.

Restrictive Postmodification	#	%
Prepositional Phrases	152	77%
Relative Clauses	45	23%
Total	197	100 %

Table 4.1: Distribution of prepositional phrases and relative clauses.

Prepositional Phrases	#	%
Of-phrases	120	79%
Other prepositions	32	21%
Total	152	100%

Table 4.2: Distribution of prepositions (1)

Non-finite post-modifiers include *ing, ed* (participle), and infinitive clauses.

(4.29) a. *The man writing the letter is my friend.*
b. *The train just arrived at platform 1 is from York.*
c. *The man to consult is Wilson.*

Prepositional phrases Quirk et al. (1985) claim that prepositional phrases are the commonest type of postmodification in English, three or four times more frequent than either finite or non-finite clausal post-modification. This was confirmed by our corpus study: Tables 4.1, 4.2 and 4.3 show the distribution of postmodifiers, and the types of prepositions observed for 188 postmodified descriptions.

The full range of prepositions is involved, as illustrated below. The preposition *of* is the commonest of all.

(4.30) a. *The book on grammar...*
b. *The issue of student grants...*
c. *The years before the war...*
d. *The man behind the door...*

Hawkins mentioned referent establishing relative clauses and associative clauses as two constructions that license an unfamiliar definite, but also warned that not all relative clauses are referent establishing.

Fraurud (1990), similarly, observed that 75% of the complex definite NPs (genitives, postposed PPs, restrictive adjectival modifiers) were first mention in her corpus. As it turns out, a great number of definite descriptions with restrictive post-modifiers are unfamiliar in our corpus. We found instances of many different cases of postmodification, as shown below:

Other prepositions	#	%
in	8	25%
for	7	22%
on	7	22%
to	4	12%
others	6	19%
Total	32	100%

Table 4.3: Distribution of prepositions (2)

(4.31) a. Santa Fe Energy Co. bought from Amoco *the rights that allowed it to drill the Sharpshooter*.

b. During the past three months there have been several demonstrations at *the office complex where the Land Bureau is housed*.

c. Considered as a whole, Mr. Lane said, *the filings required under the proposed rules* will be at least as effective.

d. They wonder whether he has *the economic know-how to steer the city through a possible fiscal crisis*.

e. What the investors object to most is *the effect they say the proposal would have on their ability to spot telltale clusters of trading activity*.

f. Some in Big Oil are easing *the grip on their wallets*.

g. Mideast politics have calmed down and *the squabbling within the Organization of Petroleum Exporting Countries* seems under control for now.

h. *The appetite for oil-service stocks* has been especially strong, although some got hit yesterday when Shearson Lehman Hutton cut its short-term investment ratings on them.

i. If you know you've got stability in price, you can do things you wouldn't do with *the volatility of the past few years*.

j. For the Parks and millions of other young Koreans, *the long-cherished dream of home ownership* has become a cruel illusion.

Our program used the following patterns to identify relative and associative clauses:

(4.32) a. `[NP,the,_Premodifiers_,_Head_,[SBARQ|_]|_];`

b. `[NP,the,_Premodifiers_,_Head_,[SBAR|_]|_];`

c. `[NP,the,_Premodifiers_,_Head_,[S|_]|_];`

d. `[NP,the,_Premodifiers_,_Head_,[VP|_]|_];`

e. `[NP,the,_Premodifiers_,_Head_,[PP,_|_]|_];`

f. `[NP,the,_Premodifiers_,_Head_,[WHPP,_|_]|_].`

Sometimes the modified NP is embedded in another NP in the Treebank, so structures like the one below are also considered (again for all types of clauses just shown above):

(4.33) `[NP,[NP,the,_Premodifiers_,_Head_],[Clause]].`

An NP may have zero, one, or more premodifiers, as shown in the following examples from the corpus. The actual procedure looks for lists such as the ones above following the head noun.

(4.34) a. the squabbling within the Organization;
b. the *economic* know-how to steer the city;
c. the *flourishing, high-production* trait known as hybrid vigor.

Non-restrictive postmodification Non-restrictive postmodifiers in definite descriptions are usually differentiated from restrictive post-modifiers by the use of commas, as seen in the following examples.

(4.35) a. The apple tree, swaying in the breeze, had a good crop of fruit.
b. The substance, discovered almost by accident, is very important.
c. The book, on grammar, ...
d. The issue, of no importance, ...
e. He met Mary, who invited him to a party.

The system will not consider such modifications as an indication of discourse new descriptions because they are usually additional information that is not essential for identifying the referent. These cases are encoded in the Penn Treebank as follows:

(4.36)
```
[NP,the,proposal,',',[SBAR,[WHNP,which],also,[S,[NP,T],would,
                [VP,create,[NP,a,new,type,[PP,of,[NP,individual,
                                retirement,account]]]]]], ',' ]...
```

Restrictive premodification

Restrictive modification in pre-head position is not so common as in the post-head position, but it is often used. These structures are interesting because they also correlate well with larger situation and unfamiliar uses of definite descriptions. A restrictive premodifier may be a noun or a proper noun. The examples below are extracted from the corpus.

(4.37) a. Mr. Koch already has announced he will drop 3,200 jobs from *the city payroll*, but that won't be enough.

b. A native of the area, he is back now after riding *the oil-field boom* to the top, then surviving the bust running an Oklahoma City convenience store.

c. Norman Young, a "mud-logger" at *the Sniper well*, has worked all but about nine days of this year.

d. About the same time, *the Iran-Iraq war*, which was roiling oil markets, ended.

e. In the process, "Batibot," an archaic Filipino word meaning "strong" or "enduring," has become a powerful advocate of the use of *the Filipino language*.

f. The two sides also traded accusations about the cost of *the Packwood plan*.

The heuristic we used was to classify definite descriptions premodified by a proper noun as larger situation. We could not distinguish adjectives or verb from nouns in premodification because this information was not present in the version of the Treebank that we used, as shown by the examples in (4.38). Sometimes numbers (usually referring to dates) also work as restrictive premodification.

(4.38) a. `[NP,the,1987,stock,market,crash];`

b. `[NP,The,proposed,changes];`

c. `[NP,the,soft-spoken,clarinetist].`

4.4.3 Apposition

Definite descriptions occurring in appositive constructions are usually discourse new and are resolved locally. Appositive constructions are treated in the Treebank as NP modifiers; therefore the system recognises an apposition by checking whether the definite is inserted in a complex noun phrase with structure like those in (4.39), consisting of a sequence of noun phrases (which might be separated by comma, or not) one of which is a name or is premodified by a name.

(4.39) a.
```
[NP,[NP,Glenn,Cox],',',[NP,the,president,
                        [PP,of,[NP,Phillips,Petroleum]]]];
```

b.
```
[NP,[NP,the,oboist],[NP,Heinz,Holliger]].
```

In fact the description may itself be a name in an appositive construction with an indefinite NP, as shown in (4.40). Such cases of appositive constructions were also taken into account.

(4.40) *the Sandhills Luncheon Cafe*, a tin building in midtown.

The apposition may be an embedded NP; an example is the definite noun phrase *the former crime buster* in (4.41) below. The system takes these cases into account.

(4.41)
```
[NP,[NP,Rudolph,Giuliani],',',
            [NP,[NP,the,former,crime,buster]...]].
```

Other examples of apposition recognised by the system are:

(4.42) a. *the very countercultural chamber group* Tashi;
b. *the new chancellor*, John Major;
c. *the Sharpshooter*, a freshly drilled oil well two miles deep;
d. *the Bilbrey well*, a 15,000-foot, $ 1-million-plus natural gas well;
e. *the Vivaldi-at-brunch set*, *the yuppie audience that has embraced New Age as its very own easy listening*.[15]

4.4.4 Copular constructions

Definites occurring in copular constructions such as *the Prime Minister is Tony Blair* do not necessarily involve a relation with a textual antecedent; many of them should be classified as discourse new.

Our heuristic for handling copula constructions works as follows. If a description occurs in subject position, the system looks at the VP. If the head of the VP is the verb *to be*, *to seem*, or *to become* and the complement of the verb is not an adjectival phrase, the system classifies the description as discourse new. See for instance the structure in 4.43.

[15] In this example both descriptions are identified as being in an appositive construction.

(4.43) `[S,[NP,The,fact],[VP,is,[NP,[SBAR,that... ]]]].`

Examples from the corpus are:

(4.44) a. *The bottom line* is that he is a very genuine and decent guy.
b. When the dust and dirt settle in an extra-nasty mayoral race, *the man most likely to gain custody of all this* is a career politician named David Dinkins.

If the complement of the verb is an adjective, the subject is typically interpreted referentially and should not be consider as discourse new on the basis of its complement, as in *The president of the US is tall*. Examples are:

(4.45) a. *The missing watch* is *emblematic* of the problems Mr. Wathen encountered.
b. *The new stirrings* are *faint*.
c. *The activity* is *enough* to move some oil-service prices back up a little.
d. *The guy* is so personally *decent*...
e. *The earnings* were *fine* and above expectations.

The adjectival complement is represented as follows in the Treebank:

(4.46) `[S,[NP,The,missing,watch],[VP,is,[ADJP,emblematic...]]].`

The definite descriptions in object position of the verb *to be*, such as the one shown in (4.47), are also considered discourse new.

(4.47) What the investors object to most is *the effect they say the proposal would have on their ability to spot telltale "clusters" of trading activity*.

4.4.5 Proper names

Proper names preceded by the definite article are often used in the genre we are dealing with, newspaper articles. They name entities supposedly known by the readers, although they might be new for some readers. Their first appearance in the text is usually a discourse new description. Subsequent mentions of proper names are regarded as cases of anaphora. To recognize proper names the system checks whether the head is a proper noun by checking if it is capitalised. If the test succeeds, the definite is classified as a larger situation use[16]. Examples include:

(4.48) a. *the General Accounting Office*;
b. *the Wall Street Journal*;
c. *the Securities and Exchange Commission*.

This concludes the discussion of our heuristics for the identification of discourse new descriptions. Their performance is discussed in the next chapter.

4.5 Bridging descriptions

Linguistic and computational theories of bridging descriptions identify two main subtasks involved in their resolution: first, finding the element in the text to which the bridging description is related (ANCHOR) and second, finding the relation (LINK) holding between the bridging description and its anchor (Clark, 1977; Sidner, 1979; Heim, 1982; Carter, 1987; Fraurud, 1990; Strand, 1997). A speaker is licensed to use a bridging description when he/she can assume that the common-sense knowledge required to identify the relation is shared by the listener (Hawkins, 1978; Clark and Marshall, 1981; Prince, 1981).

As discussed in Chapter 3, bridging descriptions are the most complex class of definite descriptions. They are not an homogeneous class, and their interpretation is heavily dependent on inference. Furthermore, many kinds of relations are possible between bridging descriptions and their anchors, and the same description may relate to different anchors in a text. For all these reasons, this class has been the most challenging of the problems we dealt with in the development of our system, and the results of our initial experiments are not as good as those we obtained for the other classes. The main result is what we learned about this class.

As discussed in Section §3.4, instead of adopting a semantic classification of the possible relations between descriptions and their anchors, the types of bridging descriptions found in our corpus were listed according to the kind of processing they required. In summary, we found:

[16]Note that this test is performed just after trying to find an antecedent, so that the second instance of the same proper (head) noun will be classified as an anaphoric use.

- cases of description/antecedent (anchor) pairs based on well-defined lexical relations, such as synonymy, hypernymy and meronymy;
- cases in which the antecedent is a proper name and the description a common noun;
- cases in which the anchor is not the head noun but a noun modifying an antecedent;
- cases in which the antecedent (anchor) is not introduced by an NP but by a VP;
- cases in which the antecedent is only implicitly available, e.g., because it is a discourse topic;
- finally, cases in which the relation with the anchor is based on deeper (i.e., non-lexical) inferences, such as set-subset or cause-consequence relations.

For some of the cases above, we proposed and implemented heuristics which were tested against our data (Poesio, Vieira and Teufel, 1997). We describe these heuristics in the rest of the section.

4.5.1 Bridging descriptions and WordNet

The dependence of bridging descriptions on common sense knowledge means that, in general, a system can only resolve bridging references when supplied with an adequate knowledge base; for this reason, the typical way of implementing a system for resolving bridging references has been to restrict the domain and feed the system with hand-coded world knowledge. (This approach, already proposed by Sidner, is developed in detail in (Carter, 1987)). In order to get a system capable of performing on unrestricted text, we decided to use WordNet (WN) (Miller et al., 1993) as an approximation of a knowledge base containing generic information.

We developed a WordNet interface (Vieira and Teufel, 1997) that reports a possible semantic link between two nouns when one of the following is true:

- the nouns are in the same synset (i.e., they are synonyms of each other), as in *suit/lawsuit*;
- the nouns are in a hyponymy/hypernymy relation with each other, for instance, *dollar/currency*;
- there is a direct or indirect meronymy/holonymy (part of/has parts) relation between them, as in *house/door*;
- the nouns are *coordinate sisters*, i.e. hyponyms of the same hypernym, such as *home/house*, which are hyponyms of *housing*, *lodging*.

We adopted a recency rule and the WordNet interface to identify bridging descriptions' anchors; i.e., the system would go back one sentence at a time, and stop as soon as a relation with a potential anchor was found.

Sometimes, a relation between two head nouns is not encoded in WN directly, but there is a relation between compound nouns in which these nouns appear. Thus, although there is a semantic relation between *record/album*, we find a synonymy relation only between *record_album/album*. But this extended search as well as the search for indirect meronymy relations yielded extremely low recall and precision at a very high computational cost; both types of search were dropped at the beginning of the tests we ran to process the corpus consulting WN (our automatic search for anchors).[17] The results of our tests with WordNet are presented in the next chapter (Section §5.5.1).

4.5.2 Bridging descriptions and proper names

Definite descriptions which refer back to proper names (such as *Pinkerton Inc... the company*) are very common in newspaper articles. Processing such descriptions requires determining an entity type for each name in the text. If we get the entity type *company* for a name such as *Pinkerton Inc.*, we can then resolve the subsequent description *the company* on the basis of this information. Or else we could resolve a description (such as *the firm*) using an entity type (*company*) by finding out a synonymy relation between them using WordNet.

In order to find out the type of named entities, we can consult WordNet: a few names are available—typically, of famous people, countries, states, cities and languages. Other entity types can be identified using appositive constructions and abbreviations like *Mr., Co., Inc.* etc. as cues. The algorithm for assigning a type to proper names was based on a mixture of the heuristics just described. The system first looks for the above mentioned cues to try to identify the name type. If no cue is found, pairs consisting of the proper name and each of the elements from the list *country, city, state, continent, language, person* are consulted in our WordNet interface to verify the existence of a semantic relation.

Including a back-tracking mechanism which re-processes a text filling in the discourse representation with missing name types increased our recall. With this mechanism we identify the type for the name *Morishita* in a textual sequence like *Morishita — Mr. Morishita*. The first occurrence of the name has no surface indication of the entity type, but the subsequent mention has (*Mr.*). By processing the text twice we recover such missing types.

After finding the types for names, we use same head matching or WordNet lookup to match descriptions with the types found for previous named entities.[18]

[17] They were only used when testing if WordNet encoded the semantic relations that we manually identified.

[18] The problem of named entity recognition and categorization has received considerable attention recently (Mani and MacMillan, 1996; McDonald, 1996; Paik et al., 1996; Bikel et al., 1997; Palmer and Day, 1997; Wacholder et al., 1994). It was also one of the tasks of the Sixth Message Understanding Conference (MUC-6), and 15 different systems participated in the competition for this task (Sundheim, 1995).

4.5.3 Compound nouns

Sometimes, a bridging description may be linked to a non-head noun in a compound noun:

(4.49) a. *stock market crash... the markets*;
b. *rule changes... the rules*;
c. *discount packages... the discounts*.

One way of processing these definite descriptions would be to update the discourse model with discourse referents not only for the NP as a whole, but also for the embedded nouns: for example, after processing *stock market crash*, we could introduce a discourse referent for *stock market*, and another discourse referent for *stock market crash*. The description *the markets* would be co-referring with the first of these referents (with identical head noun), and then we could simply use our anaphora resolution algorithms. This solution, however, makes available discourse referents that are generally inaccessible for pronominal anaphora[19] (Postal, 1969; Ward, 1991). For example:

(4.50) I saw *[a deer$_i$ hunter]$_j$. It$^*_i$* was dead.

We therefore followed a different route: our algorithm for identifying anchors attempts to match not only heads with heads, but also the following.

The head of a description with the pre-modifiers of a previous NP:

(4.51) a. *the stock market crash... the markets*;
b. *rule changes... the rules*.

The pre-modifiers of a description with the pre-modifiers of its antecedents:

(4.52) a. *most oil companies... the oil fields*;
b. *his art business... the art gallery*.

And finally, the pre-modifiers of the description with the head of a previous NP:

(4.53) a. *New York City... the city council district lines*;
b. *a 15-acre plot and main home... the home site*.

[19]Note that the collection of potential antecedents containing all NPs will just have the NP head *crash* for *stock market crash*. The system considers the whole NP structure as one only discourse referent, according to the structure of the Penn Treebank: [NP,the,1987,stock,market,crash].

Same head antecedent, bridging reference There are also cases in which the pre-modifiers together with the head noun of a description may indicate a bridging reference: we may find an antecedent NP with the same head noun for a description but referring to a different entity, this being signalled by the pre-modification. Some examples are:

(4.54) a. *the company's abrasive segment... the engineering materials segment*;
b. *Italy's unemployment rate... the southern unemployment rate*;
c. *Pinkerton... the new Pinkerton*;
d. *increases of 3.9 %... the actual wage increases may have been bigger*.

Our previous heuristics for treatment of pre-modifiers in anaphoric resolution handled the first two examples correctly: as they present different pre-modifiers we did not treat them as anaphoric in the first version of our system. Such cases, as well as descriptions modified by adjectives such as *new* and *actual* (last two examples), may now be treated as bridging references[20].

4.5.4 Bridging descriptions based on VPs

There is no information about relations between nouns and verbs in WordNet. To process definite descriptions based on VPs (referring to events, situations or propositions), one would have to transform verbs into their nominalization, and then look for a relation in WordNet. Some nominalizations can be generated by general procedures or learned by means of a stochastic method: e.g., we could use WordNet's morphology component as a stemmer, and augment the verbal stems with the most common suffixes for nominalizations which could be kept in a list, like *-ment, -ion*. These ideas have not been implemented. Instead, we simply tested the matching of truncated verbs and descriptions, and this has worked well for the few such cases found in our corpus. Cases of definite descriptions based on events which are resolved with such devices are:

(4.55) a. changes were proposed... *the proposals*;
b. something has changed... *the change*;
c. the government will penalize offenders... *the penalties*;
d. he plans to ... *the plan*.

There are bridging descriptions based on VPs that, however, require reasoning based on the compositional meaning of the phrases (as in *It went looking for a partner... pitching the prospect*); in fact, most of the cases are of this type. These cases are out of reach just now, as well as the cases listed, in Section §3.4, under discourse topic (those with no explicit textual antecedent) and inference (other complex non-lexical relations between NPs).

[20] This idea is not implemented.

4.6 Integration of the heuristics

In this work we discuss different implemented versions of the system (as presented in Chapter 4). Two of them resolve direct anaphora and identify discourse new descriptions; another version also deals with bridging descriptions (and accesses WordNet). The general structure of the implemented algorithm is summarised as follows. For each NP of the input:

1. The system assigns it an index.
2. NPs which are taken as potential antecedents (as described in Section §4.3.2) are made available for description resolution.
3. If the NP is a definite description, the system applies to it the following tests. The first test passed by the definite (if any) determines its classification, and after that the next NP is processed.
   (a) Examine a list of special predicates in order to identify some of the unfamiliar and larger situation uses.
   (b) Check whether the definite NP occurs in an appositive construction; there is no need to find an antecedent for those either. They are classified as discourse new, unfamiliar uses.
   (c) Try to find an antecedent for the definite description by matching head nouns and dealing with premodification, postmodification and respecting segmentation and recency. If the test succeeds the description is classified as direct anaphora and the relation of co-reference between the two NP indexes is asserted.
   (d) Verify if the head of the NP is a proper noun (by checking whether it's capitalised). If so, the description is classified as discourse new, larger situation use.
   (e) Check if the definite presents restrictive postmodification. Definites which are not anaphoric and have restrictive postmodifiers are classified as discourse new, unfamiliar uses.
   (f) The system verifies if there is a proper noun in premodifier position; if so, it is considered as a restrictive premodification, and the definite description is classified as discourse new, larger situation use.
   (g) Check if the definite occurs in a copula construction. If so, the description is classified as discourse new, unfamiliar use.
   (h) If the tests above failed, the version of the system which deals with bridging references initiates a search for an anchor according to the following heuristics (respecting their order):
      i. proper names
      ii. compound nouns
      iii. WordNet look-up

> If one of the three tests above succeeds the description is classified as bridging and the association between description and anchor indexes is asserted.

The system is not able to classify all occurrences of definite descriptions: when all tests fail the definite description is not classified. This algorithm's corresponding decision tree is presented in Figure 4.2.

Note that before trying to find an antecedent, the system executes a few tests for identifying discourse new descriptions; the strategy adopted is:

- eliminate some non-anaphoric cases (first two tests)[21],
- try to find a same head antecedent (third test),
- look for an indication that the description is discourse new (following four tests),
- try to find an anchor for the definite description (last test).

The heuristics for recognizing bridging descriptions are only applied when the other heuristics fail. This is because the performance of these heuristics is very poor and also because some of the heuristics which deal with bridging descriptions are computationally expensive; the idea was to eliminate those cases less likely to be bridging before applying these heuristics. We observed in our first tests (see next chapter) that definite descriptions which are not resolved as direct anaphora and not identified as discourse new by the previously presented heuristics were (according to the corpus analysis), mostly, bridging descriptions or discourse new[22].

For each text processed the system counts and displays the number of sentences, number of NPs considered as antecedents (indefinites, possessives, definites and others), and the number of definite descriptions processed. The system also displays its classification of descriptions, number of larger situation uses (first occurrence of proper names, time references and restrictive premodification), number of unfamiliar uses (NP complements, explanatory modifiers, appositive clauses, restrictive postmodification and copula constructions), number of resolved anaphoric descriptions, and the number of indefinites, possessives and definites identified as antecedents in the resolution process, and finally the number of non-identified description. (As shown in Section §5.4.) The user can visualise the co-referential classes achieved in the processing of a text. The user can also check what was found for other classes of uses of the definite article. Examples of the system's functions are presented in Appendix C.

[21] We considered special predicates and apposition as reliable indications of discourse novelty, also some of them produced some errors in anaphora resolution which were eliminated by processing them first.

[22] Examples of discourse new descriptions not identified by our heuristics are larger situation uses such as *the world, the nation, the government, the economy, the marketplace, the spring, the other hand, the spot, the 1920s*, or discourse new NPs with restrictive premodification such as *the low 40% range, the defense capital good sector, the residential construction industry, the developing world, the world-wide supercomputer market*, etc.

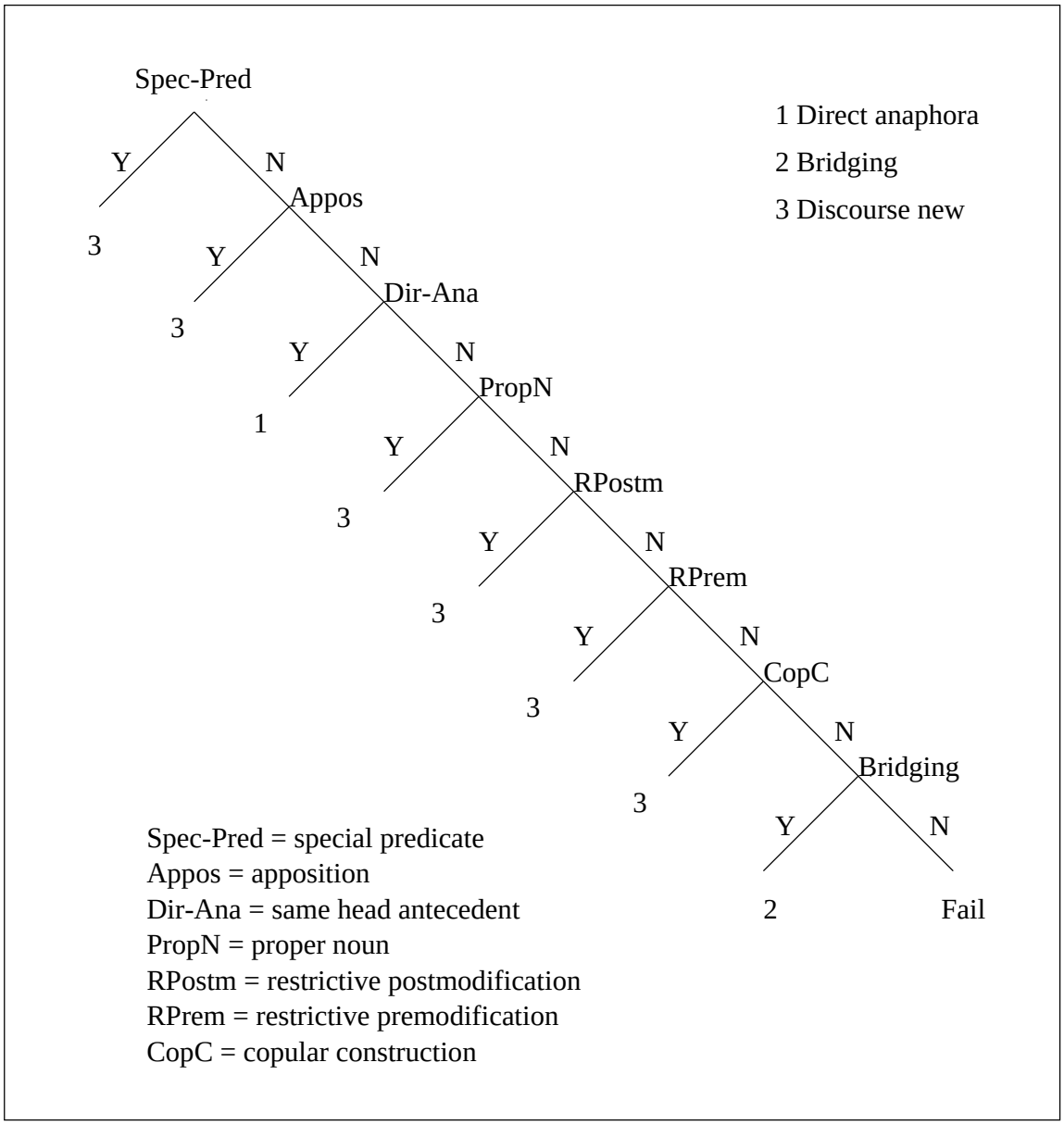

Figure 4.2: Heuristics Integration

4.6.1 An experiment with multiple classification

A version of the system which performs multiple classification was implemented. If a definite description has an antecedent but at the same time has some of the features signalling discourse novelty it is classified as ambiguous. The following indicators of discourse new use were verified:

- special predicates,
- appositive and copula constructions or
- postmodification.

The results are presented in Section §5.4.3.

4.7 An inductive decision tree

The order of application of the heuristics mentioned above was arrived at by trial and error. We also tried to arrive at this order automatically, as follows. We used a modified version of the system to assign feature values (yes, no) to definite descriptions in the training corpus. The following features were checked (the system checks if the features apply to a definite description instance or not):

1. Special predicates: the presence of a special predicate (according to the specification in Section §4.4.1), and verification of complement when needed (Spec-Pred).
2. Direct anaphora: existence of an antecedent with same head noun (respecting the established constraints) for that description (Dir-Ana).
3. Apposition: when the description is in appositive construction (Appos).
4. Proper noun: when the description has a capitalized initial (PropN).
5. Restrictive postmodification: the presence of relative or associative clauses(RPostm).

This list of features[23] was fed together with the classification given by the manual annotation of the corpus (DDUse) as shown in example (4.56) to an implementation of the Quinlan's learning algorithm ID3 (Quinlan, 1993).

(4.56)

Spec-Pred	Dir-Ana	Appos	PropN	RPostm	DDUse
no	no	no	yes	no	3
no	no	no	no	yes	3
no	no	no	no	no	2
no	no	no	no	no	2
no	no	no	no	no	1
no	yes	no	no	no	1

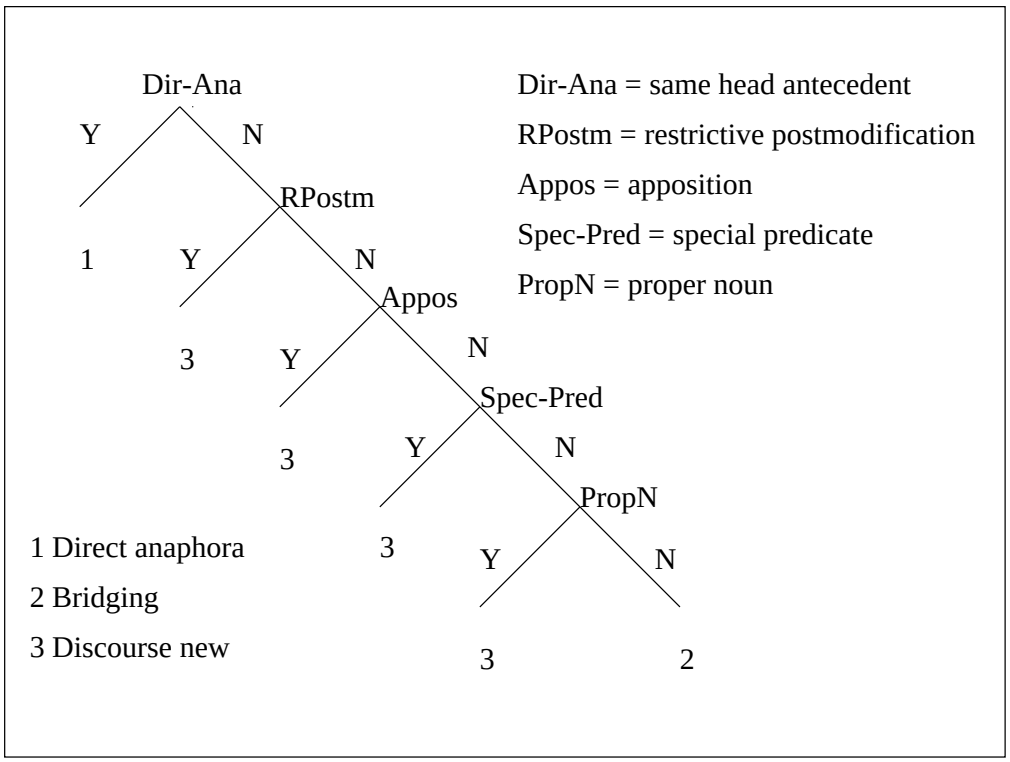

Figure 4.3: Generated Decision Tree

The algorithm generates a decision tree based on the cases given. The resulting decision tree is presented in Figure 4.3. The evaluation of the results are discussed in Chapter 5.

Note that the first distinguishing feature in the decision tree generated by the algorithm is the presence of an antecedent with same head noun; this might be so because it is probably the most regular parameter. The presence of special predicates which we adopted as the first test in our procedure appears as the fourth test in that tree. The order of presentation of the tests are different for the two trees but the answers for each test are basically the same. In Section §5.7 we present the evaluation of the generated algorithm and compare the results with the results of the algorithm we presented previously.

We have presented in this chapter our heuristics to deal with definite descriptions in domain independent written discourse. We ran a series of tests of these heuristics varying some of their parameters in order to arrive at a version of the system that resolves the maximum number of cases, with the least possible number of errors. In the next chapter we present our tests and evaluation of the heuristics just presented.

[23] We exclude from the parameters verification of restrictive premodification and copula constructions, since these parameters presented the poorest results when the heuristics were evaluated (see Chapter 5).

Chapter 5

Evaluation of the System

In this chapter we discuss the tests we ran to arrive at a final configuration of the system and we present an evaluation of the results obtained by our prototype. The performance of the heuristics was verified against the human annotation of the corpus presented in Chapter 3. Several versions of our heuristics were tried with respect to the training data, the corpus from the first experiment. After deciding upon an optimal version, our algorithms were evaluated against the test data, the corpus from the second experiment.

Different forms of evaluation are presented. First, we computed recall and precision figures on the basis of a standard annotation of the corpora (discussed in Section §5.1.2; as a second form of evaluation, we calculated the coefficient of agreement (K) among coders and the system and compared it with the agreement of coders among themselves. We also compared the results of the algorithm we developed by hand with the results produced by the algorithm learned by induction, according to Quinlan's ID3 (Quinlan, 1993).

This chapter is organized in the following way. In Section §5.1 we explain the evaluation methods we adopted. Before presenting the evaluation of each of our heuristics, we present in Section §5.2 the results of a previous prototype (Vieira and Poesio, 1997), that we call version 1. After that we developed a new version of the system: in Section §5.3.1, we present a comparative analysis of alternative heuristics dealing with the resolution of same head anaphora which guided our choice of an optimal version (version 2). In Section §5.3.2 we present an evaluation of our heuristics for identifying discourse new descriptions. Then we present the overall results of version 2 of our system in Section §5.4. After that, in Section §5.5.1 we analyse our heuristics for bridging descriptions and in Section §5.6 we present the overall results of another version of the system which includes these heuristics, version 3. Our algorithms are compared to the learned algorithm in Section §5.7. Finally, in Section §5.8 we present another form of evaluation: the coefficient of agreement K among coders and the system.

5.1 Evaluation methods

5.1.1 Recall and precision

Recall and precision are measures commonly used in Information Retrieval to evaluate systems' performance. Recall is the percentage of correct answers reported by the system in relation to the number of cases indicated by the annotated corpus:

$$R = \frac{number\ of\ correct\ responses}{number\ of\ cases}$$

Precision is the percentage of correctly reported results in relation to the total reported.

$$P = \frac{number\ of\ correct\ responses}{number\ of\ responses}$$

These two measures may be combined to form one measure of performance, the F measure, which is computed as follows:

$$F = \frac{(W+1)RP}{(WR)+P}$$

W represents the relative weight of recall to precision, and typically it has the value 1. A single measure gives us a balance between the two results; 100% of recall may be due to a precision of 0% and vice-versa. The F measure penalises both very low recall or precision.

5.1.2 Standard annotation

The system and its variants were evaluated by comparing the classification they produced with our standard corpus annotation. The standard annotation aims at capturing majority judgement: as we had 3 different coders, whenever at least two of them agreed on a class that class was adopted.[1] The standard annotation was obtained by merging the coders' annotations as described below.

Standard Annotation of the Training Data

Our standard annotation of corpus 1 was obtained out of the annotations made by three different coders: annotator A, annotator B and the author (see Chapter 3). If at least two coders agreed on a class, this class was chosen for the standard annotation. The cases of total disagreement were analysed by the author: if no errors were observed we just respected the preference ranking indicated in the instructions. The results of the standard annotation of the training data are shown in Table 5.1.

[1] An alternative way of doing this would be to give fractional values for a classification depending on the number of agreements, as done for instance in (Hatzivassiloglou and McKeown, 1993).

Class	#	%
I. Direct anaphora	312	30%
II. Bridging	204	20%
III. Discourse new	492	47%
IV. Idiom	22	2%
V. Doubt	10	1%
Total	1040	100%

Table 5.1: Standard classification of definite descriptions - training data

Identification of the antecedent The standard classification enables us to assess the performance of the system with regards to the classification task only. In order to test whether the system actually picks up a correct antecedent, the author identified the antecedent for those cases classified as direct anaphora in the standard annotation.

Standard Annotation of the Test Data

The three annotations of the second corpus were also merged into a standard annotation against which our computational experiments could be evaluated. The compilation followed the general guidelines of the standard annotation of the first corpus. Co-referential descriptions, however, were all checked, and those based on a different head noun from their antecedents were marked as bridging instead of co-referential (the whole co-referential chain was checked). Co-referential descriptions for which a previous same head noun phrase was included in the co-referential chain were marked as direct anaphora. This was done in order to be able to evaluate our heuristics for resolving direct anaphora. The results of the standard annotation of the test data are shown in Table 5.2.

Class	#	%
I. Direct anaphora	154	33%
II. Bridging	81	17.5%
III. Discourse new	218	47%
V. Doubt	11	2.5%
Total	464	100%

Table 5.2: Standard classification of definite descriptions - test data

Considering the difference between the average relative importance of the co-referent class in the second experiment (44%) and the direct anaphora class (same head antecedent only) resulting from the standard annotation (33%), it is estimated that approximately 11% of definite descriptions were co-referential and had a different head from their antecedents.

Identification of the antecedent To evaluate the resolution of anaphoric descriptions on the test data, we considered the co-referential chains indicated by the three coders, and checked the system's responses with them. If the antecedent found by the system was indicated by any of the coders as the right antecedent, we considered that response to be right. This was done based on the fact that the annotation of the test data resulted in complete agreement on antecedents for 95% of the cases.

5.1.3 Automatic evaluation

The system's results were automatically evaluated against the standard annotation of the corpus, in the way described below.

When a text is processed, each NP is ascribed an index number:

(5.1) A house^{106}... The house^{135}...

When a text is processed/annotated the system/coder associates each description index with a type of use. The system's results and the standard annotation are represented by prolog assertions, each one into a different file which can then be compared:

(5.2) a. `system: dd_class(135,anaphoric).`

b. `coder: dd_class(135,anaphoric).`

This enabled us to evaluate the system as a classifier. In order to assess the system's performance with regards to the identification of a co-referential antecedent, we needed to compare the links that indicate the antecedent of each description classified as anaphora. These links were represented by prolog assertions, as follows:

(5.3) a. `coder: coref(135,106).`

b. `system: coref(135,106).`

When comparing an antecedent indicated by the system with the one in the annotated corpus, the corresponding co-reference chain is checked (Section §4.3.5). The system's indexes and the annotated indexes do not need to be exactly the same as long as they belong to the same co-reference chain. In this way, both (5.5.a) and (5.5.b) would be evaluated as correct answers if the corpus is annotated with the links shown in (5.4) .

(5.4) A house^{106}... The house^{135}... The house^{154}...

```
coder: coref(135,106).
coder: coref(154,135).
```

(5.5) a. `system: coref(154,135).`

b. `system: coref(154,106).`

Because our experiments did not involve the annotation of all types of anaphoric expressions, an element outside an annotated co-reference chain, such as a bare noun or possessive, may be indicated as an antecedent by the coder or system:

(5.6) A house^{106}... The house^{135}... His house^{140}... The house^{154}...

```
coref(154,140).
```

If the NP (135) is indicated as the antecedent for NP (154) in the corpus annotation and the system indicates (140) as the antecedent for (154), an error is reported by the automatic evaluation, even though all of these NPs refer to the same entity. The problem is that the NP (140) is not included in the chain. Consequently, the system's errors in finding antecedents indicated by the automatic evaluation had to be manually checked. A second consequence of this problem is that in the evaluation of direct anaphora resolution we only verify if the antecedents indicated are correct; we do not evaluate how complete the co-referential chains are.[2]

However, even the limited notion of co-reference chains adopted here was very helpful in the automatic evaluation, reducing considerably the number of cases to be checked manually.

5.2 A previous prototype

In (Vieira and Poesio, 1997) we presented our first prototype and a preliminary analysis of the effectiveness of our heuristics. Our first prototype was directly related to Heim's ideas (Heim, 1982): of all definite descriptions, only those which were not resolved with a same head antecedent were considered as potential antecedents. Resolved definite descriptions would be linked to previous NPs, but would not be made available for subsequent resolution; the idea was that the same antecedent used in one resolution

[2]For the coreference task in the MUC-6 competition, which considers all types of referring expressions, the resulting co-reference chains are evaluated instead of just the indicated antecedent (Vilain et al., 1995).

could be used to resolve all subsequent mentions co-specifying with that definite description. The algorithm of our first prototype follows the description given in Section §4.6.

The results of those experiments are presented in Tables 5.3 and 5.4. They were obtained by processing the corpus from experiment 1, consisting of 1040 definite descriptions.

In Table 5.3 the column headed by (#) shows the total number of cases for each class[3] according to the standard annotation; in the column headed by (+) we listed the number of correct results; in (-) we listed the number of errors; R is recall; P is precision; and F represents the combined recall and precision measures, considering a relative weight W of recall to precision equal to 1.

System's tasks	#	+	-	R	P	F
A. Direct anaphora (classification)	312	244	29	78%	89%	83%
B. Direct anaphora (resolution)	312	225	48	72%	82%	77%
C. Discourse new	492	364	64	74%	85%	79%
Overall	1040	589	112	57%	84%	68%

Table 5.3: Evaluation of the first prototype

There is a difference in the results of anaphora classification and anaphora resolution. When the system finds a wrong antecedent, and there is one that is right, the classification as anaphora is correct, but the resolution is not. The overall performance of the system is measured on the basis of the total number of right answers for anaphora resolution and for the identification of discourse new descriptions. The descriptions not classified by the system (339 out of 1040) were distributed into different classes, as shown in Table 5.4. Most of the cases not classified by the system were bridging descriptions.

Non classified	#	% of total
Direct anaphora	39	11%
Bridging	161	48%
Discourse new	114	34%
Idiom	20	6%
Doubt	5	1%
Total	339	100%

Table 5.4: Distribution of descriptions not classified by the system

In the rest of the chapter we discuss further tests regarding anaphora resolution that motivated our version 2 of the system. We also present a detailed analysis of the results of our heuristics for identifying discourse new descriptions and the overall results of version 2. We also discuss a version of the system dealing with bridging descriptions (version 3).

[3]Besides, direct anaphora (312) and discourse new (492), another 236 descriptions were classified as bridging, idiom or doubt, totalling 1040.

5.3 Version 2

5.3.1 Anaphora resolution

In this section we discuss the performance of alternative heuristics dealing with the resolution of direct anaphora. We show the evaluation of our heuristics for segmentation, selection of potential antecedents and premodification. The optimal version of our system is based on the best results we could get for resolving same head anaphora, because we wanted to establish the co-referential relations among discourse NPs as precisely as possible.

An important difference between our first prototype and version 2 is the way we treated definite descriptions for further resolution. As said earlier, our first prototype considered only definite descriptions which were not resolved as potential antecedents. Resolved definite descriptions would be linked to previous NPs, but would not be made available for subsequent resolution. In version 2 we abandoned that approach, and the definites resolved by the system are also made available for the resolution of subsequent definites. The reason for this was that in our previous prototype an error in identifying an indefinite antecedent was sometimes propagated through a co-referential chain, and the right antecedent was missed.

In the next sections we present the tests we used to choose the heuristics for resolving anaphoric descriptions used in version 2 of the system.

Segmentation and recency

Our first set of tests shows the effects of considering different window sizes. Note that the restriction on sentence distance is relaxed (i.e., the resolver will consider an antecedent outside the window) when either:

- the antecedent is itself a subsequent mention; or
- the antecedent is identical to the definite description being resolved (including the article).

This relaxed rule was called “loose segmentation heuristic”.

This segmentation heuristic allows the assertion of more than one coreference link to the same description (all antecedents attending the requirements will be indicated as a possible antecedent); therefore, when evaluating the system’s results we may find that all antecedents indicated for the resolution of a description were right, or some were right and some wrong, or that all were wrong. The recall and precision figures reported here relate to those cases were all resolutions indicated were right according to the annotated corpus.

We compared the loose segmentation heuristic with another approach, which we called "recency".[4] The results are shown in Table 5.5.

In these tests, indefinites (i.e., NPs with determiners a, an, some, or bare and cardinal plurals), possessives and definite descriptions head nouns were considered as a potential antecedent, as we did in our first prototype. We also adopted the same premodification heuristics we had before. Alternatives to these heuristics were also evaluated and are presented later in this section.

Heuristics	R	P	F
Segmentation: 1 sentence window	71.79%	86.48%	78.45%
Segmentation: 4-sentence window	76.92%	82.75%	79.73%
Segmentation: 8-sentence window	78.20%	80.26%	79.22%
Recency: all sentences	80.76%	78.50%	79.62%

Table 5.5: Evaluation of loose segmentation and recency heuristics

The resulting F measures were almost the same for all heuristics. There was clearly an increase in recall with a loss of precision when enlarging the window size. The recency heuristic had the best recall, but the lowest precision; not much lower than the others, however. The best precision was achieved with 1 sentence-window, and recall was not dramatically affected, but this only happened because the window size constraint was relaxed. We present below the results reported when the constraint on window size is strict, which also serves as an illustration of why we adopted the loose segmentation heuristic.

Strict segmentation Table 5.6 shows the results when a strict segmentation heuristic is adopted. Strict segmentation means that the system only considers those antecedents that are inside the sentence-window for resolving a description, with no exceptions. As the table shows, this form of segmentation presents higher precision but has a strong effect on recall. The overall results (F) are all worse than in the previous tests.

Strict segmentation	R	P	F
1 sentence window	29.48%	89.32%	44.33%
4 sentence window	57.69%	88.23%	69.76%
8 sentence window	67.94%	84.46%	75.31%

Table 5.6: Evaluation of the strict segmentation heuristic

Combining segmentation and recency Finally, we tried a combination of the recency and segmentation heuristics: just one potential antecedent for each different head noun

[4]We recall that by recency we mean that we do not collect all candidate NPs as potential antecedents but we keep only the last occurrence of an NP from all those having the same head noun, and there are no restrictions regarding the antecedent's distance.

is available for resolution, it is always the last occurrence of that head noun. The resolution still respects the segmentation heuristic (loose version). The results are presented in Table 5.7.

Combined heuristics	R	P	F
4 sentences + recency	75.96%	87.77%	81.44%
8 sentences + recency	77.88%	84.96%	81.27%

Table 5.7: Combining loose segmentation and recency heuristics

By combining these two heuristics we obtained a better trade-off between recall and precision with respect to the alternatives presented in Table 5.5. We chose the version with higher recall (4 sentence-window plus recency) to perform further tests presented in the next sections.

Potential antecedents

We also evaluated various ways of choosing the set of potential antecedents to be taken into account. We tested the impact of different selections of potential antecedents together with the segmentation heuristic chosen above (4 sentence-window with recency). The results are shown in Table 5.8.

Antecedents selection	R	P	F
Indefinites, definites, and possessives	75.96%	87.77%	81.44%
All NPs	77.88%	86.17%	81.81%
Indefinites and definites	73.39%	88.41%	80.21%
Indefinites only	12.17%	77.55%	21.05%

Table 5.8: Evaluation of the collection of potential antecedents

If we only consider indefinites as potential antecedents the recall is extremely low (12%), and the precision is also the worst. In other words, considering only indefinites for the resolution of definite descriptions is too restrictive (this fact is also observed in (Fraurud, 1990)).

The alternative with the highest precision (88%) is the one that only considers indefinites and definite descriptions as antecedents, but the recall is lower compared to that which considered other NPs. We chose the alternative which combines higher F measure and precision, which is the one in the first row in Table 5.8 to use in our last test concerning anaphoric descriptions, which deals with premodifiers.

Premodifiers

Our tests with premodifiers were first presented in (Vieira and Poesio, 1997). At that time we arrived at the following heuristic matching algorithm, discussed in Section §4.3.4:

1. allow an antecedent to match with a definite description if the premodifiers of the description are a subset of the premodifiers of the antecedent;
2. allow a non-premodified antecedent to match with any same head definite.

We show in Table 5.9 the results for these heuristics in that first prototype. Our heuristics are called:

1. Ant-set/Desc-subset (the description's premodification is a subset of the antecedent's premodification);
2. Ant-empty (the antecedent has no premodification).

The highest precision is achieved when no new information is allowed in the anaphoric description, i.e., the modification of the description is a subset of the modification contained in the antecedent (Ant-set/Desc-subset), which is in accordance with the intuition behind Sidner's co-specification rule 1 (see Section §2.4.2). But the best overall results of recall and precision are achieved by also allowing a non-modified antecedent to be further complemented by new information in the anaphoric description.

Antecedents selection	R	P	F
Ant-set/Desc-subset & Ant-empty	72.43%	82.48%	77.13%
Ant-set/Desc-subset	67.94%	83.79%	75.04%
None	73.07%	76.51%	74.75%

Table 5.9: Evaluation of the heuristics for premodification (version 1)

We repeated these tests with version 2 of the system. The results are presented in Table 5.10. We consider here a third rule allowing a pre-modified antecedent to match with a definite whose set of pre-modifiers is a superset of the set of modifiers of the antecedent (an elaboration of rule 2). This new heuristic is called Ant-subset/Desc-set. We tested each of the heuristics alone and their combinations. Note that the combination of heuristics 2 and 3 is equivalent to heuristic 3 alone (rule 3 subsumes rule 2). Heuristic 2 and 3 alone are counter-intuitive and indeed give the poorest results; however, the impact is greater on recall than precision, which suggests that the introduction of new information in the noun modification is just less frequent rather than commonly used to refer to a different entity.

The fact that we are using recency has reduced the impact of the heuristics for premodification on the performance of the algorithm with respect to the first prototype. The best

Antecedents selection	R	P	F
1. Ant-set/Desc-subset	69.87%	91.21%	79.12%
2. Ant-empty	55.12%	88.20%	67.85%
3. Ant-subset/Desc-set	64.74%	88.59%	74.81%
1 and 2 (basic v.)	75.96%	87.77%	81.44%
1 and 3	75.96%	87.13%	81.16%
None	78.52%	81.93%	80.19%

Table 5.10: Evaluation of the heuristics for premodification (version 2)

precision is still achieved when no new information is accepted in the anaphoric expression, but the best results overall are again obtained with the choice of heuristics adopted in our optimal version, although either 2 or 3 works equally well when combined with 1.

Finally, in Table5.11 we present the results of anaphora classification and anaphora resolution[5] with the basic version (version 2) of the system for both training and test data.

Anaphora classification	#	+	-	R	P	F
Training data	312	243	27	78%	90%	83%
Test data	154	103 12		67%	90%	77%
Anaphora resolution	#	+	-	R	P	F
Training data	312	237	33	76%	88%	81%
Test data	154	96	19	62%	83%	71%

Table 5.11: Evaluation of the heuristics for direct anaphora (version 2)

The recall figures are much lower for the test data. By analysing the system's output we noted that several cases were missed due to segmentation. We then repeated the tests with some of the variations of the algorithm, to see if a different version would produce better results with the test data. As shown in Table 5.12, adopting a 8 sentence-window produced an increase in recall and a better F figure for the test data, but with a decrease in precision. Considering all NPs produces a negative effect. Not considering the heuristics for premodification resulted in better recall but worse precision. The basic version produces the highest precision. The evaluation of other heuristics and the analysis of the global results presented in the next sections are therefore based on this basic version.[6]

To summarize, version 2, the optimal version we arrived at through our tests, includes the following heuristics:

[5] Anaphora classification differs from anaphora resolution, when the system finds a wrong antecedent, the classification as anaphora may be correct but the resolution is wrong.

[6] In our experiments small differences in recall, precision and F measures are frequent. Further statistical investigation, as done for instance in (Chinchor, 1995), might shed light on the actual significance of these differences.

Alternative versions	R	P	F
1. version 2	62%	83%	71%
2. v.2 but 8 sentence-window	67%	79%	73%
3. v.2 but all NPs	61%	80%	69%
4. v.2 but no premodification	68%	74%	71%

Table 5.12: Evaluation of alternative heuristics for the test data

1. combined segmentation and recency,
2. 4 sentence-window,
3. considering indefinites, definites and possessives as potential antecedents,
4. the premodification of the description must be contained in the premodification of the antecedent or else the antecedent has no premodifiers.

Before presenting the evaluation of heuristics for recognizing discourse new descriptions, we discuss some examples of errors in the resolution of anaphoric descriptions with version 2.

Examples of errors in anaphora resolution

The errors in direct anaphora resolution have more than one cause. The system may fail to classify a direct anaphora as such, or it may classify as direct anaphora descriptions classified differently in the standard annotation, or it may just get the wrong antecedent. Examples are given below.

Some errors are caused by the selection of potential antecedents. If we select indefinites, definite descriptions and possessives, we miss proper names such as *Toni Johnson* in (5.7). The following definite description is then classified by the system as larger situation/unfamiliar.

(5.7) *Toni Johnson* pulls a tape measure across the front of what was once a stately Victorian home.
...
The petite, 29-year-old Ms. Johnson...

A noun coordination such as the one in (5.8) is recognized in the human annotation but is missed by the system:

(5.8) The owners, *William and Margie Hammack*, are luckier than many others.
...
The Hammacks...

Some errors are caused by the premodification heuristics. Examples in which they prevent the system from finding the right antecedent include the rare cases of co-referent descriptions that present different premodification, such as (5.9).

(5.9) *The Victorian house* that Ms. Johnson is inspecting has been deemed unsafe by town officials.
...
Once inside, she spends nearly four hours measuring and diagramming each room in *the 80-year-old house*.

Misspelling in the Treebank also causes some errors, as seen in the example below.

(5.10) *A Lorillard spokewoman... The Lorillard spokeswoman*

But the most common problems are the cases in which the system fails to find the antecedent due to the segmentation heuristics, such as those shown in example (5.11) (the number of each sentence is indicated).

(5.11) 7. She has been on the move almost incessantly since last Thursday, when *an army of adjusters*, employed by major insurers, invaded the San Francisco area.
...
30. Aetna, which has *nearly 3,000 adjusters*, had deployed about 750 of them
...
53. Many of *the adjusters* employed by Aetna and other insurers

In the following example the system classifies the description as anaphoric, but it is discourse new according to the standard annotation.

(5.12) Most companies still are trying to sort through the wreckage caused by Hurricane Hugo in the Carolinas last month.
Aetna, which has nearly 3,000 adjusters, had deployed about 750 of them in Charlotte, Columbia, and Charleston.
Adjusters who had been working on the East Coast say the insurer will still be processing *claims from that storm* through December.
It could take six to nine months to handle *the earthquake-related claims*.

Finally, an example of right classification but wrong resolution (i.e., wrong antecedent) is given in (5.13). The system suggests as antecedent *an income tax law* whereas the *a money lending law* is more correct.[7]

(5.13) Nearly 20 years ago, Mr. Morishita, founder and chairman of Aichi Corp., a finance company, received a 10-month suspended sentence from a Tokyo court for violating *a money-lending law* and *an income tax law*.
He was convicted of charging interest rates much higher than what *the law* permitted, and attempting to evade income taxes by using a double accounting system.

5.3.2 Identification of discourse new descriptions

Discourse new	#	+	-	R	P	F
Training data	492	368	60	75%	86%	80%
Test data	218	151	58	69%	72%	70%

Table 5.13: Evaluation of the heuristics for identifying discourse new descriptions

The recall and precision results for identification of discourse new descriptions, according to version 2, following the heuristics presented in Section §4.4 are shown in Table 5.13, for both the training and the test data. The column headed by (#) represents the number of cases of descriptions classified as discourse new in the standard annotation.

The performance of each of the heuristics[8], separately, according to the training data, is presented in Table 5.14 (larger situation uses) and Table 5.15 (unfamiliar uses). Here, only precision figures can be reported, since we do not have a manual counting of how many of each type is actually present in the corpus. The most common feature of discourse new descriptions is postmodification. The least satisfactory results are those for proper names in premodification. In general, heuristics for recognising unfamiliar uses present better precision than those for larger situation uses.

Tables 5.16 and 5.17 summarise the results of the heuristics for discourse new descriptions for the test data. The heuristics gave similar results for both corpora; the best results were obtained by the heuristics related to unfamiliar uses. Copula constructions gave good precision in the training data, but poor precision for the test data. As a very low recall was reported for both training and test data, the actual performance of that heuristic is difficult to evaluate.

[7] However, if the expression *the law* is interpreted as referring to "the law system in general", any of the antecedents would be correct.

[8] These figures represent an estimated evaluation since the heuristics are applied in a deterministic sequential order and sometimes more than one may apply to the same instance of definite description, but the system only counts the first feature it finds.

Larger Situation	Total found	Errors	Precision
Names	73	10	86%
Time references	50	7	86%
Premodification	41	19	54%
Total	164	36	78%

Table 5.14: Evaluation of heuristics for larger situation uses (training data)

Unfamiliar	Total found	Errors	Precision
NP compl/Unexp mod	32	2	93%
Apposition	27	2	92%
Copula	8	2	75%
Postmodification	197	18	91%
Total	264	24	91%

Table 5.15: Evaluation of heuristics for unfamiliar uses (training data)

Larger Situation	Total found	Errors	Precision
Names	44	14	68%
Time references	21	5	64%
Premodification	17	9	47%
Total	82	28	66%

Table 5.16: Evaluation of heuristics for larger situation uses (test data)

Unfamiliar	Total found	Errors	Precision
NP compl/Unexp mod	16	2	87%
Apposition	10	2	80%
Copula	6	4	33%
Postmodification	95	22	77%
Total	127	30	76%

Table 5.17: Evaluation of heuristics for unfamiliar uses (test data)

Examples of errors in identifying discourse new descriptions

Apposition A problem with this heuristic is encountered when processing the coordinated NP in the sentence: *G-7 consists of the U.S., Japan, Britain, West Germany, Canada, France and Italy*. The problem is that this coordinated NP in the Treebank has the same structure that is checked by the system for appositions, as shown in (5.14).

(5.14) `[NP,[NP,the,U.S.],,,[NP,Japan],,,[NP,Britain],,,[NP,West, Germany],,,[NP,Canada],,,[NP,France],and,[NP,Italy]]`

Copula This heuristic was difficult to evaluate because it has shown very low recall and very different precision for the two data sets (see Tables 5.15 and 5.17 above). One problem is that the descriptions in copula constructions may as well be bridging references. For instance, the description *the result* in (5.15.a) below is the result of something mentioned previously, while the copula construction specifies its referent.

Other ambiguous examples are (5.15.b) and (5.15.c):

(5.15) a. *The result* is that those rich enough to own any real estate at all have boosted their holdings substantially.

b. *The chief culprits*, he says, are big companies and business groups that buy huge amounts of land not for their corporate use, but for resale at huge profit.

c. *The key man* seems to be the campaign manager, Mr. Lynch.

Restrictive premodification One problem with this heuristic is that although often proper nouns in premodifier positions are used with discourse new definites (e.g., *the Iran-Iraq war*), they may also be used as additional information in associative or anaphoric uses:

(5.16) Others grab books, records, photo albums, sofas and chairs, working frantically in the fear that an aftershock will jolt *the house* again.
...
As Ms. Johnson stands outside *the Hammack house* after winding up her chores there, the house begins to creak and sway.

Restrictive postmodification If the system fails to find an antecedent or anchor and the description is postmodified, it is classified wrongly as discourse new. In (5.17) *the filing on the details of the spinoff* was classified as bridging on *documents filed ...* by the coders, but the system classified it as discourse new.

(5.17) *Documents filed with the Securities and Exchange Commission on the pending spinoff* disclosed that Cray Research Inc. will withdraw the almost $100 million in financing it is providing the new firm if Mr. Cray leaves or if the product-design project he heads is scrapped.
...
The filing on the details of the spinoff caused Cray Research stock to jump $2.875 yesterday to close at $38 in New York Stock Exchange composite trading.

Proper noun As we have already seen—(5.7), repeated below (5.18)— a proper noun may be anaphoric, but if the antecedent is missed it is classified wrongly as discourse new.

(5.18) *Toni Johnson* pulls a tape measure across the front of what was once a stately Victorian home.
...
The petite, 29-year-old Ms. Johnson...

Special predicates In this example the system classifies as discourse new a time reference which is classified as bridging in the standard annotation.

(5.19) Newsweek's circulation for *the first six months of 1989* was 3,288,453, flat from the same period last year.
U.S. News' circulation in *the same time* was 2,303,328, down 2.6%.

5.4 Overall results of version 2

We present in this section the overall results of our optimal version dealing with direct anaphora and discourse new descriptions, version 2. We will discuss later version 3 that also handles bridging descriptions.

```
NR. OF TEXTS: 20       NR. OF NOUN PHRASES: 6831

NR. OF ANTECEDENTS CONSIDERED: 2911
       Indefinites:            1569
       Possessives:             388
       Definites:               954

NR. OF DEFINITE DESCRIPTIONS: 1040

DIRECT ANAPHORA: 270  ANTECEDENTS FOUND: Indefinites: 49
                                         Possessives:  9
                                         Definites:  212
DISCOURSE NEW DESCRIPTIONS: 428

 LARGER SITUATION  USES: 164       UNFAMILIAR  USES    : 264
      NAMES             :  73           NP COMP./UN.MOD.:  32
      TIME REFERENCES   :  50           APPOSITIONS     :  27
      REST.PREMOD.      :  41           REST. POSTMOD.  : 197
                                        COPULA          :   8
NON-IDENTIFIED: 342

TOTAL ESTIMATED ERRORS (for anaphora classification)    : 27
TOTAL ESTIMATED ERRORS (for anaphora resolution)        : 33
TOTAL ESTIMATED ERRORS (for larger situation/unfamiliar): 60
```

Table 5.18: Global results for training data

5.4.1 Training data

Table 5.18 shows the output of the system for the training data: a total of 20 texts were processed, containing 6831 NPs. Almost half of them (2911) were considered as potential antecedents. A total of 1040 descriptions were processed by the system. An antecedent was identified for 270 of them, and the antecedents identified were mostly definites themselves. For 212 out of the 270 definite descriptions classified as anaphoric same-head by the system the antecedent was a definite NP.[9]

Table 5.19 shows a summary of the results reported by the system and those reported by the standard annotation. It also shows how descriptions which were not resolved by the system were classified in the standard annotation. The largest number of descriptions missed by the system were bridging descriptions. In Section §5.5.1 we present the results of our tests for that class.

The final figures for the training data when processed by version 2 of the system are presented in Table 5.20. Note that because a large number of definite descriptions is not classified, the overall recall is only 59%, even though the recall for both anaphoric and discourse new descriptions is much higher.

[9] According to the annotation of one of our coders (not the system's output), for 312 anaphoric descriptions we had 164 co-referential chains. Taking the first NP of a coref chain, we observed that there were 86 chains initiated by definite descriptions.

```
TOTAL TYPES IDENTIFIED BY THE SYSTEM
   anaphoric: 270
   larger sit./unfam: 428
   total: 698

TOTAL NON CLASSIFIED
   anaphoric: 41
   larger sit./unfam: 113
   associative: 162
   idiom: 20
   doubt: 6
   total: 342

TOTAL TYPES CLASSIFIED BY HAND
   anaphoric: 312
   larger sit./unfam: 492
   associative: 204
   idiom: 22
   doubt: 10
   total: 1040
```

Table 5.19: Summary of the results for training data

System's tasks	R	P	F
Anaphora classification	78%	90%	83%
Anaphora resolution	76%	88%	81%
Discourse new	75%	86%	80%
Overall	59%	88%	70%

Table 5.20: Global results for training data

5.4.2 Test data

The system was evaluated against the test data used in our second annotation experiment. The results are shown in Tables 5.21 and 5.22. The recall and precision figures of the system's performance over the test data are presented in Table 5.23. This set contained 14 texts and 2990 NPs. Again, almost half of the NPs were considered as potential antecedents. There were 464 definite descriptions processed by the system, 115 were classified as direct anaphora and 88 antecedents were definites themselves. As before, there were just a few more errors in anaphora resolution than in anaphora classification. Overall recall for the test data was 53% and precision was 76%.

A difference observed between the results of the two data sets is related to the distribution into classes of those descriptions that the system fails to classify. In the first corpus, the larger number of cases not classified are bridging descriptions. This proportion is

```
NR. OF TEXTS: 14       NR. OF NOUN PHRASES: 2990

NR. OF ANTECEDENTS CONSIDERED: 1226
       Indefinites:             657
       Possessives:             144
       Definites:               425

NR. OF DEFINITE DESCRIPTIONS: 464

DIRECT ANAPHORA: 115  ANTECEDENTS FOUND: Indefinites: 21
                                         Possessives:  6
                                         Definites:   88
DISCOURSE NEW DESCRIPTIONS: 209

 LARGER SITUATION  USES: 82       UNFAMILIAR  USES     : 127
      NAMES            : 44            NP COMP./UN.MOD.:  16
      TIME REFERENCES  : 21            APPOSITIONS     :  10
      REST.PREMOD.     : 17            REST. POSTMOD.  :  95
                                       COPULA          :   6
NON-IDENTIFIED: 140

TOTAL ESTIMATED ERRORS (for anaphora classification)    : 12
TOTAL ESTIMATED ERRORS (for anaphora resolution)        : 19
TOTAL ESTIMATED ERRORS (for larger situation/unfamiliar): 58
```

Table 5.21: Global results for test data

```
TOTAL TYPES IDENTIFIED BY THE SYSTEM
   anaphoric: 115
   larger sit./unfam: 209
   total: 324

TOTAL NON CLASSIFIED
   anaphoric: 29
   larger sit./unfam: 61
   associative: 46
   doubt: 4
   total: 140

TOTAL TYPES CLASSIFIED BY HAND
   anaphoric: 154
   larger sit./unfam: 218
   associative: 81
   doubt: 11
   total: 464
```

Table 5.22: Summary of the results for test data

System's tasks	R	P	F
Anaphora classification	67%	90%	77%
Anaphora resolution	62%	83%	71%
Discourse new	69%	72%	70%
Overall	53%	76%	63%

Table 5.23: Evaluation of the system according to the test data

not repeated for the second corpus, where the larger number of cases not classified are discourse new.

5.4.3 Multiple classification

As a further test, we ran a modified version of the system which performs multiple classification: descriptions were classified as ambiguous if both:

- a same head antecedent was found and
- one of the following indicators of discourse new use was verified,
  - special predicates,
  - appositive and copula constructions or
  - postmodification.

Ambiguous resolutions would not be taken into account for evaluation. The results thus obtained for anaphora resolution in the training data were recall = 69% and precision = 88%. Compared to the previously obtained figures (76% and 88%), it does not actually show the improvement we expected in precision. In the test data, this modified version presented recall = 61% and precision = 86% (before we had 62% and 83%), here we can see signs of the expected effect.

5.5 Version 3

Version 3 of the system includes also a treatment of bridging descriptions.

5.5.1 Bridging references

Our experiments in corpus annotations showed the bridging class to be the most difficult for humans to agree on. Even when there is agreement on that class, different anchors may be available in the text for the interpretation of a bridging description. This makes the results on this class very difficult to evaluate; also, the results have to be evaluated by hand.

We first tested the heuristics separately, by adding them to our system one at a time. These separate tests, which are more detailed, were manually evaluated. They are based on the training data, the same data that our previous analysis of bridging descriptions (Vieira and Teufel, 1997) was based on. Our heuristics were just experimental; the evaluation was mostly helpful to indicate to us the problems we still have to solve. Nevertheless, we tested the integration of these heuristics into the main system, using both automatic and manual evaluation.

Bridging Class	Relations Found	Right Anchors	% Right
Syn	11	4	36%
Hyp	59	18	30%
Mer	6	2	33%
Sister	30	6	20%
Total	106	30	28%

Table 5.24: Evaluation of the search for anchors using WordNet

Evaluating the use of WordNet

We implemented an automatic search for anchors using WordNet which looks for a semantic relation between descriptions (those 204 descriptions classified as bridging in the standard annotation) and one of the NPs in the previous five sentences. Table 5.24 shows the results of this search over our training corpus. It is interesting to note that the semantic relations found in this automatic search were not always those observed in our manual analysis.

We found that the existence of a semantic relation in WordNet is not a sufficient condition (nor a strong indication) to establish a link between an antecedent and a bridging description. Only about a third of the semantic relations found in WordNet between descriptions and antecedents were right anchors. An example of a semantic relation we found and which does not establish a link is the relation between *argument* and *information* in the sequence below. The description *the argument* relates to the VP *contend* rather than to the NP *information*.

(5.20) A SEC proposal to ease reporting requirements for some company executives would undermine the usefulness of *information* on insider trades as a stock-picking tool, individual investors and professional money managers *contend*.
They make *the argument* in letters to the agency about rule changes proposed this past summer that, among other things, would exempt many middle-management executives from reporting trades in their own companies' shares.

The problem of sense ambiguity was also responsible for some of the false positives. For instance, the noun *company* has at least two distinct senses: *visitor* and *business*. A relation of hypernymy was found between *company* and *human* (its *visitor* sense), whereas in the text the noun *company* was used in the *business* sense.

In another test, we checked if WordNet encoded the semantic relations which were manually identified in the analysis of bridging descriptions, to have an idea of how well WordNet represents those relations. We selected from our two corpora 70 bridging descriptions linked to their anchors via semantic relations of synonymy, hypernymy (hyponymy) and meronymy (holonymy). The results are presented in Table 5.25. We

Bridging Class	Anchor/DD pairs	Found in WN	Found Sister	%
Syn	20	5	2	35%
Hyp	32	17	1	56%
Mer	18	5	2	38%
Total	70	27	5	46%

Table 5.25: Evaluation of the encoding of semantic relations in WordNet

also indicate in that table when we found that the expected relation was not encoded, but the two nouns were sisters in the hierarchy.

The recall figure was quite disappointing, specially for synonymy relations. Some of the problems were due to the use of specific sub-language terminology which sometimes may have context dependent senses, such as *slump, crash* and *bust*, all synonyms in Economics jargon. Sometimes the relations were missing due to WN structure or incompleteness. For instance, in WN the nouns *room, wall, floor* are encoded as part of *building* but not of *house* (Fig. 5.1).

Some of the words we looked for were not in WN, examples are *newsweekly (news-weekly), crocidolite, countersuit (conter-suit)*. Sometimes the word we looked for was encoded but not in the same manner as it was presented in the text; for example we had *spinoff* in a text, whereas WordNet had only an entry for *spin-off*.

In summary, our tests have shown that the knowledge encoded in WordNet is insufficient for the semantic relations expressed in the kind of texts we are dealing with: only 46% of the relations observed were encoded in WordNet. Besides that, only looking for the closest semantic relation is not enough as a heuristic for finding anchors of bridging descriptions.

Evaluating the results for bridging descriptions based on proper names

Identifying named entity types is a pre-requisite for resolving descriptions based on names. Our simple heuristics identified entity types for 66% (535/814) of all names in the corpus (organizations, persons and locations). The precision was 95%. We could have had a better recall if we had adopted more comprehensive lists of cue words, or consulted dictionaries of names as done for the systems participating in MUC-6. There, recall in the named entity task varies from 82% to 96%, and precision from 89% to 97%. The system from Sheffield (Gaizauskas et al., 1995), for instance, used a list of 2600 names of organizations, 94 company designators (Co., Ltd, PLC, etc.), 160 titles (Dr. Mr., etc.), about 500 human names from the Oxford Advanced Learner's Dictionary, 2268 place names (country, province and city names), and other trigger words for location, government institutions and organizations (Golf, Mountain, Agency, Ministry, Airline, etc.). We only used a handful of cue words and WordNet search.

The errors we found were sometimes due to name or sense ambiguity. In the same text a name may refer both to a person and a company, as in *Cray Computers Corp.*

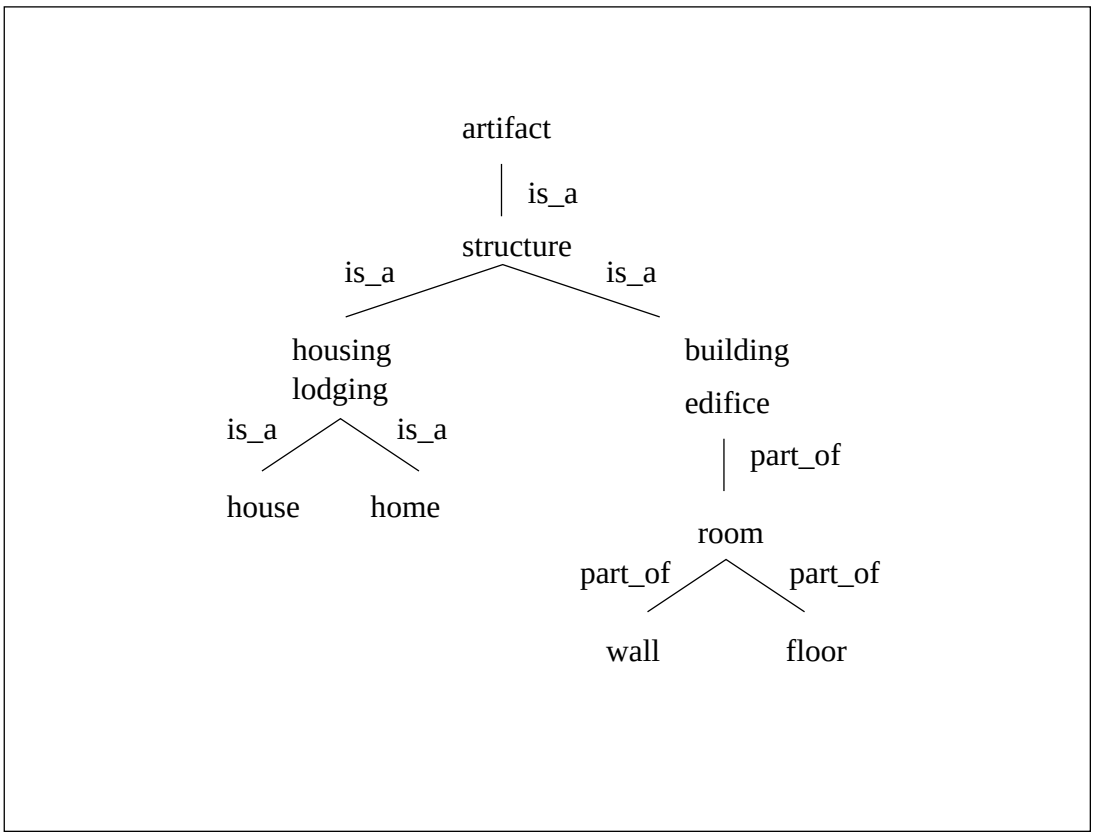

Figure 5.1: WordNet hierarchy

and *Seymour Cray*. When looking for a type in WordNet for the name *Steve Reich* we found the type country for the name *Reich*. (These problems have also been noted by the authors of systems participating in MUC-6 (Appelt et al., 1995)). We also found undesirable relations such as hypernymy for *person* and *company*.

Evaluating the results for bridging descriptions based on compound nouns

We had 25 definite descriptions manually identified as based on compound nouns. For these 25 cases our implemented heuristics achieved a recall of 36% (9/25) but found in some cases valid relations other than the ones we identified. The low recall was due sometimes to the segmentation, since we had a fixed-size window. Sometimes the spelling of the premodification was slightly different from the description, as in *a 15-acre plot... the 15 acres*. Another reason for the low recall was that we have also included in this class those cases in which the head nouns of antecedent and description are identical, but the premodification indicates that the entity referred to is not the same: as in *Italy's unemployment rate... the southern unemployment rate*.

Some errors presented in our tests were due to the lack of part-of-speech tags in the Treebank. If we could force the heuristic to accept only nouns in the premodifier position, we could avoid errors such as: *her second meeting - the second floor*, where the premodification is not a noun but an adjective.

Analysis of bridging descriptions based on VPs

We observed that only 34% of the cases based on events in our corpus could be solved through the technique of nominalization transformations (or alternatively, word truncation) suggested in Section §4.5.4. The remaining 66% require knowledge representation and inference, as shown in the examples below:

(5.21) a. *It went looking for a partner for the Sharpshooter*. " We went to six companies over two days pitching *the prospect*" says Tim Parker, a Santa Fe exploration manager.

b. *The FCC allowed AT&T Co. to continue offering discount phone services for large-business customers and said it would soon re-examine its regulation of the long-distance market. The moves* were good news for AT&T.

In (Humphreys, Gaizauskas and Azzam, 1997) a general approach for performing event coreference and for constructing event representation is proposed. They note that recent work on coreference has concentrated on noun phrases or pronouns, but recognize that coreference involving events, expressed via verbs or nominalized verb forms is also common.

Bridging Class	Found by System	False Posit.
Names	12	14
C. Nouns	15	10
WN Rel.	34	76
Total	61	100

Table 5.26: Evaluation of the bridging heuristics all together

5.6 Overall results of version 3

The overall results of the heuristics for bridging descriptions presented in the last section were not very good. Nevertheless, we ran some tests combining the heuristics just described.

Version 3 of our prototype included the heuristics for bridging descriptions. They were applied only to those descriptions which failed to be treated as direct anaphora or discourse new. The heuristics were applied in the following order:

1. proper names,
2. compound nouns,
3. WordNet,

5.6.1 Training data

For the training data the results were manually evaluated (the anchors suggested by the system were analysed) and the results are presented in Table 5.26. In that table we present the number of acceptable anchors and the number of false positives found by each heuristic. Note that the right anchors found by the system do not always correspond to those identified manually. We found fewer bridging relations than the number we observed in the corpus analysis (204); besides, the number of false positives produced by such heuristics is almost twice the right answers.

5.6.2 Test data

Our version 3 was tested over the test data using the automatic evaluation. In this case, the system was only evaluated as a classifier, the anchors found were not analysed. A total of 57 bridging relations were found, but only 19 of them had been classified as bridging descriptions in the standard annotation. Compared to version 2 of the system, which does not resolve bridging descriptions, version 3 presents higher recall but lower precision, as shown in Table 5.27.

System's versions	R	P	F
V.2 Overall	53%	76%	62%
V.3 Overall	57%	70%	62%

Table 5.27: Comparative evaluation of the system's versions (test data)

5.7 Evaluation of the automatically learned algorithm

We measured the performance of the learned algorithm described in Section §4.7 and compared it with the algorithm we arrived at by trial and error. The first fourteen texts of corpus 1 (845 descriptions) were used as training data to generate the decision tree. They were fed to the learning algorithm together with the results of the standard annotation for the corresponding texts. We then tested the learned algorithm over the other 6 texts of that corpus (195 instances of definite descriptions).

Two different tests were undertaken:

- first, we gave as input to the learning algorithm all cases classified as direct anaphora, discourse new or bridging (818 in total)[10];
- in a second test, the algorithm was trained only with direct anaphora and discourse new descriptions (639 descriptions); all cases classified as bridging, idiom or doubt in the standard annotation were not given as input in the learning process. This algorithm was then only able to classify descriptions as one of those two classes.[11].

Here we present the results evaluated all together considering the system as a classifier only, i.e., without considering the tasks anaphora resolution and identification of discourse new descriptions separately. The output produced by the learned algorithm is compared to the standard annotation. Recall, precision and the F measure results are all equal in these tests, since the learned algorithm classifies all cases[12]. The tests over 6 texts with 195 definite descriptions resulted in:

- $R = P = F = 69\%$ when the algorithm was trained with three classes;
- $R = P = F = 75\%$, when training with two classes only.

[10]This test produces the decision tree presented in Section §4.7.

[11]The resulting decision tree classifies descriptions with same head antecedent as anaphoric, all the rest as discourse new.

[12]The number of responses is equal to the number of cases, therefore recall is the same as precision and so is the F measure.

The best results were achieved when the algorithm was trained with two classes only. When bridging descriptions were given as input the learned algorithm produced more errors. This suggests that the parameters taken into consideration are not as uniform for bridging descriptions as they are for other classes. Our own algorithm (version 2) was used for the same 6 texts and resulted in 62% recall and 85% precision ($F = 71.70\%$). Overall, our algorithm built by hand presented a lower F measure, due to a lower recall (unlike the learned algorithm it does not classify all instances of definite descriptions), but the precision was higher. However, if we take the class discourse new as a default for all cases of definite descriptions not resolved by the system, our recall and precision go to 77%.

As the generated decision tree takes the test for a same head antecedent as the first test, we modified our algorithm to work in the same way, and tested it with the two corpora. The results with this configuration were:

- $R = 0.75$, $P = 0.87$, $F = 0.80$, for the training data (compared with $R = 0.76$, $P = 0.88$, $F = 0.81$) ;
- $R = 0.59$, $P = 0.83$, $F = 0.69$, for the test data (compared with $R = 0.62$, $P = 0.83$, $F = 0.71$).

The results were about the same, although a slightly better performance was obtained when the tests to identify discourse new descriptions were tried first (as described in Section §4.6).

5.8 Agreement between system and coders

Another way to evaluate the results of the system is to compare the agreement between the coders with the agreement between coders and system. In this section, we present such results concerning our second corpus (the test data) for versions 2 and 3.

To compare the system's classification with the coders' classification of the test data, we had first to transform each coder's annotation of that corpus to an annotation corresponding to the system's output, i.e., an annotation like the one specified in the first experiment, moving from a co-referent/associative distinction to a direct anaphora/bridging one.[13] (This is the reason why the figures presented here concerning the coder's agreement differ slightly from the figures presented in Section §3.3. It is interesting to note that, in this way, the agreement among coders was slightly higher.)

[13] As we did for the compilation of the standard annotation of the test data, we verified whether the co-referent descriptions were referring to a same head antecedent; if so, these cases were classified as direct anaphora. Otherwise, the descriptions were classified as bridging.

5.8.1 Version 2

Measuring Kappa for 3 categories (direct anaphora, bridging, discourse new), between coders only, and taking into account only those cases classified by the system (314 definite descriptions, excluding 10 doubts), the agreement coefficient was K = 0.70. The agreement among the three coders and the system, was K = 0.64.

The most frequent type of disagreement among the system and coders were those cases classified by the system as discourse new and classified by the coders as associative (40 instances of descriptions resulted in this type of disagreement). Examples are presented in (5.22). The antecedents indicated by the coders are presented in parentheses; the descriptions classified as bridging by coder and as discourse new by the system are presented in brackets.

(5.22) a. The Hammacks' own home, also in Los Gatos, suffered (*comparatively minor damage.*)
...
Because of the difficulty of assessing [*the damages caused by the earthquake*], Aetna pulled together a team of its most experienced claims adjusters from around the country.

b. (*New England Electric System bowed out of the bidding for Public Service Co. of New Hampshire*), saying that the risks were too high and the potential payoff too far in the future to justify a higher offer.
...
Wilbur Ross Jr. of Rothschild Inc., the financial adviser to the troubled company's equity holders, said [*the withdrawal of New England Electric*] might speed up the reorganization process.

c. (*The documents*) also said that Cray Computer anticipates needing perhaps another $120 million in financing beginning next September.
...
The filing on [*the details of the spinoff*] caused Cray Research stock to jump $2.875 yesterday to close at $38 in New York Stock Exchange composite trading.

The second most common cases of disagreement were those in which the system classified a description as discourse new and the coders as direct anaphora (24 cases).

(5.23) a. She has been on the move almost incessantly since last Thursday, when (*an army of adjusters*), employed by major insurers, invaded the San Francisco area to help policyholders sift through the rubble and restore some order to their lives.
...
Many of [*the adjusters employed by Aetna and other insurers*] have some experience with construction work or carpentry.

b. Newsweek, trying to keep pace with rival Time magazine, announced new advertising rates for 1990 and said it will introduce (*a new incentive plan for advertisers*).
...
[*The new ad plan from Newsweek*], a unit of the Washington Post Co., is the second incentive plan the magazine has offered advertisers in three years.

There were then cases of disagreement in which the system classified a description as direct anaphora and the coders classified it as bridging (8).

(5.24) a. (*The morbidity rate*)$_{coder}$ is (*a striking finding*)$_{system}$ among those of us who study asbestos-related diseases, said Dr. Talcott.
...
[*The finding*] probably will support those who argue that the U.S. should regulate the class of asbestos including crocidolite more stringently than the common kind of asbestos, chrysotile, found in most schools and other buildings, Dr. Talcott said.

b. Japanese investors nearly single-handedly bought up (*two new mortgage securities-based mutual funds*)$^{coder}_{system}$ totaling $701 million, the U.S. Federal National Mortgage Association said.
...
He said more than 90% of [*the funds*] were placed with Japanese institutional investors.

In the last example above (5.24.b), the antecedent found is the same for system and coder, but the coder classifies it as bridging while the system classifies it as anaphoric. Only 4 cases were classified by the system as direct anaphora and by the coders as discourse new.

If we measure K for 2 categories (direct anaphora and non-anaphoric[14]), for the cases handled by the system, we have K = 0.81 for coders only and K = 0.78 among coders and system. There were 15 cases classified by the system as non-anaphoric and by coders as direct anaphora, and 10 cases in which the system classified a description as direct anaphora and the coder's classified it as non-anaphoric: a total of 25 disagreements. Among the three coders there were 33 cases where one marked a description as anaphoric and the other as non-anaphoric.

If we count two classes only, we may consider the cases not resolved by the system as belonging to the class non-anaphoric, and compute K on the complete set of 447 descriptions (which excludes 17 cases of doubt). In this case we get K = 0.72 for coders and system against 0.78 for coders only. Table 5.28 summarises the figures just presented.

[14]Bridging and discourse new are grouped together into the same class.

K	coders & system	coders only
3 categories (314 dds)	0.64	0.70
2 categories (314 dds)	0.78	0.81
2 categories (447 dds)	0.72	0.78

Table 5.28: Comparison of the K coefficient for coders and system v.1, corpus 2

5.8.2 Version 3

For the third version of the system, which implements some heuristics to resolve bridging descriptions and therefore classifies a definite description into one of three classes, the coefficient of agreement was lower than for version 2, which classifies into one of two classes. The number of cases treated is larger (370 against 314) but the resulting agreement is K = 0.61 (previously K = 0.64). The agreement among coders over these 370 instances is K = 0.70 (there were 11 cases of doubt). The results are summarized in Table 5.29.

Kappa	coders & system	coders only
3 categories (370 dds)	0.61	0.70

Table 5.29: Comparison of the K coefficient for coders and system v.2, corpus 2

5.8.3 Discussion

If only two categories are considered, and only the data handled by the system is taken into account, we have little difference[15] in the agreement among the coders compared to the agreement among coders and systems, and in both cases there are considerably consistent agreement levels. This indicates that our system's disagreements with human judgements is not worse than the disagreement observed among humans themselves when resolving direct anaphora, at least when the subset it handles is considered. Such a system could be used in a tool for semi-automatic annotation of co-reference to identify some of the co-reference relations without introducing too many errors.

The worse agreement results, when the whole set of descriptions is considered are due to the low recall for anaphora resolution: 62% in the test data. When trying to distinguish three different classes, the system performance gets rather worse, and these problems with the third class (bridging) are also observed among coders. The figures indicate that there is less disagreement for those cases that the system handles, since the coders' agreement in the subset handled by the system is higher.

[15]The actual significance of this difference might be found with the help of further statistical investigation.

Chapter 6

Conclusions

We have presented a domain independent system for definite description interpretation whose development was based on a study of definite description use that included multi-annotator experiments, and on corpus-based experiments in which we compared the performance of alternative heuristics. To our knowledge, our studies have been the first to consider inter-annotator agreement for the analysis of definite description use.

We reviewed the linguistic literature on definite description use and found several related classifications which we have compared and simplified. Our simplified classification was used to annotate a corpus consisting of 33 Wall Street Journal articles with approximately 1400 definite descriptions. We presented two corpus annotation experiments which adopted slightly different methods. Our results have confirmed earlier findings that first mention definite descriptions are very common (Fraurud, 1990). According to our corpus studies, about 50% of definite descriptions analysed were new in the discourse. This gave us evidence that definite descriptions are not primarily anaphoric; they are often used to introduce a new entity in the discourse.

We implemented and tested a system which deals with different uses of definite descriptions: in the model that we proposed, recognising discourse new descriptions plays a role as important as identifying the antecedent of those used anaphorically.

For direct anaphora, we evaluated the effect of shallow heuristics dealing with recency, segmentation and noun modification. Our system achieved 62% recall and 83% precision for direct anaphora resolution on test data. For identifying discourse new descriptions, the largest class in our corpus, we have exploited the correlation between syntax and type of use noted by Hawkins (1978) and semantically explained by Löbner (1985). Our system achieved 69% recall and 72% precision for this class.

Previous approaches to discourse processing tend to emphasise the role of commonsense inference mechanisms. We proposed techniques which avoid encoding specific knowledge and common sense reasoning techniques; the only lexical source we used was WordNet (Miller et al., 1993). As a consequence, our system can process definite descriptions efficiently in any domain. In section Section §6.1 we briefly review other systems performing similar tasks.

We measured the coders agreement on the annotation exercise and found out that it was more difficult for the annotators to agree on bridging descriptions than on the other classes (anaphoric and discourse new descriptions). This class is also the most difficult to process. However, in the kind of texts we worked with bridging descriptions are a small class in comparison to the others; failing to resolve them does not dramatically affect the overall performance of a shallow system such as the one proposed here. On the other hand, this class is theoretically interesting due to its great complexity and the difficulties it imposes for processing; but because they are less frequent, the data for analysis and tests were limited. In other text genres the distribution of definite descriptions into the classes might change; but we believe that the ordering of the heuristics we propose here will still be adequate. Direct anaphora and discourse new descriptions can be processed with much simpler methods and it seems that the distinguishing features do not usually overlap. Our tests with bridging descriptions, in which we used WordNet and other heuristics, resulted in a great number of false positives. This suggests that a focusing mechanism for selecting discourse referents is needed, as proposed in (Grosz, 1977; Sidner, 1979; Grosz et al., 1983; Grosz and Sidner, 1986; Grosz et al., 1983). In Section §6.2 we discuss the role of focus on definite description processing and other themes for future work.

6.1 Comparison with other systems

Several systems for processing definite NPs have been proposed. One difference between our work and those previous ones is that our system is only concerned with definite descriptions; other systems have usually treated referring or anaphoric expressions in general. The crucial difference from most existing systems is that they typically work in a specific domain and exploit hand coded common-sense knowledge.

Usually, these systems use discourse that is especially built for the purpose of testing the system, such as (Carter, 1987; Carbonell and Brown, 1988). Exceptions are the recently proposed systems dealing with textual coreference which participate in the Sixth Message Understanding Conference (Sundheim, 1995). We now review several of these systems.

6.1.1 Sidner's theory of definite anaphora comprehension

Sidner's algorithms are heavily dependent on the availability of a knowledge network and associated inference mechanism. In general, her rules would require a powerful knowledge engine to work on unrestricted domains, and this is out of reach just now. Her (co)specification rules are sometimes too restrictive[1] (see Section §2.4.2). Carter (1987), who re-interpreted her proposals, seems to agree with this, since he has weakened some of her proposed restrictions in his anaphor resolver. Also in (Grosz et al., 1983) it is recognised that an anaphoric full noun phrase may include some new and unshared information about a previously mentioned entity.

[1]Co-specification 1: Definite description and focus have the same head and no new information is introduced by the definite.

Sidner (1979) discusses in detail some aspects of anaphoric definite NPs, such as generic/specific and referential/attributive distinctions, proposing sophisticated techniques to deal with the problem with a semantic precision which is outside the scope of the present thesis. The reasons for that are threefold. First, we consider primarily linguistically represented entities, and our aim is mainly identifying discourse relations between definite descriptions and their discourse antecedents or else verifying the absence of such relations. Second, our goal is to test and evaluate the performance of a system which does not rely on specific world knowledge representation and associated inference mechanisms. Third, the sophistication she proposes did not show itself to be required in the task at hand for the largest amount of cases occurring in the type of texts we studied. Nevertheless, her discussion is certainly relevant and should constitute a theme for future work. (See Section §6.2.3.)

Although her emphasis is on anaphoric relations and associations, she acknowledges the occurrence of non-anaphoric definite NPs. She comments on how a definite that specifies outside the discourse should be treated and how to relate it to an associated data base, while in this thesis we aim at simply identifying the definite descriptions that specify outside the discourse as such.

The main contribution of Sidner's work is her theory of focus and its role in resolving definite NPs; we discuss this in Section §6.2.2 below.

6.1.2 Carter's shallow processing anaphor resolver

Carter's system implements a modified version of Sidner's focusing algorithm. Carter (1987) proposes a shallow processing anaphor resolver in which reasoning is minimally considered. The system avoids it when possible but does make use of specific hand coded knowledge. His system is tested over short stories specifically designed for the testing of the system. Definite descriptions are just one type of anaphoric expression among several dealt with his system.

6.1.3 The Core Language Engine

The Core Language Engine (CLE) (Alshawi, 1990) is a domain independent system which translates English sentences into formal representations. The construction of this formal representation passes through an intermediate stage called Quasi Logical Form (or QLF). The QLF may contain unresolved terms corresponding to anaphoric NPs including, among others, definite descriptions.

The resolution process which transforms QLF into resolved logical form representation of sentences is described in (Alshawi, 1990). Definite descriptions are represented by quantified terms. Referential readings of definite descriptions are handled by proposing referents from the external application context as well as the CLE context model. Attributive readings may also be proposed during QLF resolution. This means that the identification of an external or contextual referent is not necessary for the resolution of the QLF. Both referential reading resolution with the external application context and the attributive reading seem to account for discourse new descriptions. Although this

is claimed to be a domain independent system, CLE relies on a core lexicon (that allows new entries to be added) and world knowledge reasoning is used to verify the plausibility of choice of referents from an ordered list; the required world knowledge has to be added by hand for each application. The CLE seems to account well for discourse new descriptions, although they are not explicitly mentioned. Unlike us, they require world knowledge representation and world model (the external application context).

6.1.4 Probabilistic methods in anaphora resolution

Aone and Bennet (1995) propose an automatically trainable anaphora resolution system. They train a decision tree using the C 4.5 algorithm by feeding feature vectors for pairs of anaphor and antecedent. They use 66 features, including lexical, syntactic, semantic, and positional features. Their overall recall and precision figures are 66.56% and 72.18%. Considering only definite NPs whose referent is an organisation (that is the only distinction available in their report), recall is 35.19% and precision 50% (measured on 54 instances). Their training and test texts were newspaper articles about joint ventures, and they claim that because each article always talked about more than one organisation, finding the antecedents of organisational anaphora was not straightforward.

In (Burger and Connolly, 1992) a Bayesian network is used to resolve anaphora by probabilistically combining linguistic evidence. Their sources of evidences are: c-command (syntactic constraints), semantic agreement (gender, person, and number plus a term subsumption hierarchy), discourse focus, discourse structure, recency, centering. Their methods are described and exemplified but not evaluated. A Bayesian framework is also proposed by (Cho and Maida, 1992) for the identification of definite descriptions' referents.

6.1.5 MUC-6 systems in the coreference task

There were seven systems participating in the coreference task in the MUC-6 competition. They presented recall scores ranging from 35.69% to 62.78% and precision scores ranging from 44.23% to 71.88%. It is important to note that the evaluation in MUC-6 differs from ours in three important aspects. First of all they have to parse the texts. Secondly, the evaluation there considers the co-referential chain, and not only one correct antecedent. The third difference is that they annotate a wider range of referring expressions (pronouns, bare nouns), while we annotate only definite NPs, which makes comparison difficult; on the other hand, not all definite descriptions are marked in the MUC-6 coreference task: they mark only identity relations, and the relation is not marked if the antecedent is a clause or a conjoined NP. This leaves out bridging references which, as we have seen, are by far the most difficult cases.

Kameyama (1997) analyses in more detail the co-reference module of the SRI system presented in MUC-6 (Appelt et al., 1995). This system presented one of the top scores for the co-reference task (a recall of 59% and precision of 72%). Their system uses a sort hierarchy claimed to be sparse and incomplete. For definite descriptions they report

a test on five articles in which recall was 46% (28/61); and for proper names recall was 69% (22/32). Their analyses rely on a much lower number of cases than ours. The precision figures for these two sub-classes are not reported. Some of the errors in definite descriptions are said to be due to non-identity referential relations; however, there is no mention of differences between discourse new and bridging descriptions. Other errors were said to be related to failure in recognising synonyms.

Coreference annotation

In MUC-6, an inter-annotator variability test was also conducted. Two independent manual annotations of 17 articles were scored against each other. The agreement was measured in terms of recall and precision. The result was 80% recall and 82% precision. The 20% disagreement was attributed to problems such as overlooking coreferential NPs, using different interpretations of the guidelines, and making different subjective decisions for ambiguous, sloppy texts. They observed that most human errors occurred with definite descriptions and bare nominals, and not with names and pronouns. These results are comparable to ours: our three coders in the second experiment had a percentage agreement of 82% which, after excluding expected chance agreement, results in a coefficient of agreement $K = 0.68$.

6.2 Future work

6.2.1 Theoretical developments

We have defended the importance of identifying discourse new descriptions, and we believe that there is still need for research into the semantics of this class; the role of premodification and postmodification should also be further examined. Postmodification is one of the most frequent features of discourse new descriptions; additional empirical studies considering a detailed subclassification of discourse new descriptions would give a better understanding of the problem. The distribution of copula constructions, appositions, premodification, postmodification, proper nouns we have presented were just based on the results of our system. The postmodification of a description is like a self-contained anchor (what Löbner (1985) calls 'disambiguating arguments and attributes'); how the head noun of a postmodified description relates 'semantically' with its complement is a problem similar to how a bridging description relates to its anaphoric anchor. To date, there hasn't been much research on this topic. Even a relation of coreference can occur between an NP's head noun and its complement, as seen in the examples in (6.1):

(6.1) a. the dream of home ownership
b. the issue of student grants

We also observed that definite descriptions with premodification were responsible for considerable amount of disagreement among the annotators. The reasons for that are still to be explained.

6.2.2 The role of focus in definite descriptions processing

Our tests with bridging descriptions resulted in a great number of false positives. This suggests that a focusing mechanism for selecting discourse referents is needed.

Sidner's set of (co-)specification rules (presented in Section §2.4.2) relies on algorithms for tracking local focus. Sidner claims that a definite description co-specifies with one of the elements in one of the focus structures. Her rules of specification are even more restrictive since they do not consider stacked focus. This means that her algorithms only look for an associated antecedent in the previous sentence.

Grosz et al. (1983) claim that local focus (or centering[2]) has greater effect on pronominal expressions while global focus has major effects in the interpretation of non-pronominal definite referring expressions. Fraurud (1990) is also critical of the idea that local focus plays a role in the resolution of definite descriptions. She claims that although it plays a central role in current theories of anaphora, the set of possible anchors for definite NP interpretation is wider and more differentiated than the set proposed in such theories.

Furthermore, Sidner's algorithms, as stated, are difficult to implement, since reasoning and information about the thematic role of the verbs is needed for the identification of focus; also, there is not a precise definition of how to deal with embedded sentences or clauses in Sidner's proposal.[3]

Difficulties such as the above mentioned are the reason why our system does not include a focus tracking mechanism. However, we believe that a less restrictive and more practical way to select and order candidate antecedents, possibly integrating local and global focus, should be taken into account, specially for dealing with bridging descriptions; but this still remains to be defined.

6.2.3 Further issues in annotation

The cases of disagreement among annotators (and annotators and system) which involve direct anaphora might be related to problems discussed by Sidner, such as generic/specific and referential/attributive distinctions. In the MUC-6 coreference task definition, it was observed that two expressions are coreferential if they both refer to types and the types are identical or if both refer to sets and the sets are identical. An example given in the task guidelines is:

(6.2) ... *producers* don't like to see a hit wine increase in price... *Producers* have seen this market opening up and *they* are now creating wines that appeal to these people.

They say that *producers, Producers* and *they* in the example above refer to types and they all refer to the same type; if they were interpreted as referring to sets, they would not all refer to the same set. They mention also a difference between functions and values:

[2]Centering theory is a reformulation of Sidner's theory concerned more specifically with local focus.

[3]Revisions and extensions of Sidner's proposal related to these problems have been proposed (Suri and McCoy, 1993; Suri and McCoy, 1994).

(6.3) *The temperature* is 90... *The temperature* is rising.

They say that the first occurrence of *The temperature* refers to the value and the second one to the function, so they are not coreferential. Also, mention is made of problems raised by metonymy:

(6.4) *The White House* sent its health care proposal to Congress yesterday. Senator Dole said *the administration*'s bill had little chance of passing.

They say that in (6.4) there is a coercion from the White House to the administration operating out of the White House, and that they are therefore coreferent. Our text annotation instructions do not mention these problems. The analyses of cases like these and their effect on our experiments on text annotation, the adoption of detailed explanations and diversified examples in the subjects' training for text annotation, as well as the implications of these problems for processing are a theme for future work.

6.2.4 The system: extensions and applications

A natural application of this work is to include the ideas developed here in a tool for semi-automatic coreference annotation. It would be important to test the functioning of the system with a partial parser. Also, some specific functionalities were only minimally developed here, such as detecting copula constructions and the treatment of named entities. Some basic operations are only implemented very crudely: e.g., we assume as plural any word ending in 's', and as proper name any capitalised word. Another aspect of the system that deserves further examination is the construction of coreference chains, and cases of multiple resolutions. We did not get a clear picture of how complete/incomplete, or how broken the co-referential chains resulting from the processing of one text are, and we did not relate them with the chains of the annotated texts. In order to do this, the system would have to be extended to cover all cases of anaphoric expressions.

Appendix A

Text Annotation Instructions 1

Classification of uses of "the"-phrases

You will receive a set of texts to read and annotate. From the texts, the system will extract and present you "the"-phrases and will ask you for a classification. You must choose one of the following classes:

1. ANAPHORIC (same noun): For anaphoric "the"-phrases the text presents an antecedent noun phrase which has the same noun of the given "the"-phrase. The interpretation of the given "the"-phrase is based on this previous noun-phrase.

2. ASSOCIATIVE: For associative "the"-phrases the text presents an antecedent noun phrase which has a different noun for the interpretation of the given "the"-phrase. The antecedent for the "the"-phrase in this case may

a) allow an inference towards the interpretation of the "the"-phrase,

b) be a synonym,

c) be an associate such as part-of, is-a, etc.

d) a proper name

3. LARGER SITUATION/UNFAMILIAR: For larger situation use of "the"-phrases **you do not find an explicit antecedent in the text**, because the reference is based on basic common knowledge:

a) first occurrences of proper names (subsequent occurrences must be considered as anaphoric),

b) reference to times,

c) community common knowledge;

d) proper names in premodifier position.

Also for unfamiliar uses of "the"-phrases **the text does not provide an antecedent**. The "the"-phrase refers to something **new** to the text. The help for the interpration may be given together with the "the"-phrase as in

e) restrictive relative clauses (the ... that ... - RC in general)

f) associative clauses (the ... of ... - PP in general)

g) NP complements (the fact that ...,the conclusion that ...)

h) unexplanatory modifiers (the first ..., the best ...)

i) appositive structures (James Dean , the actor)

j) copulas (the actor is James Dean)

4. IDIOM: "The"-phrases can be used just as idiomatic expressions, indirect references or metaphorical uses.

5. DOUBT: When you are in doubt about the classification: a comment on your doubt is requested.

PREFERENCE ORDER FOR THE CLASSIFICATION: In spite of the fact that definites often fall in more than one class of use, the identification of a unique class is required. In order to make the choices uniform, priority is to be given to anaphoric situations. According to this ordering, cases like "the White House" or "the government" are anaphoric rather than larger situation, **when it has already occurred once in the text**. When a "the"-phrase seems to belong both to larger sit./unfamiliar and associative classes, preference is given to larger sit./unfamiliar.

Examples

[Examples from the corpus were given as in section 3.2.1.]

Summary

WHEN AN ANTECEDENT IS GIVEN EXPLICITLY IN THE TEXT:(1,2)	WHEN THE REFERENT FOR THE DESCRIPTION IS KNOWN OR NEW:(3,4)
1.: ANAPHORIC There is an antecedent in the text which has the same descriptive noun of the "the"-phrase.	3.: LARGER SIT./UNFAMILIAR The "the"-phrase is novel in the text, unique identifiable, or based on common knowlege or is given with its referent
2.: ASSOCIATIVE There is an antecedent in the text which has a different noun, but it is a synonym or associate to the description.	4.: IDIOM The "the"-phrase is an idiomatic expression

1. (a) a house: **the house**
2. (a) something has changed: **the change**
   (b) a home: **the house**
   (c) a house: **the door**
   (d) Kadane Co.: **the company**
3. (a) the White House (first occurrence)
   (b) the third quarter
   (c) the nation
   (d) the Iran-Iraq war
   (e) **the woman** he likes
   (f) **the door** of the house
   (g) **the fact** that
   (h) the first, the best, the highest, the tallest ...
   (i) James Dean, **the actor**
   (j) **the actor** is James Dean
4. (a) back into **the soup**

Appendix B

Text Annotation Instructions 2

Text Annotation of Definite Descriptions

This material provides you with instructions, examples and some training for the text-annotation task. The task consists of reading newspaper articles and analysing occurrences of DEFINITE DESCRIPTIONS, which are expressions starting with the definite article THE. We will call these expressions DDs or DD. DDs describe things, ideas or entities which are talked about in the text. The things, ideas or entities being described by DDs will be called ENTITIES. You should look at the text, carefully in order to indicate whether the ENTITY was mentioned before in the text and if so, to indicate where. You will receive a set of texts and their corresponding tables to fill in. There are basically four cases to be considered:

1. Usually DDs pick up an entity introduced before in the text. For instance, in the sequence:

"*Mrs. Park* is saving to buy an apartment. *The housewife* is saving harder than ever."

the ENTITY described by the DD "*the housewife*" was mentioned before as "*Mrs. Park*".

2. If the ENTITY itself was not mentioned before but its interpretation is based on , dependent on, or related to some other idea or thing in the text, you should indicate it. For instance, in the sequence:
" The Parks wanted to buy *an apartment* but *the price* was very high.
the ENTITY described by the DD *the price* is related to the idea expressed by *an apartment* in the text.

3. It may also be the case that the DD was not mentioned before and is not related to something in the text, but it refers to something which is part of the common knowledge of the writer and readers in general. (The texts to be analysed are Wall Street Journal articles - location and time, for instance, are usually known to the general reader from sources which are outside the text). Example:
"During *the past 15 years* housing prices increased nearly fivefold".
here, the ENTITY described by the DD *the past 15 years* is known to the general reader of the Wall Street Journal and was not mentioned before in the text.

4. Or it may be the case that the DD is self-explanatory or it is given together with its own identification. In these cases it becomes clear to the general reader what is being talked about even without previous mention in the text or without previous common knowledge of it. For instance:
"The proposed legislation is aimed at rectifying some of *the inequities in the current land-ownership system*."
the ENTITY described here is new in the text, and is not part of the knowledge of readers but the DD *the inequities in the current land-ownership system* is self-explanatory.

The texts will be presented to you in the following format: on the left, the text with its DDs in evidence; on the right, the keys (number of the sentence/number of DD) and the DD to be analysed. The key is for internal control only, but it may help you to find DDs in the table you have to fill in.

Text 0

1 Y. J. Park and her family scrimped for four years to buy a tiny apartment here, but found that the closer they got to saving the $40,000 they originally needed, the more **the price** rose.	**(1/1) the price**
...	
3 Now **the 33-year-old housewife**, whose husband earns a modest salary as an assistant professor of economics, is saving harder than ever.	**(3/2) the 33-year-old housewife**
...	
9 During **the past 15 years**, the report showed, housing prices increased nearly fivefold.	**(9/3) the past 15 years**
...	
22 The proposed legislation is aimed at rectifying some of **the inequities in the current land-ownership system**.	**(22/4) the inequities in the current land-ownership system**

You can draw arrows, use colours, whatever you like over the text and the list of DDs to help your analysis and then you should complete a table in the format below.

Text 0 **Key**	**DEFINITE DESCRIPTION**	**LINK** **=/R**	**LINK** **Sentence no./** **previous mention**	**NO LINK** **K/D**
1/1	**the price**			
3/2	**the 33-year-old housewife**			
⋮				

Each case (1 to 4, above) is to be indicated on the table according to the following (see examples in the table below):

Whenever you find a previous mention in the text of the DD you should mark the column **LINK**:
1. Mark "=" if the ENTITY described was mentioned before.
2. Mark "R" if the ENTITY described is new but it is related/based/dependent on something mentioned before).
In the case of both **1** and **2** you should provide the sentence number where the previous/related mention is and write down the previous/related mention of it (see example in the table below).

If the entity was not previously mentioned in the text and it is not related to something mentioned before, then mark the column **NO LINK**:
3. Mark "K" if it is something of writer/readers' common knowledge.
4. Mark "D" if it is new in the text and the readers have no previous knowledge about it but the description is enough to make readers identify it.

Text 0 **Key**	**DEFINITE DESCRIPTION**	**LINK** **=/R**	**LINK** **Sentence no./** **previous mention**	**NO LINK** **K/D**
1/1	**the price**	R	*1/apartment*	
3/2	**the 33-year-old housewife**	=	*1/Y.J. Park*	
9/3	**the past 15 years**			K
22/4	**the inequities in the current land-ownership system**		— —	 D

In case of doubt just leave the line in blank and comment at the back of the page using the key number to identify the DD you are commenting on.

Examples

Next we present some examples and further explanation for each one of the four cases that are bein considered.

Case 1 - LINK (=)

For case no. 1 you may find a previous mention that may be equal or different from the DD (for instance, the government - **the government**, a report - **the report**, and three bills - **the proposed legislation** in the examples below); distances from previous mentions and DDs may also vary.

- Meanwhile, the government's Land Bureau reports that only about a third of Korean families own their own homes. Last week, **the government** took three bills to the National Assembly.
- Last May, a government panel released a report on the extent and causes of the problem. During the past 15 years, **the report** showed, housing prices increased nearly fivefold.

- Last week, the government took three bills to the National Assembly. **The proposed legislation** is aimed at rectifying some of the inequities in the current land-ownership system.

Case 2 - LINK (R)

Here are cases of DDs which are related to something that was present in the text. If you ask for the examples below, "Which *government, population, nation* is that?","Which *blame* is that?" the answer is given by something previously mentioned in the text (Koreans, and the increase of housing prices, respectively) [1].

- For the Parks and millions of other young Koreans, the long-cherished dream of home ownership has become a cruel illusion. For **the government**, it has become a highly volatile political issue. In 1987, a quarter of **the population** owned 91% of **the nation**'s 71,895 square kilometers of private land.
- During the past 15 years, the report showed, housing prices increased nearly fivefold. The report laid **the blame** on speculators, who it said had pushed land prices up ninefold.

Case 3 - NO LINK (K)

These cases of DDs are based on the common reader's knowledge. The texts to be analysed are Wall Street Journal articles - location and time, for instance, are usually known to the general reader from sources which are outside the text [2].

- For example , officials at Walnut Creek office learned that the Amfac Hotel near **the San Francisco airport**, which is insured by Aetna, was badly damaged when they saw it on network television news.
- Adjusters who had been working on **the East Coast** say the insurer will still be processing claims from that storm through December .

Case 4 - NO LINK (D)

These cases of DDs are self-explanatory or accompained by their identification. For instance if you ask "Which *difficulty* is that?", "Which *fact* is that?", "Which *know-how* is that?" etc. for the examples below, the answer is given by the DD itself. In the last example the DD is accompained by its explanation.

[1] Note that DDs like *the blame, the government, the population*, which are case 2 in their first occurrence, are to be considered case 1 in possible posterior occurrences.

[2] Note that a DD like "the government" may belong to case 2 as exemplified, but it may refer to the U.S.A. in another text, without any explicit mention of U.S.A in the text, since it is the country where the newspaper is produced. In such a situation the DD "the government" belongs to case 3. It may also be the case that the entity is part of the readers' knowledge but was mentioned before, in this situation it belongs to case 1.

- Because of **the difficulty of assessing the damages caused by the earthquake**, Aetna pulled together a team of its most experienced claims adjusters from around the country .
- They wonder whether he has **the economic know-how to steer the city through a possible fiscal crisis**.
- Mr. Dinkins also has failed to allay Jewish voters' fears about his association with the Rev. Jesse Jackson, despite **the fact that few local non-Jewish politicians have been as vocal for Jewish causes in the past 20 years as Mr. Dinkins has**.
- But racial gerrymandering is not **the best way to accomplish that essential goal**.
- **The first hybrid corn seeds produced using this mechanical approach** were introduced in the 1930s and they yielded as much as 20 % more corn than naturally pollinated plants.
- **The Citizens Coalition for Economic Justice**,*a public-interest group leading the charge for radical reform*, wants restrictions on landholdings, high taxation of capital gains, and drastic revamping of the value-assessment system on which property taxes are based.

SCRIPT

In order to help you filling in the table, answer the YES-NO questions below for each one of the DDs in the text. When the answer for the question is YES (Y) you have an action to follow, if the answer is NO (N), skip to the next question.

1. **Does the DD describe an ENTITY mentioned before?**

   Y Mark "**=**" (column LINK) to indicate that the same entity was mentioned before and tell where by providing the sentence number and the words used in the previous mention.

   N Go to question no. 2.

2. **Is the ENTITY new but related to something mentioned berfore? If you ask: "Which entity is that?", is the answer based on previous text** [3]**?**

   Y Mark "**R**" (column LINK) to indicate related entity and provide the sentence number and the previous mention on which the DD is based .

   N Go to question no. 3.

3. Is the ENTITY new in the text? If it was not mentioned before and its interpretation is not based on the previous text, then: **is it something mutually known by writer and general readers of the Wall Street Journal?**

[3]For instance if you ask: "Which *price* is that?" for *the price* in sentence number 1, given above, your answer is based on *apartment* in the text.

Y Mark "**K**" (column NO LINK) to indicate general knowledge about the entity.

N Go to question no. 4.

4. Is the ENTITY new in the text? If it was not mentioned before and its interpretation is not based on the previous text, then: **is it self-explanatory or accompanied by its identification?**

   Y mark "**D**" (column NO LINK) to indicate that the description is enough to make readers identify the entity.

   N Leave the line in blank and comment at the back of the page using the key number to identify the DD."

Appendix C

Interactions with the Working System

Our system runs over WSJ articles from Penn Treebank I (first version) converted into a Prolog format by our converter program. The system is loaded by:

```
| ?- ['ddr-main'].
```

To process one text the user may call:

```
| ?- defres(File_name).
```

The system displays its results for each definite description individually and globally; to suppress/activate the display one has to call:

```
| ?- verbose(0).  % suppress all displays

| ?- verbose(1).  % suppress the display of individual results
                  % activate the global results

| ?- verbose(2).  % activate all displays
```

After processing one text, the results can be examined through these functions:

```
show_functions. : (shows system functions) - as listed below

show_definite_descriptions. : (lists all definite descriptions
                               found in the text)

show_potential_antecedents. : (lists all NPs considered as potential
                               antecedents by the system)

show_possible_antecedents.  : (lists those potential antecedents
                               actually used in the resolutions)
```

```
show_dds_classification.    : (shows dds, antecedents and
                                classification)

show_resolved_dds.     : (shows resolved dds and antecedents)

show_equi_classes.     : (shows equivalence classes)

show_equi_classes_l.   : (as above without showing syntax)

show_ls_unf_dds.       : (shows larger sit./unfamiliar dds )

show_ls_unf_dds_l.     : (as above without showing syntax )

show_unresolved_dds.   : (shows the list of unresolved dds)

show_unresolved_dds_l. :(as above without showing syntax)

show_def_names.        : (shows definite names)

show_timerefs.         : (shows definite time references)

show_dd_sentence.      : (asks DD index and shows DD sentence)

show_sentence_syntax.  : (asks sentence number,
                          shows sentence syntax)

show_text.             : (asks DD index and prints all
                          previous text)

show_previous_sents.   : (asks DD index and prints
                          previous sentences)

show_all_text.         : (prints whole text)

show_all_text_dd.      : (prints whole text highlighting dds)

total_types.           : (displays dd types)

total_types_unf.       : (displays unfamilar types)
```

C.1 An example

This example refers to text wsj_0761. The complete text is presented in the next section.

```
| ?- defres(w0761.par.np).


Sentence 1:
```

```
Y.J. Park and her family scrimped for four years to buy
a tiny apartment here , but found that the closer they
got to saving the $ 40,000 they originally needed , the
 more ** the price ** rose .

<< the price  >> *** UNRESOLVED ***
...

Sentence 6:
For the Parks and millions of other young Koreans , ** the
long-cherished dream of home ownership ** has become a cruel
illusion .

<< the long-cherished dream of home
   ownership  >> *** FIRST MENTION ***
...

Sentence 9: During the past 15 years , ** the report ** showed ,
housing prices increased nearly fivefold .

<< the report  >> *** DIRECT ANAPHORA ***

Antecedent: a report on the extent and causes of the problem
Sentence 8: Last May , a government panel released ** a report
 on the extent and causes of the problem ** .
...

 *********  TEXT  RESULTS *********

File: wsj/w0761.par

NR. OF SENTENCES: 48

NR. OF NOUN PHRASES: 368

NR. OF ANTECEDENTS CONSIDERED: 173

        Indefinites:81
        Possessives:22
        Definites:70

NR. OF DEFINITE DESCRIPTIONS: 78

** DISCOURSE NEW DESCRIPTIONS : 29

LARGER SITUATIONS  USES : 7

    NAMES          : 4
    TIME REFERENCES: 3
    REST. PREMOD.  : 0

UNFAMILIAR  USES : 22
```

```
    NP COMP. and UNEXP. MOD.: 4
    APPOSITIONS             : 1
    RESTRICTIVE POSTMOD.    : 15
    COPULA                  : 2

**  DIRECT  ANAPHORA : 27

     ANTECEDENTS FOUND - Indefinites: 3
                       - Possessives: 1
                       - Definites:  23

** NON-IDENTIFIED : 22

yes
| ?- show_equi_classes.


43 [NP,a,government,panel] s.8

59 [NP,The,panel] s.11


48 [NP,a,report,[PP,on,[NP,[NP,the,extent],and,
[NP,causes,[PP,of,[NP,the,problem]]]]]] s.8

50 [NP,the,report] s.9
52 [NP,The,report] s.10
95 [NP,the,report] s.14


89 [NP,the,population] s.14

96 [NP,the,population] s.14


45 [NP,the,problem] s.8

112 [NP,the,problem] s.17


103 [NP,[NP,the,government],'s,Land,Bureau] s.15

123 [NP,the,Land,Bureau] s.19


138 [NP,three,bills] s.21

146 [NP,the,bills] s.23
267 [NP,the,bills] s.36


98 [NP,the,land] s.14
```

```
279 [NP,The,land] s.38


39 [NP,the,government] s.7

66 [NP,the,government] s.11
102 [NP,the,government] s.15
127 [NP,the,government] s.19
137 [NP,the,government] s.21
158 [NP,The,government] s.24
173 [NP,the,government] s.26
178 [NP,The,government] s.27
197 [NP,the,government] s.29
243 [NP,the,government] s.33
248 [NP,the,government] s.34
264 [NP,the,government] s.36
273 [NP,the,government] s.37
313 [NP,The,government] s.42


125 [NP,the,National,Assembly] s.19

139 [NP,the,National,Assembly] s.21
316 [NP,the,National,Assembly] s.42


91 [NP,the,nation] s.14

151 [NP,the,nation] s.23
320 [NP,the,nation] s.42

yes

| ?- show_ls_unf_dds_l.

12 the closer they got to saving the $ 40,000 they originally needed s.1

14 the more the price rose s.1

37 the long-cherished dream of home ownership s.6

49 the past 15 years s.9

65 the 1988 Seoul Olympics s.11

73 The result s.12

87 the prospects of buying a home s.13

98  the land devoted to housing s.14

120 the past three months s.19
```

```
122  the office complex where the Land Bureau is housed s.19

125 the National Assembly s.19

133 the past year s.20

143 the inequities in the current land-ownership system s.22

156 the amount of real estate one family can own , to 660
square meters in the nation ‘s six largest cities , but more
in smaller cities and rural areas s.23

171 the resale of property s.26

174 the sale of idle land to the government s.26

180  the average realized for other similar-sized properties
in an area s.27

190 the full scope of the penalties s.28

204 the popular standing of President Roh s.29

225  The Citizens Coalition for Economic Justice , a
public-interest group leading the charge for radical
reform , s.32

235  the value-assessment system on which property
taxes are based s.32

246 the Federation of Korean Industries s.34

255 the arguments of business leaders s.35

259 the capitalistic principle of private property s.36

278 the shortage of land s.37

295 The chief culprits s.40

319 the first half of 1989 s.42

326 The Ministry of Finance s.43

354 The maximum allowable property holdings for
insurance companies s.46

yes

| ?- show_ls_unf_dds.

12 [NP,the,closer,[SBAR,0,[S,[NP,they],[VP,got,[NP,T],
[S,[NP,*],to,[VP,saving,[NP,the,[NP,[NP,$,40,000],
```

```
[SBAR,0,[S,[NP,they],originally,[VP,needed]]]]]]]]]]s.1

14 [NP,the,more,[SBAR,0,[S,[NP,the,price],[VP,rose]]]]s.1

37 [NP,the,long-cherished,dream,[PP,of,[NP,home,ownership]]]s.6

49 [NP,the,[ADJP,past],15,years]s.9

65 [NP,the,1988,Seoul,Olympics]s.11

...

yes

| ?- show_all_text_dd.

1 Y.J. Park and her family scrimped for four years to buy a
tiny apartment here , but found that  **/ the closer they got to
saving the $ 40,000 they originally needed /** ,  **/ the more
the price rose /** .

2 By this month , it had more than doubled .

...
```

C.2 Text wsj_0761

1 Y.J. Park and her family scrimped for four years to buy a tiny apartment here, but found that **the closer they got to saving the $ 40,000 they originally needed** , **the more the price rose** .

2 By this month, it had more than doubled.

3 Now **the 33-year-old housewife** , whose husband earns a modest salary as an assistant professor of economics, is saving harder than ever.

4 I am determined to get an apartment in three years, she says.

5 It's all I think about or talk about.

6 For **the Parks and millions of other young Koreans** , **the long-cherished dream of home ownership** has become a cruel illusion.

7 For **the government** , it has become a highly volatile political issue.

8 Last May, a government panel released a report on **the extent and causes of the problem** .

9 During **the past 15 years** , **the report** showed, housing prices increased nearly fivefold.

10 **The report** laid **the blame** on speculators, who it said had pushed land prices up ninefold.

11 **The panel** found that since 1987, real-estate prices rose nearly 50 % in a speculative fever fueled by economic prosperity, **the 1988 Seoul Olympics** and **the government** 's pledge to rapidly develop Korea's southwest.

12 **The result** is that those rich enough to own any real estate at all have boosted their holdings substantially.

13 For those with no holdings, **the prospects of buying a home** are ever slimmer.

14 In 1987, a quarter of **the population** owned 91 % of **the nation** 's 71,895 square kilometers of private land, **the report** said, and 10 % of **the population** owned 65 % of **the land devoted to housing**.

15 Meanwhile, **the government** 's Land Bureau reports that only about a third of Korean families own their own homes.

16 Rents have soared along with house prices.

17 Former National Assemblyman Hong Sa-Duk, now a radio commentator, says **the problem** is intolerable for many people.

18 I 'm afraid of a popular revolt if this situation is n't corrected, he adds.

19 In fact, during **the past three months** there have been several demonstrations at **the office complex where the Land Bureau is housed**, and at **the National Assembly** , demanding **the government** put a stop to real-estate speculation.

20 President Roh Tae Woo's administration has been studying **the real-estate crisis** for **the past year** with an eye to partial land redistribution.

21 Last week, **the government** took three bills to **the National Assembly** .

22 **The proposed legislation** is aimed at rectifying some of **the inequities in the current land-ownership system** .

23 Highlights of **the bills** , as currently framed, are : – A restriction on **the amount of real estate one family can own, to 660 square meters in the nation's six largest cities, but more in smaller cities and rural areas** .

24 **The government** will penalize offenders, but wo n't confiscate property.

25 – A tax of between 3 % and 6 % on property holdings that exceed **the government set ceiling** .

26 – Taxes of between 15 % and 50 % a year on excessive profits from **the resale of property** , or **the sale of idle land to the government** .

27 **The government** defines excessive profits as those above **the average realized for other similar-sized properties in an area**.

28 – Grace periods ranging from two to five years before **the full scope of the penalties** takes effect.

29 **The administration** says **the measures** would stem rampant property speculation, free more land for **the government** 's ambitious housing-construction program, designed to build two million apartments by 1992 – and, perhaps, boost **the popular standing of President Roh** .

30 But opposition legislators and others calling for help for South Korea's renters say **the proposed changes** do n't go far enough to make it possible for ordinary people to buy a home.

31 Some want lower limits on house sizes others insist on progressively higher taxation for larger homes and lots.

32 **The Citizens Coalition for Economic Justice** , a public-interest group leading **the charge** for radical reform, wants restrictions on landholdings, high taxation of capital gains, and drastic revamping of **the value-assessment system** on which property taxes are based.

33 But others, large landowners, real-estate developers and business leaders, say **the government** 's proposals are intolerable.

34 Led by **the Federation of Korean Industries** , **the critics** are lobbying for **the government** to weaken its proposed restrictions and penalties.

35 Government officials who are urging real-estate reforms balk at **the arguments of business leaders** and chafe at their pressure.

36 There is no violation of **the capitalistic principle of private property** in what we are doing, says Lee Kyu Hwang, director of **the government** 's Land Bureau, which drafted **the bills** .

37 But, he adds, **the constitution** empowers **the government** to impose some controls, to mitigate **the shortage of land** .

38 **The land available for housing construction** stands at about 46.2 square meters a person – 18 % lower than in Taiwan and only about half that of Japan.

39 Mr. Lee estimates that about 10,000 property speculators are operating in South Korea.

40 **The chief culprits** , he says, are big companies and business groups that buy huge amounts of land not for their corporate use, but for resale at huge profit.

41 One research institute calculated that as much as 67 % of corporate-owned land is held by 403 companies – and that as little as 1.5 % of that is used for business.

42 **The government** 's Office of Bank Supervision and Examination told **the National Assembly** this month that in **the first half of 1989** , **the nation** 's 30 largest business groups bought real estate valued at $ 1.5 billion.

43 **The Ministry of Finance** , as a result, has proposed a series of measures that would restrict business investment in real estate even more tightly than restrictions aimed at individuals.

44 Under those measures, financial institutions would be restricted from owning any more real estate than they need for their business operations.

45 Banks, investment and credit firms would be permitted to own land equivalent in value to 50 % of their capital – currently **the proportion** is 75 %.

46 **The maximum allowable property holdings for insurance companies** would be reduced to 10 % of their total asset value, down from 15 % currently.

47 But Mrs. Park acknowledges that even if **the policies** work to slow or stop speculation, apartment prices are unlikely to go down.

48 At best, she realizes, they will rise more slowly – more slowly, she hopes, than her family's income.

References

Alshawi, H. 1990. Resolving Quasi Logical Forms. *Computational Linguistics*, 16(3):133–144.

Anderson, A. H. et al. 1991. The HCRC map task corpus. *Language and Speech*, 34(4):351–366.

Aone, C. and S. W. Bennet. 1995. Automated acquisition of anaphora resolution strategies. In *Proc. AAAI Spring Symposium on Empirical Methods in Discourse Interpretation and Generation*, pages 1–7, Stanford.

Appelt, D. 1985. *Planning English Sentences*. Cambridge: Cambridge University Press.

Appelt, D. et al. 1995. SRI International FASTUS system MUC-6 test results and analysis. In *Proc. of the Sixth Message Understanding Conference*, pages 237–248. Columbia, Maryland. November 6-8.

Barker, C. 1991. *Possessive Descriptions*. Ph.D. thesis, University of California at Santa Cruz, Santa Cruz, CA.

Barwise, J. and J. Perry. 1983. *Situations and Attitudes*. Cambridge Mass. London: MIT Press.

Bikel, D. M. et al. 1997. Nymble: a high-performance learning name-finder. In *Proc. of the Fifth Conference on Applied Natural Language Processing*, pages 194–201. Washington, DC. March 31-April 3.

Birner, B. and G. Ward. 1994. Uniqueness, familiarity, and the definite article in English. In *Proc. of the Annual Meeting of the Berkeley Linguistic Society*, pages 93–102.

Bobrow, D. G. and T. Winograd. 1977. An overview of KRL, a knowledge representation language. *Cognitive Science* 1.

Bosch, P. and B. Geurts. 1989. Processing definite NPs. *IWBS Report 78*, IBM Germany, July.

Burger, J. D. and D. Connolly. 1992. Probabilistic resolution of anaphoric reference. In *Proc. AAAI Fall Symposium on Probabilistic Approaches to Natural Language*, pages 17–24, Cambridge, Massachusetts.

Carbonell, J. G. and R. D. Brown. 1988. Anaphora Resolution: a Multi-Strategy Approach. In *Proceedings of the 12th International Joint Conference on Computational Linguistics*, pages 96–101.

Carletta, J. 1996. Assessing agreement on classification tasks: the Kappa statistic. *Computational Linguistics*, 22(2):249–254.

Carletta, J. et al. 1997. The Reliability of a Dialogue Structure Coding Scheme. *Computational Linguistics*, 23(1):13–32.

Carter, D. M. 1987. *Interpreting Anaphors in Natural Language Texts*. Chichester, UK: Ellis Horwood.

Chinchor, N. A. and B. Sundheim. 1995. Message Understanding Conference (MUC) tests of discourse processing. In *Proc. AAAI Spring Symposium on Empirical Methods in Discourse Interpretation and Generation*, pages 21–26, Stanford.

Chinchor, N. A. 1995. Statistical significance of MUC-6 results. In *Proc. of the Sixth Message Understanding Conference* (MUC-6) , pages 39-44. Columbia, Maryland. November 6-8.

Cho, S. and A. S. Maida. 1992. Using a bayesian framework to identify the referents of definite descriptions. In *Proc. AAAI Fall Symposium on Probabilistic Approaches to Natural Language*, pages 39–46, Cambridge, Massachusetts.

Christopherson, P. 1939. *The Articles: A Study of Their Theory and Use in English*. Copenhagen : E. Munksgaard.

Clark, H. H. 1977. Bridging. In P. N. Johnson-Laird and P.C. Wason, editors, *Thinking: Readings in Cognitive Science*. London New York: Cambridge University Press.

Clark, H. H. and C. R. Marshall. 1981. Definite reference and mutual knowledge. In *Elements of Discourse Understanding*. New York: Cambridge University Press.

Cohen, P. R. 1978. On knowing what to say: Planning speech acts. *Technical Report 118*, Department of Computer Science, University of Toronto, January.

Cooper, R. 1993. Generalised quantifiers and resource situations. In P. Aczel, D. Israel, Y. Katagiri, and S. Peters, editors, *Situation Theory and its Applications, v.3*. CSLI and University of Chicago, Stanford, pages 191–212.

Dale, R. 1992. *Generating Referring Expressions*. Cambridge, Mass.: The MIT Press.

Donnellan, K. 1972. Proper names and identifying descriptions. In D. Davidson and G. Harman, editors, *Semantics of Natural Language*. Dordrecht: Reidel, pages 356–379.

Donnellan, K. 1977. Reference and Definite Descriptions. In S. P. Schwartz, editor, *Naming, Necessity and Natural Kinds*. Ithaca: Cornell University Press.

Francis, W. N. and H. Kucera. 1982. *Frequency Analysis of English Usage: lexicon and grammar*. Boston: Houghton Mifflin.

Fraurud, K. 1990. Definiteness and the processing of NPs in natural discourse. *Journal of Semantics*, 7:395–433.

Gallaway, C. 1996. Children's and adults' use of 'the'- how anaphoric is it? In S. Botley, J. Glass, T. McEnery, and A. Wilson, editors, *Approaches to Discourse Anaphora– Proceedings of the Discourse Anaphora and Resolution Colloquium*, pages 318–330. University of Lancaster, UCREL.

Gaizauskas, R. and K. Humphreys. 1997. Quantitative Evaluation of Coreference Algorithms in an Information Extraction System. In S. Botley and T. McEnery, editors, *Corpus-based and Computational Approaches to Discourse Anaphora*. UCL Press, London. Forthcoming.

Gaizauskas, R. et al. 1995. University of Sheffield: description of the LaSIE system as used fo MUC-6. In *Proc. of the Sixth Message Understanding Conference*, pages 207–220. Columbia, Maryland. November 6-8.

Gamut, L. T. F. 1991. Logic, language and meaning. Vol. 2: Intensional logic and logic grammar. Chicago, Ill. London: University of Chicago Press.

Grosz, B. J. 1977. *The Representation and Use of Focus in Dialogue Understanding*. Ph.D. thesis, Stanford University.

Grosz, B. J. et al. 1983. Providing a unified account of definite noun phrases in discourse. *Technical Note (SRI)*, 292. Artificial Intelligence Center, SRI International, Menlo Park, California.

Grosz, B. J. and C. L. Sidner. 1986. Attention, intention, and the structure of discourse. *Computational Linguistics*, 12(3):175–204.

Grosz, B. J. et al. 1995. Centering: A framework for modelling the local coherence of discourse. *Computational Linguistics*, May.

Halliday, M. A. K. and R. Hasan. 1976. *Cohesion in English*. London: Longman.

Hawkins, J. A. 1978. *Definiteness and Indefiniteness: a study in reference and grammaticality prediction*. London: Croom Helm.

Hatzivassiloglou, V. and McKeown, K. 1993 Towards the Automatic Identification of Adjectival Scales: Clustering Adjectives According to Meaning. In *Proceedings of the 31st Annual Meeting of the ACL*, pages 172–182. Association for Computational Linguistics, Columbus, Ohio, June.

Hearst, M. A. 1997 TextTiling: segmenting text into multi-paragraph subtopic passages. *Computational Linguistics*, 23(1):33–64.

Heeman, P. A. and J. F. Allen. 1995. The TRAINS-93 dialogues. TRAINS *Technical Note TN 94-2*, University of Rochester, Dept. of Computer Science, Rochester, NY.

Heim, I. 1982. *The Semantics of Definite and Indefinite Noun Phrases*. Ph.D. thesis, University of Massachusetts at Amherst.

Hendrix, G. 1975. *Partitional Networks for the Mathematical Modeling of Natural Language Semantics*. Technical Report NL-28, Department of Computer Science, University of Texas, Austin, Texas.

von Heusinger, K. 1997. Salience and Definiteness. *The Prague Bulletin of Mathematical Linguistics, 67*.

Hirst, G. 1981. *Anaphora in natural language understanding: a survey*. Berlin New York: Springer Verlag. Lecture notes in computer science; V.119.

Hirschberg, J. and B. Grosz. 1992. Intonational features of local and global discourse structure. In *Proceedings of the DARPA Workshop on Speech and Language Processing*, Harriman, NY.

Humphreys, K., R. Gaizauskas and S. Azzam. 1997. Event coreference for information extraction. In *Proceedings of the Workshop on Operational Factors In Practical, Robust, Anaphora Resolution for Unrestricted Texts*, pages 75–81. Sponsored by Association for Computational Linguistics, Madrid, Spain. July, 11.

Johansson, S. and K. Hofland. 1989. *Frequency Analysis of English vocabulary and grammar, based on the LOB corpus, I: Tag Frequencies and Word frequencies*. Oxford: Clarendon Press.

Kadmon, N. 1987. *On Unique and Non-Unique Reference and Asymmetric Quantification*. Ph.D. thesis, University of Massachusetts at Amherst.

Kamp, H. 1981. A Theory of Truth and Semantic Representation. *Formal Methods in the Study of Language*. J. Groenendijk, T. Janssen, and M. Stokhof, editors. Amsterdam: Mathematisch Centrum.

Kamp, H. and U. Reyle. 1993. *From Discourse to Logic*. Dordrecht London: Kluwer Academic.

Kameyama, M. 1997. Recognizing referential links: an information extraction perspective. In *Proceedings of the Workshop on Operational Factors In Practical, Robust, Anaphora Resolution for Unrestricted Texts*, pages 46–53. Sponsored by Association for Computational Linguistics, Madrid, Spain. July, 11.

Korbayová, I. 1994. Contextual Reference of Noun Phrases in Plinius. *The Prague Bulletin of Mathematical Linguistics* 61:23–46 and 62:47–72.

Kronfeld, A. 1990. *Reference and Computation*. Cambridge, UK: Cambridge University Press.

Kowtko, J. C., S. D. Isard, and G. M. Doherty. 1992. Conversational games within dialogue. *Research Paper HCRC/RP-31*, Human Communication Research Centre.

Kripke, S. A. 1977. Speaker reference and semantic reference. In P. A. French, T. E. Uehling, and H. K. Wettstein, editors, *Contemporary Perspectives in the Philosophy of Language*. Minneapolis:University of Minnesota Press, pages 6–27.

Krippendorff, K. 1980. *Content Analysis: An introduction to its methodology*. Beverly Hills London: Sage Publications.

Landis, J. R., and G. G. Koch. 1977. The measurement of observer agreement for categorial data. *Biometrics*, 36:159-174.

Löbner, S. 1985. Definites. *Journal of Semantics*, 4:279–326.

Löbner, S. 1996. Associative Anaphora. In *Proceedings of the Workshop on Indirect Anaphora*, Lancaster.

Mani, I. and T. R. MacMillan. 1996. Identifying unknown proper names in newswire text. In B. Boguraev and J. Pustejovsky, editors, *Corpus processing for lexical acquisition*, pages 41–59.

McDonald, D. 1996. Internal and external evidence in the identification and semantic categorization of proper names. In B. Boguraev and J. Pustejovsky, editors, *Corpus processing for lexical acquisition*, pages 21–39.

McEnery, T., Tanaka, I. and Botley, S. 1997. Corpus annotation and reference resolution. In *Proceedings of the Workshop on Operational Factors In Practical, Robust, Anaphora Resolution for Unrestricted Texts*. Sponsored by Association for Computational Linguistics, Madrid, Spain. July, 11.

Miller, G. et al. 1993. Introduction to WordNet: an on-line lexical database, five papers on WordNet. *Technical Report CSL Report 43*, Cognitive Science Laboratory, Princeton University.

Montague, R. Formal philosophy: selected papers. 1974. Edited by R. H. Thomason, New Haven: Yale University Press.

Moore, J. and M. Walker, editors. 1995. *Empirical Methods in Discourse Interpretation and Generation - Papers from the 1995 AAAI Spring Symposium*, Stanford, March. AAAI.

Neale, S. 1990. *Descriptions*. Cambridge, Mass.: MIT Press.

Nunberg, G. D. 1978. *The pragmatics of reference*. Bloomington, Ind.: Indiana University Linguistics Club.

Paik, W. et al. 1996. Categorizing and standardazing proper nouns for efficient information retrieval. In B. Boguraev and J. Pustejovsky, editors, *Corpus processing for lexical acquisition*, pages 61–73.

Palmer, D. D. and D. .S. Day. 1997. A statistical profile of the named entity task. In *Proc. of the Fifth Conference on Applied Natural Language Processing*, pages 190–193. Washington, DC. March 31-April 3.

Passonneau, R. and D. Litman. 1993. Feasibility of automated discourse segmentation. In *Proceedings of 31st Annual Meeting of the ACL*. Association for Computational Linguistics. Ohio State University, Columbus Ohio, June 22-26.

Poesio, M. 1993. A situation-theoretic formalization of definite description interpretation in plan elaboration dialogues. In P. Aczel, D. Israel, Y. Katagiri, and S. Peters, editors, *Situation Theory and its Applications*, vol.3. CSLI, Stanford, pages 339–374.

Poesio, M. 1994. Weak definites. In *Proceedings of the Fourth Conference on Semantics and Linguistic Theory, SALT-4*. Cornell University Press.

Poesio, M., Vieira, R. and Teufel, S. 1997. Resolving bridging descriptions in unrestricted texts. Edinburgh University, HCRC Research Paper, HCRC/RP-87. In *Proceedings of the Workshop on Operational Factors In Practical, Robust, Anaphora Resolution for Unrestricted Texts*, pages 1–6. Sponsored by Association for Computational Linguistics, Madrid, Spain. July, 11.

Poesio, M., Vieira, R. 1997. A Corpus-based investigation of definite description use. Edinburgh University, CCS Research Paper, Ref:EUCCS-RP-1997-1. In *Computational Linguistics*, forthcoming.

Postal, P. 1969. Anaphoric Islands. In R. I. Binnick et al., Eds. *Papers from the Fifth Regional Meeting of the Chicago Linguistic Society*. University of Chicago IL, pages 205-235.

Prince, E. F. 1981. Toward a taxonomy of given-new information. In P. Cole, editor, *Radical Pragmatics*. Academic Press, New York, pages 223–256.

Prince, E. F. 1992. The ZPG letter: subjects, definiteness, and information status. In S. Thompson and W. Mann, editors, *Discourse description: diverse analyses of a fund-raising text*. John Benjamins, pages 295–325.

Quinlan. J. R. 1993. *C4.5: programs for machine learning*. San Mateo: Morgan Kaufmann Publishers.

Quirk, R. et al. 1985. *A Comprehensive Grammar of the English language*. London: Longman.

Reinhart, T. 1981. Pragmatics and Linguistics: An Analysis of Sentence Topics *Philosophica*, 27(1).

Richmond, K. et al. 1997. Detecting subject boundaries within text: a language independent statistical approach. In *Proceedings of The Second Conference on Empirical Methods in Natural Language Processing* (EMNLP-2). Brown University, Providence, Rhode Island, USA.

Roberts, R. B. and I. P. Goldstein. 1977. The FR1 Manual. *AIM-409* Artificial Intelligence Lab. M.I.T., Cambridge, Ma.

Russell, B. 1905. On denoting. *Mind*, 14:479–493. Reprinted in 1985, *Logic and Knowledge*, ed. R. C. Marsh. London: George Allen and Unwin.

Russell, B. 1919. Descriptions. In *Introduction to Mathematical Philosophy*. George Allen & Unwin Publishers. Reprinted in 1993, London: Routledge.

Sidner, C. L. 1979. *Towards a computational theory of definite anaphora comprehension in English discourse*. Ph.D. thesis, MIT.

Sidner, C. L. 1978. The use of focus as a tool for disambiguation of definite noun phrases. In Waltz, D. L. (Ed.), *Theoretical issues in natural language processing -2*. University of Illinois at Urbana-Champaign, 25-27, July.

Siegel, S. and N. J. Castellan. 1988. *Nonparametric statistics for the Behavioral Sciences*. 2nd edition. New York London: McGraw-Hill.

Sorace, A. and E. Bard. 1996. Magnitude estimation of linguistic acceptability. *Language*, 72(1):32–68.

Strand, K. 1997. A Taxonomy of Linking Relations. Manuscript.

Strand, K. 1997. Personal communication, October, 29.

Strawson, P. F. 1950. On referring. *Mind*, 59:320–344. Reprinted in *The Philosophy of Language*, ed. A. P. Martinich. New York: Oxford University Press, 1985.

Sundheim, B. M. 1995. Overview of the Results of the MUC-6 Evaluation. In *Proc. of the Sixth Message Understanding Conference* (MUC-6) , pages 13–31. Columbia, Maryland. November 6-8.

Suri, L. and K. McCoy. 1993. Focusing and Pronoun Resolution in Particular Kinds of Complex Sentences. In *Proc. of the Workshop on Centering Theory in Naturally-occurring Discourse*, Institute for Research in Cognitive Science, University of Pennsylvania, Philadephia, PA, May 8–10.

Suri, L. and K. McCoy. 1994. RAFT/RAPR and Centering: A Comparison and Discussion of Problems Related to Processing Complex Sentences. In *Computational Linguistics*, Vol. 20, No. 2, Squibs and Discussions, pp. 301-317, June, 1994.

Vieira, R. and M. Poesio. 1997. Processing definite descriptions in corpora. HCRC Resarch paper HCRC/RP-86.

Vieira, R. and S. Teufel. 1997. Towards Resolution of Bridging Descriptions. In *Proceedings of the ACL Student Session*. Association for Computational Linguistics. Madrid, Spain. July 7-12.

Vilain, M. et al. 1995. A model-theoretic coreference scoring scheme. In *Proc. of the Sixth Message Understanding Conference* (MUC-6) , pages 45-52. Columbia, Maryland. November 6-8.

Wacholder, N. et al. 1997. Disambiguating proper names in text. In *Proc. of the Fifth Conference on Applied Natural Language Processing*, pages 202–208. Washington, DC. March 31-April 3.

Walde, S. S. 1997. *Resolving Bridging Descriptions in High-Dimensional Space*. Studienarbeit. Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart.

Walker, M. and J. Moore. 1997. Empirical Studies in Discourse. *Computational Linguistics*, 23(1):1–12.

Ward, G. 1991. A pragmatic analysis of so-called anaphoric islands. In *Language*, September 67(3):439–473.

Webber, B. L. 1979. *A Formal Approach to Discourse Anaphora*. New York: Garland.