An empirically-based system for processing definite descriptions

Renata Vieira*
Universidade do Vale do Rio dos Sinos

Massimo Poesio†
University of Edinburgh

We present an implemented system for processing definite descriptions in arbitrary domains. The design of the system is based on the results of a corpus analysis previously reported, which highlighted the prevalence of discourse-new descriptions in newspaper corpora. The annotated corpus was used to extensively evaluate the proposed techniques for matching definite descriptions with their antecedents, discourse segmentation, recognizing discourse-new descriptions, and suggesting anchors for bridging descriptions.

1. Introduction

Most models of DEFINITE DESCRIPTION[1] processing proposed in the literature tend to emphasise their anaphoric[2] role ((Heim, 1982) is perhaps the best formalization of this type of theory). This approach is challenged by the results of experiments we reported previously (Poesio and Vieira, 1998), in which subjects were asked to classify the uses of definite descriptions in Wall Street Journal articles according to schemes derived from proposals by Hawkins (1978) and Prince (1981). The results of these experiments indicated that definite descriptions are not primarily anaphoric; about half of the time they are used to introduce a new entity in the discourse. In this paper, we present an implemented system for processing definite

* Universidade do Vale do Rio dos Sinos - UNISINOS, Av. Unisinos 950 - Cx. Postal 275, 93022-000 São Leopoldo RS Brazil. E-mail:renata@exatas.unisinos.br
† University of Edinburgh, ICCS and Informatics 2, Buccleuch Place EH8 9LW Edinburgh UK. E-mail: Massimo.Poesio@ed.ac.uk
1 We use the term definite description (Russell, 1905) to indicate definite noun phrases with the definite article *the*, such as *the car*. We are not concerned with other types of definite noun phrases such as pronouns, demonstratives, or possessive descriptions.
2 Anaphoric expressions are those linguistic expressions used to signal, evoke or refer to previously mentioned entities.

descriptions based on the results of that earlier study. In our system, techniques for recognising discourse-new descriptions play a role as important as techniques for identifying the antecedent of anaphoric ones.

A central characteristic of the work described here is that we intended from the start to develop a system whose performance could be evaluated using the texts annotated in the experiments mentioned above. Assessing the performance of an NLP system on a large number of examples is increasingly seen as a much more thorough evaluation of its performance than trying to come up with counterexamples, and as essential for language engineering applications. These advantages are thought by many to offset some of the obvious disadvantages of this way of developing NLP theories–in particular, the fact that, given the current state of language processing technology, many hypotheses of interest cannot be tested yet (see below). As a result, quantitative evaluation is now commonplace in areas of language engineering such as parsing, and quantitative evaluation techniques are being proposed for semantic interpretation as well, particularly in the Sixth and Seventh Message Understanding Conferences (MUC-6 and MUC-7) (Sundheim, 1995; Chinchor, 1997), which also included evaluations of systems on the so-called COREFERENCE TASK, which includes resolving definite descriptions as a subtask. The system we present was developed to be evaluated in a quantitative fashion, as well; but because of the problems concerning agreement between annotators observed in our previous study, we evaluated the system both by measuring precision/recall against a 'gold standard', as done in MUC, and by measuring the agreement between the annotation produced by the system and those proposed by the annotators.

The decision of developing a system that could be quantitatively evaluated on a large number of examples resulted in an important constraint: we could not make use of inference mechanisms such as those assumed by traditional computational theories of definite description resolution (e.g., (Sidner, 1979; Carter, 1987; Alshawi, 1990; Poesio, 1993)). Too many facts and axioms would have to be encoded by hand in order for theories of this type to be tested even on a medium-sized corpus. Our system therefore is based on a shallow-processing approach more radical even than that attempted by the first advocate of this approach, Carter (1987) or by the the systems that participate in the MUC evaluations (Appelt, 1995; Gaizauskas et al., 1995; Humphreys et al., 1998), since we didn't attempt to fine-tune the system to maximise performance on a particular domain. The system only relies on structural information, on the information provided by pre-existing lexical sources such as WordNet (Fellbaum, 1998), on minimal amounts of general hand-coded information, or on information that could be acquired automatically from a corpus. As a result, the system doesn't really have the resources for resolving correctly those definite descriptions whose interpretation does require complex reasoning (we grouped these in what we call the 'bridging' class). We nevertheless developed heuristic techniques for processing these types of definites as well, the idea being that these heuristics may provide a baseline against which the gains in performance due to the use of commonsense knowledge can be assessed more clearly.[3]

The paper is organised as follows. In Section §2 we summarize the results of our previous corpus study (Poesio and Vieira, 1998). In Section §3 we discuss the model of definite description processing that we adopted as a result of that work and the general architecture of the system. In Section §4 we discuss the heuristics that we developed for resolving anaphoric definite descriptions, recognizing discourse-

3 In fact, it is precisely because we are interested in identifying the types of commonsense reasoning actually used in language processing that we focused on definite descriptions rather on other types of anaphoric expressions such as pronouns and ellipsis, that can be processed much more effectively on the basis of syntactic information alone (Lappin and Leass, 1994; Hardt, 1997).

new descriptions, and processing bridging descriptions. We then discuss in Section §5 how the performance of these heuristics was evaluated using the annotated corpus. Finally, we present in Section §6 the final configuration of the two versions of the system that we developed. In Section §7 we review other systems which perform similar tasks; in Section §8 we present our conclusions and indicate future work.

2. Preliminary Empirical Work

As mentioned above, the architecture of our system is motivated by the results concerning definite description use in our corpus discussed in (Poesio and Vieira, 1998). In this section we briefly review the results presented in that paper.

The corpus. We used a subset of the Penn Treebank I corpus (Marcus, Santorini, and Marcinkiewicz, 1993) from the ACL/DCI CD-rom, containing newspaper articles from the Wall Street Journal. We divided the corpus in two parts: one, containing about 1000 definite descriptions, was used as source during the development of the system; we will refer to these texts as Corpus 1.[4] The other part, containing about 400 definite descriptions, was kept aside during development and used for testing; we will refer to this subset as Corpus 2.[5]

Classifications of Anaphoric Expressions. The best-known studies of definite description use (Hawkins, 1978; Prince, 1992; Fraurud, 1990; Loebner, 1987; Clark, 1977; Sidner, 1979; Strand, 1996) classify definite descriptions on the basis of their relation with their 'antecedent'. A fundamental distinction made in these studies is

4 The texts in question are w0203, w0207, w0209, w0301, w0305, w0725, w0760, w0761, w0765, w0766, w0767, w0800, w0803, w0804, w0808, w0820, w1108, w1122, w1124, and w1137.
5 The articles in this second subset are w0766, 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.

between descriptions which denote the same discourse entity as their antecedent (which we will call ANAPHORIC or, following Fraurud, SUBSEQUENT MENTION), descriptions which denote an object which is in some way 'associated' with the antecedent–e.g., it is part of it, as in *a car ... the wheel* (these definite expressions are called 'associative descriptions' by Hawkins and 'inferrables' by Prince), and descriptions that introduce a new entity into discourse.

In the case of semantic identity between definite description and antecedent, a further distinction can be made depending on the semantic relation between the predicate used in the description and that used for the antecedent. The predicate used in an anaphoric definite description may be a synonym of the predicate used for the antecedent (*a house...the home*), a generalization / hypernym (*a oak...the tree*), and even, sometimes, a specialization /hyponym (*a tree...the oak*). In fact, the NP introducing the antecedent may not have a head noun at all, e.g., when a proper name is used, as in *Bill Clinton...the president*. We will use the term DIRECT ANAPHORA when both description and antecedent have the same head noun, as in *a house...the house*. Direct anaphors are the easiest definite descriptions for a shallow system to resolve; in all other cases, as well as when the antecedent and the definite description are related in a more indirect way, lexical knowledge or, more in general, encyclopedic knowledge are needed.

All of the classifications mentioned above also acknowledge the fact that not all definite descriptions depend on the previous discourse for their interpretation. Some refer to an entity in the physical environment, others to objects which are assumed to be known on the basis of common knowledge (Prince's 'discourse-new / hearer-old' expressions, such as *the pope*), and others still are licensed in virtue of the semantics of their head noun and complement (as in *the fact that Milan won the Italian football championship*).

A Study of Definite Description Use. In the experiments discussed in (Poesio and Vieira, 1998) we asked our subjects to classify all definite descriptions uses in our two corpora. These experiments had the dual objective of verifying how easy it was for human subjects to agree on the distinctions between definite descriptions just discussed, and to produce data that we could use to evaluate the performance of a system. The classification schemes we used were simpler than those proposed in the literature just mentioned and was motivated, on the one hand, by the desire to make the annotation uncomplicated for the subjects employed in the empirical analysis;[6] on the other hand, by our intention to use the annotation to get an estimate of how well a system using only limited lexical and encyclopedic knowledge could do. We ran two experiments, using two slightly different classification schemes; in the first experiment we used the following three classes:[7]

- **direct anaphora**, for subsequent-mention definite descriptions that refer to an antecedent with the same head noun as the description;
- **bridging descriptions** for definite descriptions that either (i) have an antecedent denoting the same discourse entity, but using a different head noun (as in *house* . . . *building*) or (ii) are related by a relation other than identity to an entity already introduced in the discourse;[8]
- **discourse new** for first mention definite descriptions that denote objects not related by shared associative knowledge to entities already introduced in the discourse.

6 Previous attempts to annotate anaphoric relations had resulted in very low agreement levels; e.g., in the coreference annotation experiments for MUC-6 (Sundheim, 1995), relations other than identity were dropped due to difficulties in annotating them.
7 In this experiment, our subjects could also classify a definite description as 'idiomatic' or 'doubt'–see tables below.
8 In (Poesio and Vieira, 1998), Hawkins' term 'associative' was used for this class; but in fact, the definition we used for the class is closest to the sense of 'bridging' used by (Clark, 1977).

In the second experiment, we tried to treat all anaphoric definite descriptions as part of one class, and all inferrables as part of a different class, without significant changes in the agreement results.

The agreement among annotators was measured using the K statistic (Siegel and Castellan, 1988; Carletta, 1996). K measures agreement among k annotators over and above chance agreement (Siegel and Castellan, 1988). The K coefficient of agreement is defined as:

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

where P(A) is the proportion of times the annotators agree, and P(E) the proportion of times that we would expect them to agree by chance. The interpretation of K figures is an open question, but in the field of content analysis, where reliability has long been an issue (Krippendorff, 1980), $K > 0.8$ is generally taken to indicate good reliability, whereas $0.68 \leq K < 0.8$ allows tentative conclusions to be drawn. Carletta et al. (1997) observes however that in other areas, such as medical research, much lower levels of K are considered acceptable (Landis and Koch, 1977).

An interesting overall result of that study was that the only truly reliable distinction that our annotators could make was that between first mention and subsequent mention (K=.76)[9]; the measure of agreement for the 3-way distinction just discussed was K=.73. The second interesting result concerned the distribution of definite descriptions in the three classes above: we found that about half of the definite descriptions were discourse new. The distribution of the definite descriptions in classes in our first experiment according to annotators A and B are shown

9 The K coefficient of agreement measures agreement among annotators over and above chance agreement. It is defined as: is the proportion of times the annotators agree, and P(E) the proportion of times that we would expect them to agree by chance.

in Tables 1 and 2, respectively. (Class IV includes cases of idiomatic expressions or doubts expressed by the annotators).

Class	Total Number	Percentage of the total
I. Direct anaphora	294	28.27%
II. Bridging	160	15.38%
III. Discourse new	546	52%
IV. Others	40	3.84%
Total	1040	100%

Table 1
Classification of definite descriptions according to Annotator A.

Class	Total Number	Percentage of the total
I. Direct anaphora	332	31.92%
II. Bridging	150	14.42%
III. Discourse new	549	52.78%
IV. Others	9	0.86%
Total	1040	100%

Table 2
Classification of definite descriptions according to Annotator B.

The third main result of that study was that we found very little agreement between our subjects on identifying briding descriptions: in our second experiment, the agreement on bridging descriptions was K=0.24. This was due in part to the fact that many definite descriptions could be classified in more than one class –e.g., either anaphoric or bridging, depending on which antecedent was chosen–in part to the fact that in the case of descriptions indirectly related to their antecedents, the discourse might provide more than one distinct anchor all equally suitable (Poesio and Vieira, 1998). The most common classification problem is distinguishing between larger situation and bridging descriptions; see also (Fraurud, 1990; Poesio and Vieira, 1998).

3. A Model of Definite Description Processing Inspired by Empirical Studies

The results just discussed led us to adopt a model of definite descriptions processing advanced in (Fraurud, 1990) and further elaborated in (Poesio and Vieira, 1998), according to which interpreting definite descriptions in written discourse is not just a matter of checking whether there is a suitable antecedent for the description, but also involves a classification task: recognizing whether a description is, in Fraurud's terms, first mention or subsequent mention–or, in our terminology, direct anaphora, discourse new or bridging. The crucial aspect of Fraurud's proposal is the idea that interpreting definite descriptions is not just a matter of looking for an antecedent; separate rules for recognizing first mention definite descriptions are needed as well.

The fact that there was so much disagreement about bridging descriptions and their anchors led us to try to keep the rules for processing them fairly separate from those for processing other types of descriptions, and to attempt using agreement measures to evaluate the performance of the system, in addition to more traditional precision and recall figures.

In this section we discuss our model and then present the overall architecture of the system.

3.1 Fraurud's Proposal; Our Model

The results discussed above lead further support to Fraurud's (1990) criticism of the approach to processing definite NPs based on the assumption that they are primarily anaphoric. Because of the large proportion of first mention definites found in the texts she examined, Fraurud 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 ((Fraurud, 1990), page 421).

Fraurud proposes, *contra* (Heim, 1982), that processing a definite NP always involves establishing a new discourse entity.[10] This new discourse entity may then be linked to one or more ANCHORS in the text (where the term 'anchor' is used to indicate an 'antecedent' either of an anaphoric expression or of a bridging reference) or to a background referent.[11] Fraurud discusses the example of the description *the king*, interpreted relationally, encountered in a text in which no king has been previously mentioned. 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.[12]

The most interesting aspect of Fraurud's proposal is the hypothesis that first mention definites are not necessarily recognized simply because no suitable antecedent has been found; independent strategies for recognizing them may be involved. This hypothesis is consistent with Loebner's proposal (Loebner, 1987) that the fundamental property of a definite description is that it denotes a *function* (in a logical sense); this function can be part of the meaning assigned to the definite description by the grammar (as in *the beginning of X*), or can be specified by context (as in the case of anaphoric definites). Fraurud's and Loebner's ideas can be translated

10 Discourse entities are representations in the discourse model of entities explicitly mentioned (Webber, 1979; Heim, 1982).
11 'Background referents' are entities that have not been mentioned in the discourse–those entities that Grosz (1977) would call 'elements of the implicit focus'.
12 Fraurud does not explain what is it that justifies the use of definite descriptions, if not familiarity. In (Poesio and Vieira, 1998) we suggest that Loebner's proposal (Loebner, 1987) seems to be the one that accounts for the most data.

into a requirement for a system to have separate methods or rules for recognizing discourse new descriptions (and in particular, Loebner's 'semantically functional' definites) in addition to rules for resolving anaphoric definite descriptions; these rules may run in parallel with the rules for resolving anaphoric definites, rather than after them.

Rather than deciding a priori on the question of whether the (in our case) heuristic rules for identifying discourse-new descriptions should be run in parallel with, or after, resolution, we treated this as an empirical question. We made the architecture of the system fairly modular, so that we could both try different heuristics and try to apply them in a different order, using the corpus for evaluation. We discuss all the heuristics that we tried in Section §4, and our evaluation of them in Section §5.

3.2 Architecture of Our System

The overall architecture of our system is shown in Figure 1. The system attempts to classify each definite description as either direct anaphora, discourse-new, or bridging description. In addition to this classification, the system tries to identify the antecedents of anaphoric descriptions and the anchors (Fraurud, 1990) of bridging ones. The system processes parsed newswire texts from the Penn Treebank I, constructing a fairly simple discourse model, consisting of a list of discourse entities that may serve as potential antecedents (hence, POTENTIAL ANTECEDENTS *tout-court*), according to the chosen segmentation algorithm (see below). The system uses the discourse model, syntactic information and a small amount of lexical information to classify definite descriptions as discourse-new or to link them to anchors in the text; WordNet is also consulted by the version of the system that attempts to resolve bridging descriptions. The system is implemented in Sicstus Prolog.

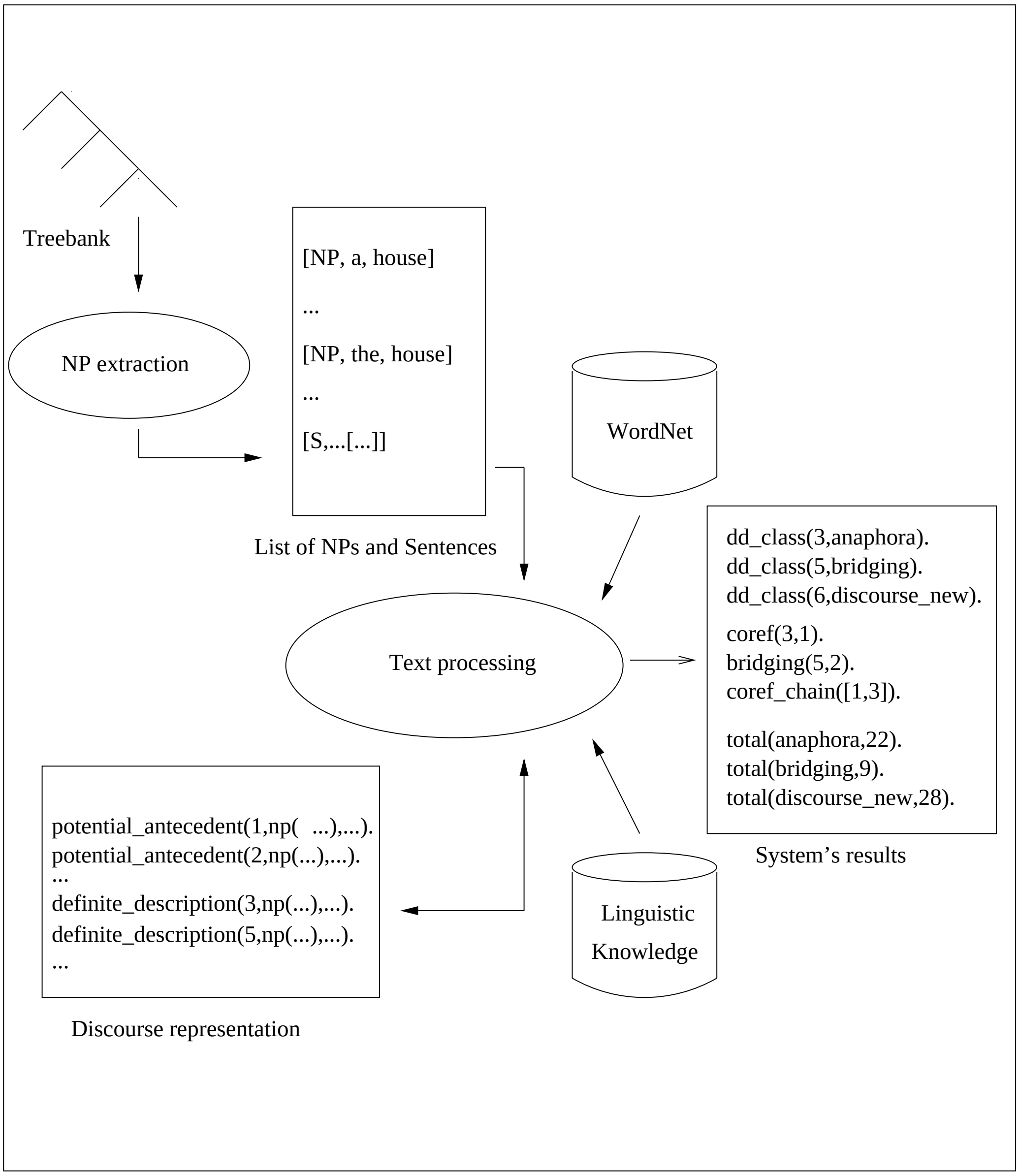

Figure 1
System architecture

Input. The texts in the Penn Treebank corpus consist 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 it contains are extracted. The output of this pre-processing phase is passed to the system proper. For example, the sentence in (1) is represented in the Treebank as (2) and the input to the system after the pre-processing phase is (3)[13]. 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*).

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

(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))))
  .)
```

(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].
```

13 Prolog variables will be indicated in the rest of the paper by the use of "_" in the beginning and in the end of the variables:; e.g., _X_ for variable X.

```
[[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]]]],.].
```

Output. The system outputs the classification it has assigned to each definite description in the text, together with the co-referential and bridging links it has identified.

4. The Heuristics

We developed three types of heuristics:

- for resolving directly anaphoric descriptions. These include heuristics for dealing with segmentation and to handle modification.
- for identifying discourse-new descriptions. Some of these heuristics attempt to recognize semantically functional definite descriptions (Hawkins, 1978; Loebner, 1987), whereas other ones try to recognize definite descriptions that are anchored via their modification (Clark and Marshall, 1981; Prince, 1981).
- for identifying the anchor of a bridging description and the semantic relation between the bridging description and its anchor. WordNet is accessed, and heuristics for named entity recognition were also developed.

We present the heuristics for each class of definite descriptions in turn in this section, and discuss their limitations in Section §4.1, Section §4.2, and Section §4.3, respectively. The final configuration of the system was arrived at on the basis of an extensive evaluation of the heuristics using the corpus annotated in our previous work (Poesio and Vieira, 1998). The evaluation was used both to determine which version of each heuristic worked better, and to identify the best order in which to try them. The results about the performance of these heuristics are presented in Section §5, and the way we organized them in an overall algorithm in Section §6.

4.1 Direct anaphora

Our system's strategy for resolving direct anaphora is very simple: it just looks for a potential antecedent whose head matches the head noun of the definite description. The key issues to address in doing this are:

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

Performing each of these tasks may potentially involve complex syntactic analysis and general reasoning; we will discuss our heuristic solutions to them in turn.

4.1.1 Identifying head nouns. In order to resolve an anaphoric description it is necessary to identify its head noun, which in the parsed texts of the Penn Treebank is generally the rightmost atom in the NP. For example, the nouns *politics* and *squabbling* are the heads of the following NPs:

(4) a. `[NP,Mideast,politics];`
    b. `[NP,the,squabbling,[PP,within,[NP,the,Organization...]]].`

Although this strategy works most of the time, it does have some problems. One of these problems are headless definites (such as *the highest in the southern region*). A second problem are definites whose head is not represented in the Treebank as an atom at the determiner level (such as, [NP,The,[NP,[NP,2000],[NP,tax]]]). Corpus 1, for example, contains 17 definite descriptions with these problems (0.2%). A third problem is coordination: for example, our algorithm does not recognize that a noun such as *reporters* below is a head noun:

[NP,reporters,and,editors,[PP,of,[NP,The,WSJ]]].

4.1.2 Potential antecedents. The second problem is to determine which NPs should be used to resolve definite descriptions, among all those in the text. The system keeps track of NP index, NP structure, head noun, and NP type (definite, indefinite, bare plural, possessive[14]) of each potential antecedent, as illustrated by (5) below.

(5)
```
potential_antecedent(I,np(NP),head(H),type(T)).
```

Examples of potential antecedents extracted from (6) are shown in (7):

(6) 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.

(7) a.
```
potential_antecedent(1,np(_NPstructure_),
                       head(reporters),
                       type(indef)).
potential_antecedent(2,np(_NPstructure_),
                       head(interview),
                       type(indef)).
```

14 Other NPs not included in any of these categories are idenfied as type(other).

b.
```
potential_antecedent(3,np(_NPstructure_),
                     head(Journal),
                     type(def)).
potential_antecedent(4,np(_NPstructure_),
                     head(candidate),
                     type(def)).
potential_antecedent(5,np(_NPstructure_),
                     head(mayoralty),
                     type(def)).
```

c.
```
potential_antecedent(6,np(_NPstructure_),
                     head(ability),
                     type(possessive)).
```

d.
```
potential_antecedent(7,np(_NPstructure_),
                     head(victory),
                     type(other)).
```

We found that different recall/precision trade-offs can be achieved depending on the choice of potential antecedents – i.e., depending on whether all NPs are considered as possible antecedents, or only indefinite NPs, or various other subsets–so we ran experiments to identify the best group of potential antecedents. Four different NP subsets were considered:

1. indefinite NPs, defined as NPs containing the indefinite articles *a, an, some* and bare/cardinal plurals, as in (7a);[15];

2. indefinite NPs and definite descriptions (NPs beginning with the definite article) ((7a) and (7b));

3. indefinite NPs, definite descriptions, and possessive NPs (NPs with a possessive pronoun or possessive mark) ((7a), (7b) and (7c));

4. all NPs ((7a), (7b), (7c))and (7d)).

The results obtained by considering each subset of the total number of NPs as potential antecedents are discussed in Section §5.2.

15 Only plural nouns ending in *s* are handled by the system.

4.1.3 Segmentation. The set of potential antecedents of anaphoric expressions is also restricted by the fact that antecedents tend to have a limited 'life span'—i.e., they only serve as antecedents for anaphoric expressions within pragmatically determined SEGMENTS of the whole text (see, e.g., (Reichman, 1985; Grosz and Sidner, 1986; Fox, 1987)). In our corpus we found that about 10% of direct anaphoric definite descriptions have more than one possible antecedent if segmentation is not taken into account (Vieira and Poesio, 1996). In (8), for example, the antecedent of *the house*$_j$ mentioned in sentence 50 is not the house mentioned earlier in sentences 2 and 19, but another (non-mobile) house implicitly introduced in sentence 49 by the reference to *the yard*.

(8) 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*$_k$ 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.

In general, it is not sufficient to look at the most recent antecedents only: this is because segments are organized hierarchically, and the antecedents introduced in a segment at a lower level are typically not accessible from a segment at a higher level (Fox, 1987; Grosz, 1977; Grosz and Sidner, 1986; Reichman, 1985), whereas the antecedents introduced in a prior segment at the same level may be. Later in (8), for example, *the house*$_j$ in sentence 50 becomes inaccessible again, and in sentence 65, the text start referring again to the house introduced in sentence 2. Recognizing the hierarchical structure of texts is still, however, an open problem, as it involves reasoning about intentions;[16] better results have been achieved on the simpler task of 'chunking' the text into sequences of segments, generally by means of lexical density measures (Hearst, 1997; Richmond, Smith, and Amitay, 1997).

The methods to limit the lifespan of discourse entity we considered for our system are even simpler. One type of heuristics we looked at are window-based techniques, i.e., considering only the antecedents within fixed-size windows of previous sentences, allowing however some discourse entities to take a longer life span: we call this method LOOSE SEGMENTATION. More specifically, a discourse entity is considered as potential antecedent for a definite description when the antecedent's head is identical to the description's head, and

- the potential antecedent is within the established window, or else
- the potential antecedent is itself a subsequent mention, or else

16 See, however, (Marcu, 1999).

- the definite description and the antecedent are identical NPs (including the article).

We also considered an even simpler RECENCY heuristic: this involves keeping a table indexed by the heads of potential antecedents, such that the entry for noun N contains the index of the last occurrence of an antecedent with head N. Finally, we considered combinations of segmentation and recency. The comparative results of the alternative heuristics just described are discussed in Section §5.2.

4.1.4 Noun modifiers. Once the head nouns of the antecedent and of the description have been identified, the system attempts to match them. This simple head-matching strategy works correctly in simple cases like (9):

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

In general, however, when matching a definite description with a potential antecedent the information provided by the prenominal and the postnominal part of the noun phrases also has to be taken into account: so, for example, *a blue car* cannot serve as the antecedent for *the red car*, or *the house on the left* for *the house on the right*. In our corpus, cases of antecedents that would incorrectly match by simply matching heads without regarding premodification include:

(10) a. *the business community... the younger, more activist black political community;*

b. *the population... the voting population.*

Again, taking proper care of the semantic contribution of these premodifiers would, in general, require commonsense reasoning; for the moment, we only developed heuristic solutions to the problem, including:

- allowing an antecedent to match with a definite description if the premodifiers of the description are a subset of the premodifiers of the antecedent. This heuristic deals with definites which contain less information than the antecedent, such as *an old Victorian house... the house*, and prevents matches such as *the business community... the younger, more activist black political community*.
- allowing a non-premodified antecedent to match with any same head definite. This second heuristic deals with definites that contain additional information, such as *a check... the lost check*.

The information that two discourse entities are disjoint may come from postmodification, as well, although same head antecedents with different postmodification are not as common as those with differences in premodification. An example from our corpus is shown in (11).

(11) *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.*

The heuristic method we developed 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.

The results obtained with these heuristics are discussed in Section §5.2, as well.

4.2 Discourse new descriptions

As mentioned above, a fundamental characteristic of our system is that it also includes heuristics for recognising discourse-new descriptions (i.e., definite descriptions that introduce new discourse entities) on the basis of syntactic and lexical features of the noun phrase. Our heuristics are based on the discussion by (Hawkins, 1978), who identified a number of correlations between certain types of syntactic structure and discourse-new descriptions, particularly those he called 'unfamiliar' definites (i.e., those whose existence cannot be expected to be known on the basis of generally shared knowledge), including:[17]

- 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* (Hawkins calls these 'unexplanatory modifiers; Loebner (1987) showed how these predicates may license the use of definite descriptions in an account of definite descriptions based on functionality');
  - — a head noun taking a complement such as *the fact that there is life on Earth* (Hawkins calls this subclass 'NP complements');
- the presence of restrictive modification, as in *the inequities of the current land-ownership system.*

17 Hawkins himself proposes a transformation-based account of unfamiliar definites, but the correlations he identified proved to be a useful source of heuristics for identifying these uses of definite descriptions even though the existence of counterexamples to these heuristics suggests that a syntactic-based account cannot be entirely correct. Most of these examples can be accounted in terms of Loebner's theory of definiteness.

Our system attempts to recognize these syntactic patterns. In addition, we also added heuristics classifying as unfamiliar some definites occurring in

- appositive constructions (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*).

(The reason why definite descriptions in appositive and copular constructions tend to be discourse-new, in fact unfamiliar, is that the information needed for the identification is given by the NP to which the apposition is attached and the predicative part of the copular construction, respectively.)[18]

Finally, we found that three classes of what Hawkins called 'larger situation' definites (those whose existence can be assumed to be known on the basis of encyclopedic knowledge, such as *the pope*) can also be recognised on the basis of heuristics exploiting syntactic and lexical features:

- definites that behave like proper nouns, like *the United States*;
- 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*.

18 In the systems participating in MUC, definite descriptions occurring in appositions are treated as anaphoric on the preceding NP; our system considers the NP and the apposition as a a unit that introduces a new referent to the discourse.

In our corpus study we found that our subjects did much better at identifying discourse-new descriptions all together (K=.68) than they did at distinguish unfamiliar from larger situation cases (K = .63). This finding was confirmed by our implementation: although each of the heuristics is designed, in principle, to identify only one of the uses (larger situation or unfamiliar), they work better as identifying all together the class of discourse new descriptions.

We discuss in detail the heuristics to recognize discourse-new definite descriptions that we tested in our system next.

4.2.1 Special predicates. Some cases of discourse-new definite descriptions can be identified by comparing the head noun or modifiers of the definite NP with a list of predicates that are either functional or likely to take a complement (Loebner, 1987). Our list of predicates that, when taking NP complements, are generally used to introduce discourse-new entities, was compiled by hand and currently includes the nouns *fact, result, conclusion, idea, belief, saying* and *remark*. In these cases, what licenses the use of a definite is not anaphoricity, but the fact that the head noun can be interpreted as semantically functional; the noun complement specifies the argument of the function. Functionality is enough to license the use of the definite description (Loebner, 1987). An example of definite description classified as discourse-new on these grounds is given in (12).

(12) 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.*

When encountering a definite whose head noun occurs in this list, the system checks if a complement is present or if the definite appears in a copular construction (e.g., *the fact is that* ...).

A second list of special predicates consulted by the system includes what Hawkins called UNEXPLANATORY MODIFIERS: these include adjectives such as *first, last, best, most, maximum, minimum, only*, and superlatives in general.[19]. All of these adjectives are predicate modifiers that turn a head noun into a function, therefore again–according to Loebner–licensing the use of a definite even when no antecedent is present (see examples below). When applying this heuristic, the system verifies the presence of a complement for some of the modifiers (*first, last*), but not for superlatives.

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

b. She jumps at *the slightest noise*.

Finally, our system uses a list of special predicates that we found to correlate well with larger situation uses (i.e., definite descriptions referring to objects whose existence is generally known). This list consists mainly of terms indicating time reference, and includes the nouns *hour, time, morning, afternoon, night, day, week, month, period, quarter, year* and their respective plurals. An example from the corpus is:

(14) Only 14,505 wells were drilled for oil and natural gas in the U.S. in the first nine months of *the year*.

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

Other definites typically used with reference to the larger situation uses are *the moon*, *the sky*, or *the pope*, *the weather*.

It should be noted that although these constructions may indicate a discourse-new interpretation, these expressions may also be used anaphorically; this is one of the cases in which a decision has to be made concerning the relative priority of different heuristics. We discuss this issue further in connection with the evaluation of the system performance in Section §5.[20]

4.2.2 Restrictive and non-restrictive modification. A second set of heuristics for identifying discourse-new descriptions that we derived from Hawkins's suggestions and from our corpus analysis look for restrictive modification.[21] We developed patterns to recognize restrictive post-modification and non-restrictive post-modification; we also tested the correlation between discourse novelty and pre-modification (which we always assumed to be restrictive). We discuss each of these heuristics in turn.

Restrictive postmodification. Hawkins (1978) pointed out that unfamiliar definites often include referent establishing relative clauses and associative clauses, while warning that not all relative clauses are referent establishing. Some statistics about this correlation were reported by Fraurud (1990): she found that in her corpus 75% of complex definite NPs (i.e., modified by genitives, postposed PPs, restrictive adjectival modifiers) were first mention. A great number of definite descriptions with restrictive post-modifiers are unfamiliar in our corpus, as well (Poesio and Vieira,

20 More recently, Bean and Riloff (1999) have proposed methods for automatically extracting from a corpus heads that correlate well with discourse novelty.
21 The term 'restrictive modification' is used when the modifier provides information which is essential to identify the discourse entity referred to by the NP (Quirk et al., 1985). The modification is non-restrictive when the head provides sufficient information to identify the discourse entity, so that the information provided by the modification is not essential for identification.

1998); in fact, restrictive postmodification was found to be the single most frequent feature of first mention descriptions. The reason why constructions of this type are good indicators of discourse novelty is that a restrictive postmodifier may license the use of a definite description either by providing a link to the rest of the discourse (as in Prince's 'containing inferrables') or by making the description into a functional concept. Looking for restrictive postmodifiers might therefore be a good way of identifying discourse-new descriptions.

The distribution of restrictive postmodifiers in our corpus is shown in Table 3; examples of each type of postmodifier are given below.

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

Table 3
Distribution of prepositional phrases and relative clauses.

Relative clauses: these are finite clauses sometimes (but not always) introduced by relative pronouns such as *who, whom, which, where, when, why, that*:

(15) a. *The place where he lives...*

b. *The guy we met...*

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

(16) a. *The man writing the letter is my friend.*

b. *The man to consult is Wilson.*

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

Table 4
Distribution of prepositions (1)

Prepositional phrases and of-clauses: Quirk et al. (1985) found that prepositional phrases are the most common 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 (see Table 3). The types of prepositions observed for 188 postmodified descriptions is shown in Table 4; *of*-clauses are the most common.

Our program uses the following patterns to identify restrictive postmodifiers:[22]

(17) 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,_|_]|_].`

In the Treebank, sometimes the modified NP is embedded in another NP, so structures like (18) are also considered (again for all types of clauses just shown above):

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

22 Note that an NP may have zero, one, or more premodifiers.

Non-restrictive postmodification. We found it important to distinguish restrictive from non-restrictive postmodification, since in our corpus definite descriptions with non-restrictive postmodifiers were generally not discourse-new. Our system recognizes non-restrictive postmodifiers by the simple, yet effective heuristic of looking for commas. This heuristic correctly recognizes non-restrictive postmodification in cases like:

(19) The substance, discovered almost by accident, is very important.

which are annotated in the Penn Treebank I as follows:

(20)
```
[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 is not as common in prenominal position as in post-head position, but it is often used, and was also found to correlate well with larger situation and unfamiliar uses of definite descriptions (Poesio and Vieira, 1998). A restrictive premodifier may be a noun (as in (21)) a proper noun, or an adjective.[23] Sometimes numerical figures (usually referring to dates) are used as restrictive premodifiers, as in (22)).

(21) 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.

(22) *the 1987 stock market crash*;

23 Our system cannot distinguish adjectives or verb from nouns in premodification because it works directly off the parsed version of the Treebank, without looking at part-of-speech tags.

The heuristic we tested was to classify definite descriptions premodified by a proper noun as larger situation.

4.2.3 Appositions. During our corpus analysis we found additional syntactic patterns that appeared to correlate well with discourse novelty yet had not been discussed by Hawkins. One such pattern are definite descriptions occurring in appositive constructions: they usually refer to the NP modified by the apposition, therefore there is no need for the system to look for an antecedent. Appositive constructions are treated in the Treebank as NP modifiers; therefore the system recognises an apposition by checking whether the definite occurs in a complex noun phrase with a structure 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, as in the examples in (23).

(23) a. *Glenn Cox, the president of Phillips Petroleum*

b.
```
[NP,[NP,Glenn,Cox],',',[NP,the,president,
                        [PP,of,[NP,Phillips,Petroleum]]]];
```

c. *the oboist, Heinz Holliger*

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

In fact a definite description may itself be modified by an apposition, e.g., an indefinite NP, as shown by (24). Such cases of appositive constructions are also recognized by the system.

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

Our system also has patterns for recognizing cases in which the definite description in appositive position is modified by a relative clause; in the Penn Treebank, these cases are given the structure in (25).

(25)
```
[NP,[NP,Rudolph,Giuliani],',',
                [NP,[NP,the,former,crime,buster]...]].
```

Other examples of apposition recognised by the system are:

(26) a. *the very countercultural chamber group* Tashi;

b. *the new chancellor*, John Major;

c. *the Sharpshooter*, a freshly drilled oil well two miles deep;

4.2.4 Copular phrases. A second type of construction that often involves discourse-new descriptions are copular phrases such as *the Prime Minister is Tony Blair*. We developed the following heuristic for handling copula constructions. 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. Two examples correctly handled by this heuristic are shown in (27); the syntactic representation of these cases in the Penn Treebank I is shown in (28).

(27) 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.

(28) `[S,[NP,The,bottom,line],[VP,is,[NP,[SBAR,that... ]]]].`

If the complement of the verb is an adjective, the subject is typically interpreted referentially and should not be considered as discourse-new on the basis of its complement (cfr. *The president of the US is tall*). Adjectival complements are represented as follows in the Treebank:

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

Definite descriptions in object position of the verb *to be*, such as the one shown in (30), are also considered discourse-new.

(30) 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*.

4.2.5 Proper names. Proper names preceded by the definite article, such as (31), are common in the genre we are dealing with, newspaper articles.

(31) *the Securities and Exchange Commission*.

The first appearance of these definite descriptions 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 simply checks whether the head is capitalised. If the test succeeds, the definite is classified as a larger situation use.[24]

This concludes the discussion of our heuristics for the identification of discourse new descriptions. Their performance is evaluated in Section §5.3.

4.3 Bridging descriptions

Bridging descriptions are the class of definite descriptions which a shallow processing system is least equipped to handle. Linguistic and computational theories of bridging descriptions identify two main subtasks involved in their resolution: finding the element in the text to which the bridging description is related (ANCHOR) and identifying the relation (LINK) holding between the bridging description and its anchor (Clark, 1977; Sidner, 1979; Heim, 1982; Carter, 1987; Fraurud, 1990; Strand, 1996). The speaker is licensed to use a bridging description when he/she can assume that the commonsense knowledge required to identify the relation is shared by the listener (Hawkins, 1978; Clark and Marshall, 1981; Prince, 1981). This dependence on commonsense knowledge means that, in general, a system can only resolve bridging descriptions when supplied with an adequate knowledge base; for this reason, the typical way of implementing a system for resolving bridging descriptions has been to restrict the domain and feed the system with hand-coded world knowledge (see, e.g., (Sidner, 1979) and especially (Carter, 1987)). A broader view on the bridging phenomena (not only bridging descriptions) is presented in (Hahn, Strube, and Markert, 1996). They make use of a khowledge base from which they extract conceptual links to feed an adaptation of the centering model (Grosz, Joshi, and Weinstein, 1995).

24 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.

Furthermore, the relation between bridging descriptions and their anchors may be arbitrarily complex (Clark, 1977; Sidner, 1979; Prince, 1981; Strand, 1996). and the same description may relate to different anchors in a text: this makes it difficult to decide what the intended anchor and the intended link are (Poesio and Vieira, 1998). For all these reasons, this class has been the most challenging problem we dealt with in the development of our system, and the results we obtained so far can only be considered very preliminary. Nevertheless, we feel that trying to process these definite descriptions is the only way to discover which types of commonsense knowledge are actually needed.

4.4 Types of bridging descriptions

Our work on bridging descriptions began with the development of a classification of bridging descriptions (Vieira and Teufel, 1997) according to the kind of information needed to resolve them, rather than on the basis of the possible relations between descriptions and their anchors as usually done in the literature. This allowed us to get an estimate of what types of bridging descriptions we might expect out system to resolve. The classification is as follows:

- cases based on well-defined lexical relations, such as synonymy, hypernymy and meronymy, that can be found in a lexical database such as WordNet (Fellbaum, 1998)–as in *the flat ... the living room*;
- bridging descriptions in which the antecedent is a proper name and the description a common noun, whose resolution requires some way of recognizing the type of object denoted by the proper name (as in *Bach ... the composer*);

- cases in which the anchor is not the head noun but a noun modifying an antecedent, as in *the company has been selling discount packages ... the discounts*

- cases in which the antecedent (anchor) is not introduced by an NP but by a VP, as in *Kadane oil is currently drilling two oil wells. The activity ...*

- descriptions whose the antecedent is not explicitly mentioned in the text, but is implicitly available because it is a discourse topic–e.g., *the industry* in a text referring to oil companies;

- cases in which the relation with the anchor is based on more general commonsense knowledge, e.g., about cause-consequence relations.

We developed heuristics for handling the first three of these classes: lexical bridges, bridges based on names, and bridges to entities introduced by non-head nouns in a compound nominal (Poesio, Vieira, and Teufel, 1997). We describe these heuristics in the rest of this section.

4.4.1 Bridging descriptions and WordNet. In order to get a system that could be evaluated on a corpus containing texts in different domains, we used WordNet (Fellbaum, 1998) as an approximation of a lexical knowledge source. 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*.

Sometimes, finding a relation between two predicates involves complex searches through WordNet's hierarchy. E.g., in some cases, there is no relation between two head nouns, but there is a relation between compound nouns in which these nouns appear: thus, there is no semantic relation between *record/album*, but only a synonymy relation between *record_album/album*. We found that extended searches of this type, or searches 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 WordNet (Poesio, Vieira, and Teufel, 1997).

The results of our tests with WordNet are presented in section Section §5.4.

4.4.2 Bridging descriptions and named entity recognition. Definite descriptions that refer back to entities introduced by 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: e.g., if we recognize *Pinkerton Inc.* as an entity of type **company**, we can then resolve the subsequent description *the company*, or even a description such as *the firm* by finding out a synonymy relation between **company** and **firm** using WordNet.

This so-called NAMED ENTITY RECOGNITION task has received considerable attention recently (Mani and MacMillan, 1996; McDonald, 1996; Paik et al., 1996; Bikel et al., 1997; Palmer and Day, 1997; Wacholder and Ravin, 1997; Mikheev, Moens, and Grover, 1999) and was one of the tasks evaluated in the Sixth and Seventh Message Understanding Conferences; in MUC-6, 15 different systems participated in the competition (Sundheim, 1995). For the version of the system discussed and evaluated here, we implemented a preliminary algorithm for named entity recognition that we developed ourselves; a more recent version of the system (Ishikawa, 1998) uses the named entity recognition software developed by HCRC for the MUC-7 competition (Mikheev, Moens, and Grover, 1999).

WordNet contains the types of a few names —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. Our algorithm for assigning a type to proper names is 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.

The recall of this algorithm was increased by including a back-tracking mechanism which re-processes a text filling in the discourse representation with missing name types. With this mechanism we can identify later the type for the name *Morishita* in a textual sequence in which the first occurrence of the name does not provide surface indication of the entity type: e.g., *Morishita* — *Mr. Morishita*. The second mention includes such a clue (*Mr.*); by processing the text twice we recover such missing types.

After finding the types for names, the system uses the techniques previously described for same head matching or WordNet lookup to match the descriptions with the types found for previous named entities.

4.4.3 Compound nouns. Sometimes, the anchor for a bridging description is a non-head noun in a compound noun:

(32) *stock market crash... the markets*;

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*.[25] 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 anaphora (Postal, 1969): for example, it is generally accepted that in (33), *a deer* is not accessible for anaphoric reference.[26]

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

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

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

(34) *the stock market crash... the markets*;

25 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].

26 These proposed constraints have been challenged by (Ward, Sproat, and McKoon, 1991).

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

(35) *his art business... the art gallery.*

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

(36) *a 15-acre plot and main home... the home site.*

5. Evaluation of the Heuristics

In this section we discuss the tests we ran to arrive at a final configuration of the system. The performance of the heuristics discussed in Section §4 was evaluated by comparing the results of the system with the human annotation of the corpus produced during the experiments discussed in (Poesio and Vieira, 1998). Several variants of our heuristics were tried using corpus 1 as training data; after deciding upon an optimal version, our algorithms were evaluated using corpus 2 as test data. Because our proposals concerning bridging descriptions are much less developed than those concerning anaphoric descriptions and discourse-new descriptions, we ran separate evaluations of two versions of the system: one version which does not attempt to resolve bridging descriptions, here called version 1, and a version 2 which does; we will point out below which version of the system is considered in each evaluation.

In this section, we first explain the evaluation methods that we used. Next, we present a comparative analysis of alternative versions of the heuristics dealing with the resolution of same head anaphora. We then discuss the results obtained by different versions of our heuristics for identifying discourse new descriptions. After that, we analyse our heuristics for bridging descriptions.

5.1 Evaluation methods

The fact that the annotators working on our corpus did not always agree either on the classification of a definite description or on its anchor raises the question of how to evaluate the performance of our system. We tried two different approaches: evaluating the performance of the system by measuring its precision and recall against a standardized annotation based on majority voting (as done in MUC), and measuring the agreement of the system with the rest of the annotators by means of the same metric used to measure the agreement among the annotators themselves (the Kappa statistic). We used the first form of evaluation both to measure the performance of the single heuristics and to measure the performance of the system as a whole; the agreement measure was only used to measure the overall performance of the system. We discuss each of these in turn.[27]

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

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

whereas PRECISION is the percentage of correctly reported results in relation to the total reported:

27 For a rather thorough discussion of the problem of evaluating anaphora resolution algorithms, see (Mitkov, 2000).

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

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 Semi-automatic evaluation against a standardized annotation. The precision and recall figures for the different variants of the system were obtained by comparing the classification produced by each version with a standardized annotation, extracted from the annotations produced by our human annotators by majority judgement: as we had 3 different coders, whenever at least two of them agreed on a class, that class was chosen. Details on how the standard annotation was obtained are given in (Vieira, 1998).[28]

The system's performance as a classifier was automatically evaluated against the standard annotation of the corpus as follows. Each NP in a text is given an index:

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

28 An alternative way of doing this has been proposed by Hatzivassiloglou and McKeown (1993). These authors give fractional values to a classification depending on the number of agreements.

When a text is annotated or processed, the coder / system associates each index of a definite description with a type of use; both the standard annotation and the system's output are represented as Prolog assertions.

(38) a. `system: dd_class(135,anaphoric).`
b. `coder:  dd_class(135,anaphoric).`

In order to assess the system's performance with regards to the identification of a co-referential antecedent, it is necessary to compare the links that indicate the antecedent of each description classified as anaphora. These links are also represented as Prolog assertions, as follows:

(39) a. `coder:  coref(135,106).`
b. `system: coref(135,106).`

The system uses these assertions to build an equivalence class of discourse entities, called a COREFERENCE CHAIN. When comparing an antecedent indicated by the system for a given definite description with that in the annotated corpus, the corresponding co-reference chain is checked–that is, 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 (41.a) and (41.b) would be evaluated as correct answers if the corpus is annotated with the links shown in (40) .

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

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

(41) a. `system: coref(154,135).`
b. `system: coref(154,106).`

In the end, we still need to check the results manually, because our annotated coreference chains are not complete: our annotators did not annotate all types of anaphoric expressions, so it may happen that the system indicates as antecedent an element outside an annotated coreference chain, such as a bare noun or possessive. In (42), for example, suppose that all references to *the house* are coreferential:

(42) A house^{106}... The house^{135}... His house^{140}... The house^{154}...
`coref(154,140).`

if NP 135 is indicated as the antecedent for NP 154 in the corpus annotation (so that 140 is not part of the annotated coreference chain), 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. A second consequence of the fact that the coreference chains in our standard annotation are not complete 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 produced by the system are. By contrast, in the evaluation of the MUC coreference task, where all types of referring expressions are considered, the resulting co-reference chains are evaluated instead of just the indicated antecedent (Vilain, 1995). Even our limited notion of coreference chain was, nevertheless, very helpful in the automatic evaluation, considerably reducing the number of cases to be checked manually.

5.1.3 Measuring the agreement of the system with the annotators. Because the agreement between our annotators in (Poesio and Vieira, 1998) was often only partial, in addition to precision and recall measures we also evaluated the system's performance by measuring its agreement with the annotators using the K statistic already used in (Poesio and Vieira, 1998) to measure the agreement among annotators.

Because the interpretation of K figures is an open question, we interpret the K figures resulting from our tests in a comparative way (by comparing better and worse agreements).

5.2 Anaphora Resolution

We now come to the results of the evaluation of alternative versions of the heuristics dealing with the resolution of direct anaphora (segmentation, selection of potential antecedents and premodification) discussed in Section §4.1. The optimal version of our system is based on the best results we could get for resolving direct anaphora, because we wanted to establish the co-referential relations among discourse NPs as precisely as possible.

Lifespan of discourse entities. In Section §4.1 we discussed two heuristics for limiting the lifespan of discourse entities. The first segmentation heuristic discussed there, 'loose segmentation', is window-based, but 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).

With loose segmentation it is possible for the system to identify more than one coreference link for a definite description: all antecedents satisfying the requirements within the current window 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.

In Section §4.1 we also discussed a second 'segmentation' heuristic, that we called "recency": the system does not collect all candidate NPs as potential antecedents, but only keeps the last occurrence of an NP from all those having the same head noun, and there are no restrictions regarding the antecedent's distance.

The results of these two methods for different window sizes are shown in Table 5. (The results in this table were obtained considering as potential antecedents indefinites (i.e., NPs with determiners *a*, *an*, *some*, bare NPs, and cardinal plurals), possessives and definite descriptions, as in (Vieira and Poesio, 1996); we also used the premodification heuristics proposed there. Alternatives to these heuristics were also evaluated; the results are discussed 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
Evaluation of loose segmentation and recency heuristics

The resulting F measures were almost the same for all heuristics, but there was clearly an increase in recall with a loss of precision when enlarging the window size.[29] 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.

To show what happens when a strict version of the window-based segmentation approach is used, consider Table 6. (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 results in higher precision, but has a strong negative effect on recall. The overall F values are all worse than for the heuristics in Table 5.

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 6
Evaluation of the strict segmentation heuristic

Finally, we tried a combination of the recency and segmentation heuristics: just one potential antecedent for each different head noun is available for resolution, the last occurrence of that head noun. The resolution still respects the segmentation heuristic (loose version). The results are presented in Table 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 7
Combining loose segmentation and recency heuristics

29 In our experiments small differences in recall, precision and F measures are frequent. We generally assume in this paper that such differences are not significant, but a more formal significance test along the lines of that done in (Chinchor, 1995) will eventually be necessary to verify this.

This table shows that by combining the recency and loose segmentation approach to segmentation we obtain a better trade-off between recall and precision than using each heuristic separately. The version with higher recall in Table 7 (4 sentence-window plus recency) was chosen as 'standard' and used in the tests discussed in the rest of this section.

Potential antecedents. Next, we evaluated the various ways of restricting the set of potential antecedents discussed in Section §4.1, using 4 sentence-window loose segmentation with recency. In an earlier version of the system (Vieira and Poesio, 1996) only those definite descriptions 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 could be used to resolve all subsequent mentions co-specifying with that definite description.) An important difference between that implementation and the current one is that in the new version, the definites resolved by the system are also made available as potential antecedents of subsequent definites. The reason for this is that in our previous prototype errors in identifying an indefinite antecedent were sometimes propagated through a coreference chain, so that the right antecedent would be missed. The results are shown in Table 8.

Antecedents selection	R	P	F
Indefinites, def. descriptions , and possessives	75.96%	87.77%	81.44%
All NPs	77.88%	86.17%	81.81%
Indefinites and def. descriptions	73.39%	88.41%	80.21%
Indefinites only	12.17%	77.55%	21.05%

Table 8
Evaluation of the heuristics for choosing potential antecedents

If we only consider indefinites as potential antecedents recall is extremely low

(12%); we also get the worst precision. In other words, considering only indefinites for the resolution of definite descriptions is too restrictive; this is because our corpus contains a large number of first mention definite descriptions which serve as antecedents for subsequent reference (similar results were also reported in (Fraurud, 1990)). The version 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 as basis for further testing a version which combines near-optimal values for F and precision, i.e., the version which takes indefinites, definite descriptions and possessives (first row in Table 8).

Premodifiers. Finally, we tested our heuristics for dealing with premodifiers. We tested again the matching algorithm from (Vieira and Poesio, 1996) in the present version of the system; the results are presented in Table 9. In that Table we also show the results obtained with a modified matching algorithm including a third rule, that allows 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). We tested each of these three heuristics alone and their combinations. (The fourth line simply repeats the results shown in Table 7.)

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 9
Evaluation of the heuristics for premodification (version 1)

The main result of this evaluation is that using a modified segmentation heuristic (including recency) reduces the overall impact of the heuristics for premodification on the performance of the algorithm in comparison with the system discussed in (Vieira and Poesio, 1996). The best precision is still achieved by the matching algorithm that does not allow for new information in the anaphoric expression, but the best results overall are again obtained by combining rule 1 and rule 2, although either 2 or 3 works equally well when combined with 1. (Note that the combination of heuristics 2 and 3 is equivalent to heuristic 3 alone, since 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 not very frequent.

One of the problems with our premodifier heuristics is that although a difference in premodification usually indicates non-coreference, as for *the company's abrasive segment* and *the engineering materials segment*, there are a few cases in our corpus in which co-referent descriptions have totally different premodification from their antecedents, as in:

(43) *the pixie-like clarinetist... the soft-spoken clarinetist.*

These cases would be hard even for a system using real commonsense reasoning, since often the information in the premodifier is new; we think of these examples as one of the best arguments for including in the system a focus-tracking mechanism along the lines of (Sidner, 1979). Our heuristic matching algorithm also suggests wrong antecedents in cases like *the rules* in (44), when the last mention refers to a modified concept (the new rules are different from the previous ones).

(44) Currently, *the rules* force executives...

The rule changes would...

The rules will eliminate...

Finally, the matching algorithm gets the wrong result in cases such as *the population... the voting population* where the new information indicates a subset, superset or part of a previous mentioned referent.

Overall results for anaphoric definite descriptions. To summarize, on the basis of the tests just discussed, the heuristics that achieve the best results, as far as anaphoric definite descriptions are concerned, are:

1. combined loose segmentation and recency,
2. four 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 when the antecedent has no premodifiers.

In Table 10 we present the overall results of this version of the system on anaphora classification and anaphora resolution for the version of the system that does not attempt to resolve bridging descriptions, for both training data and test data . The reason why there are different figures for anaphora resolution and classification is that that the system may correctly classify a description as anaphoric, but then find the wrong antecedent.

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 10
Evaluation of the heuristics for direct anaphora (version 1)

We used this set of heuristics when evaluating the heuristics for discourse-new and bridging descriptions in the rest of the paper.

Examples of errors in anaphora resolution. Before discussing the results of the other heuristics used by the system, we will discuss in more detail some of the errors in the resolution of anaphoric descriptions made by using the heuristics just discussed.

Some errors are simply caused by misspellings in the Treebank: e.g., the example below, where the antecedent is misspelled as *spokewoman*.

(45) *A Lorillard spokewoman... The Lorillard spokeswoman*

The most common problems are due to the heuristics limiting the search for antecedents. In (46), both sentence 7 and sentence 30 are outside the window considered by the system when trying to resolve *the adjusters* in 53.

(46) 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

Limiting the type of potential antecedents to indefinites, definite descriptions and possessives, while improving precision, also leads to problems, because the antecedents introduced by other NPs, such as proper names, are missed–e.g., *Toni Johnson* in (47). The following definite description is then classified by the system as larger situation/unfamiliar. Some of these problems are corrected in version 2 of the system, that also attempts to handle bridging descriptions and therefore uses algorithms for assigning a type to such entities; see below.

(47) *Toni Johnson* pulls a tape measure across the front of what was once a stately Victorian home.

...

The petite, 29-year-old Ms. Johnson...

One case in which the premodification heuristics prevent the system from finding the right antecedent are the (rare) cases of co-referent descriptions with different premodifiers, as in (48).

(48) *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*.

In the following example, it is the lack of a proper treatment of postmodification that causes the problem. The system classifies the description *the earthquake-related claims* as anaphoric to *claims from that storm*, but it is discourse new according to the standardized annotation.

(49) 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*.

Of course, the system cannot handle the cases when the modifiers specify properties of a discourse entity which were not specified by the NP which introduced it. In (50), the system correctly classifies the definite description *the law* as anaphoric, but suggests as antecedent *an income tax law*, whereas a majority of our annotators indicated *a money lending law* as the antecedent.[30]

(50) 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*

30 *The law* could also be interpreted as referring to "the law system in general," in which case none of the antecedents would be correct (or either could be taken as anchor for a bridging interpretation of the definite).

permitted, and attempting to evade income taxes by using a double accounting system.

Finally, the system is uncapable of resolving plural references to collections of objects introduced by singular NPs, even when these collections were introduced by coordinated noun phrases. Although it would be relatively easy to add rules for handling the simplest cases (possibly at the expense of a decrease in precision), many of these references can only be resolved by means of non-trivial operations.

(51) The owners, *William and Margie Hammack*, are luckier than many others.

...

The Hammacks...

5.3 Identification of discourse new descriptions

The overall recall and precision results for the heuristics for identifying discourse new descriptions presented in Section §4.2 are shown in Table 11. In this table we do not distinguish between the two types of discourse-new descriptions, 'unfamiliar' and 'larger-situation' (Hawkins, 1978). As already mentioned in Section §4.2, distinguishing between the two types of discourse-new descriptions identified by Hawkins, Prince and others isn't easy even for humans (Fraurud, 1990; Poesio and Vieira, 1998); and indeed, our heuristics for recognizing discourse-new descriptions work better when evaluated together. The column headed by (#) represents the number of cases of descriptions classified as discourse new in the standard annotation; + indicates the total number of discourse-new descriptions correctly identified; - the number of errors. These results are for the version of the system that uses the best version of the heuristics for dealing with anaphoric descriptions discussed above, and that doesn't attempt to resolve bridging descriptions (version 1).

Discourse new	#	+	-	R	P	F
Training data	492	368	60	75%	86%	80%
Test data	218	151	58	69%	72%	70%

Table 11
Evaluation of the heuristics for identifying discourse new descriptions

The performance of the specific heuristics discussed in Section §4.2 is analyzed in Tables 12 to 15. Table 12 shows the results of the heuristics for larger situation uses on the training data, whereas Table 13 reports the performance on the same data of the heuristics for unfamiliar uses. The reason why only precision figures are reported is that our standard annotation only gives us information about the classification of these discourse descriptions as larger situation and unfamiliar, not about the reason why they were classified in a certain way. The most common feature of discourse new descriptions is postmodification; the least satisfactory results are those for proper names in premodification. As expected, the heuristics for recognising unfamiliar uses (many of which are licensed by linguistic knowledge) achieve better precision than those for larger situation uses, that depend more on commonsense knowledge.

Tables 14 and 15 summarise the results of the heuristics for discourse new descriptions on the test data (corpus 2). Again, the best results were obtained by the heuristics for recognizing unfamiliar uses. The biggest difference in performance was shown by the heuristic checking the presence of the definite in a copula construction, which performed very well on the training data, but poorly on the test data. The actual performance of that heuristic is difficult to evaluate, however, as a very low recall was reported for both training and test data.

Larger Situation	Total found	Errors	Precision
Names	73	10	86%
Time references	50	7	86%
Premodification	41	19	54%
Total	164	36	78%

Table 12
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 13
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 14
Evaluation of heuristics for larger situation uses (test data)

Examples of errors in identifying discourse new descriptions. We will now analyze some of the problems encountered by the version of the system using these heuristics.

Apposition:. Coordinated NPs with more than two conjuncts are a problem for this heuristic, since in the Penn Treebank I coordinated NPs have a structure that matches the pattern used by the system for recognizing appositions. For example, the coordinated NP in the sentence *G-7 consists of the U.S., Japan, Britain, West Germany, Canada, France and Italy* has the structure in (52).

(52)
```
[NP,[NP,the,U.S.],,,[NP,Japan],,,[NP,Britain],,,[NP,West,
 Germany],,,[NP,Canada],,,[NP,France],and,[NP,Italy]]
```

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 15
Evaluation of heuristics for unfamiliar uses (test data)

Copula:. This heuristic was difficult to evaluate because there not too many examples, and the precision in the two data sets is very different (see Tables 13 and 15 above). One problem is that the descriptions in copula constructions might also be interpreted as bridging descriptions. For instance, the description *the result* in (53.a) below is the result of something mentioned previously, while the copula construction specifies its referent. Other ambiguous examples are (53.b) and (53.c):

(53) 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 proper nouns in premodifier positions are often used with discourse new definites (e.g., *the Iran-Iraq war*), they may also be used as additional information in associative or anaphoric uses:

(54) 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 may wrongly be classified as discourse new. In (55) *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.

(55) *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 nouns:. As we have already seen—(47), repeated below as (56)— a definite description that looks like a proper noun (*the petite, 29-year-old Ms. Johnson*) may in fact be anaphoric. This is not always a problem as the system does attempt to find antecedents for these definites, as well, but if the antecedent is not found (as in the example below) the description is incorrectly classified as discourse new.

(56) *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 classified as discourse new a time reference (*the same time*) which is classified as bridging in the standard annotation.

(57) 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 Bridging Descriptions

As mentioned in Section §2, our corpus annotation experiments showed bridging descriptions to be the most difficult class for humans to agree on. Even when our annotators agreed on a bridging description belonging to that class, different anchors would be available in the text for the interpretation of that bridging description. This makes the results of the system for this class very difficult to evaluate; also, the results have to be evaluated by hand.

We first tested the heuristics individually on the training data (the same data used in a previous analysis of the performance of our system on bridging descriptions, (Vieira and Teufel, 1997)) by adding them to version 1 of the system one at a time. These separate tests were manually evaluated. We then integrated all of these heuristics into a version of the system called version 2, using both automatic and manual evaluation. In this section we only discuss the results of the individual heuristics; the overall results of version 2 are discussed in Section §6.

Bridging descriptions are much more sensitive than other types of definite descriptions to the local focus (Sidner, 1979); for this reason, version 2 uses a different search strategy for bridging descriptions than for other definite descriptions. Rather than considering all definite descriptions in the current window simultaneously, it goes back one sentence at a time and stops as soon as a relation with a potential anchor is found.

5.4.1 Using WordNet to identify anchors. Our system consults WordNet to determine if a definite description may be semantically related to one of the NPs in the previous five sentences.[31] The results of this search over our training corpus, in which 204 descriptions were classified as bridging, are shown in Table 16. It is interesting to note that the semantic relations found in this automatic search were not always those observed in our manual analysis.

Bridging Class	Relations Found	Right Anchors	% Right
Synonymy	11	4	36%
Hyponymy	59	18	30%
Meronymy	6	2	33%
Sister	30	6	20%
Total	106	30	28%

Table 16
Evaluation of the search for anchors using WordNet

The main reason why the figures are so low is 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. In only about a third of the cases, a potential antecedent for which we could find a semantic relation in WordNet was an appropriate anchor. An example is (58): although there is a semantic relation between *argument* and *information* in WordNet, the description *the*

31 We found that for bridging descriptions, a five sentence window worked better than a four sentence one.

argument is related to the VP *contend* rather than to the NP *information*. Some form of focusing seems to play a crucial role in restricting the range of antecedents (see also the discussion in (Hitzeman and Poesio, 1998)).

(58) 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.

Sense ambiguity is responsible for some of the false positives. For instance, the noun *company* has at least two distinct senses: *visitor* (as in *I have company*) 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. A more important problem, however, is the incompleteness of the information encoded in WordNet. To have an idea of how complete the information in WordNet is concerning the relations that are encoded, we selected from our two corpora 70 bridging descriptions which we had manually identified as being linked to their anchors by one of the semantic relations encoded in WordNet– synonymy, hypernymy (hyponymy) and meronymy (holonymy). In Table 17 we show the percentages of such relations actually encoded in WordNet. (The fourth column in the table indicates the cases in which the expected relation is not encoded, but the two nouns are sisters in the hierarchy.)

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 17
Evaluation of the encoding of semantic relations in WordNet

As we can see from the table, the recall figure was quite disappointing, especially for synonymy relations. In some cases, the problem was simply that some of the words we looked for were not in WordNet: examples include *newsweekly (news-weekly), crocidolite, countersuit (counter-suit)*. Other times, the word we looked for was contained in WordNet, but not in the same typographic format as it was presented in the text; for example we had *spinoff* in a text, whereas WordNet had only an entry for *spin-off*. A second source of problems was the use in the WSJ articles of domain-specific terminology with context-dependent senses, such as *slump, crash* and *bust*, which in articles about the economy are all synonyms. Finally, in other cases the relations were missing due to the structure of WordNet: for instance, in WordNet the nouns *room, wall, floor* are encoded as part of *building* but not of *house* (see Figure 2).

In summary, our tests have shown that the knowledge encoded in WordNet is not sufficient to interpret all semantic relations between a bridging description and its antecedent found in the kind of texts we are dealing with: only 46% of the relations observed were encoded in WordNet. The possibility of using domain-specific, automatically acquired lexical information for this purpose is being explored: see, e.g., (Poesio, Schulte im Walde, and Brew, 1998). In addition, we found that just looking for the closest semantic relative is not enough to find anchors for bridging descriptions; this search has to be constrained by some type of focusing mecha-

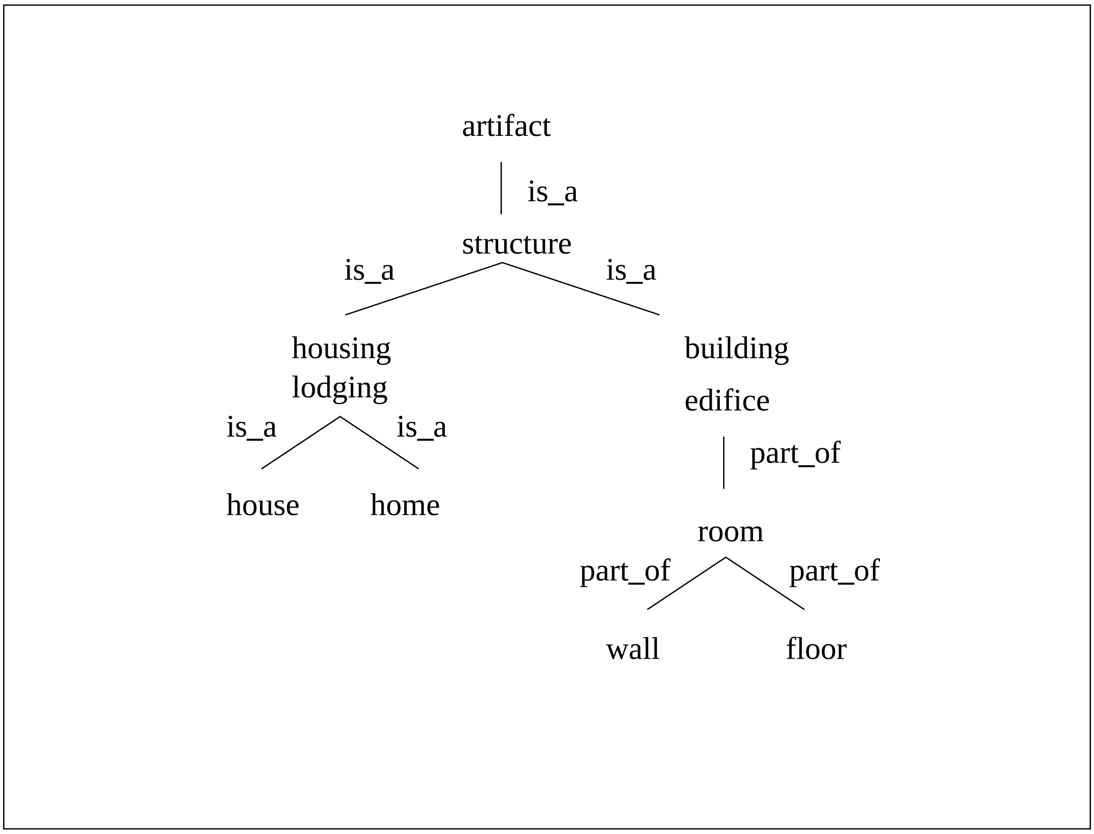

Figure 2
An Example of Problematic Organization in WordNet

nism.

5.4.2 Evaluating the results for bridging descriptions based on proper names.

Identifying named entity types is a pre-requisite for resolving descriptions based on names. The simple heuristics discussed in Section §5.4 identified entity types for 66% (535/814) of all names in the corpus (organizations, persons and locations). The precision was 95%. [32] 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.* and *Seymour Cray*. When looking in WordNet for a type for the name *Steve Reich* we found for the name *Reich* the type **country**. These problems have also been noted by the authors of systems participating in MUC-6 (Appelt, 1995). We also found undesirable relations such as hypernymy for **person** and **company**.

5.4.3 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 segmentation. Sometimes the spelling of the premodification was slightly different from the one of the description, as in *a 15-acre plot... the 15 acres*. Another reason for the low recall was that we included in this class those cases in which the head nouns of antecedent and description are identical,

32 By comparison, the systems participating in MUC-6 had a recall for the named entity task ranging from 82% to 96%, and precision from 89% to 97%, but using comprehensive lists of cue words, or consulting dictionaries of names. 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.). In MUC-7, the best combined precision score, 93.39%, was achieved by the system from LTG in Edinburgh, (Mikheev, Moens, and Grover, 1999), that doesn't use such knowledge sources. We used this system in a version of our prototype that only attempts to resolve bridging descriptions (Ishikawa, 1998).

but the premodification indicates that the entity referred to is not the same: as in *Italy's unemployment rate... the southern unemployment rate*.

6. Overall Evaluation of the System

As mentioned above, we implemented two versions of the system. Version 1 only resolves direct anaphora and identifies discourse new descriptions; version 2 also deals with bridging descriptions. Both versions of the system have at their core a decision tree in which the heuristics discussed in the previous sections are tried in a fixed order to classify a certain definite description and find its anchor. Determining the optimal order of application of the heuristics in the decision tree is crucial to the performance of the system. In version 1 and 2 of the system we used a decision tree developed by hand on the basis of extensive evaluation; we also attempted to determine the order of application automatically, by means of decision tree learning algorithms (Quinlan, 1993).

In this section we first present the hand-crafted decision tree and the results obtained using this decision tree for version 1 and version 2; we then present the results concerning the agreement between system and annotators, and we briefly discuss the results obtained using the decision tree acquired automatically.

6.1 Integration of the heuristics

The hand-crafted order of the heuristics in both versions is as follows. For each NP of the input,

1.The system assigns an index to it.

2. The NPs which may serve as potential antecedents are made available for description resolution by means of the optimal selection criterion discussed in Section §4.1.

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 (see Section §4.2).

   (b) Check whether the definite NP occurs in an appositive construction; if so, there is no need to find an antecedent for it. The NP is classified as discourse new, unfamiliar use.

   (c) Try to find an antecedent for the definite description among the antecedents that are accessible according to combined loose segmentation and recency, by matching head nouns and dealing with premodification and postmodification (see Section §4.1 and Section §5.2). If the system 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 is capitalised). If so, the description is classified as discourse new, larger situation use.

   (e) Check if the definite has a restrictive postmodifier. Definites which are not anaphoric and have restrictive postmodifiers are classified as discourse new, unfamiliar uses.

   (f) Check if there is a proper noun in premodifier position; if so, 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, Version 1 of the system stops; Version 2 start searching for an anchor going backwards one sentence at a time and according to the following heuristics (in this 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 decision tree encoded by this algorithm is shown in Figure 3.

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

- eliminate first some non-anaphoric cases using 'safe' heuristics (first two tests);[33]
- if that fails, try to find a same head antecedent (third test);
- if that doesn't work either, look for an indication that the description is discourse new (following four tests),

33 We considered special predicates and apposition as reliable indications of discourse novelty; in addition, definite descriptions that matched these patterns produced errors in anaphora resolution which were eliminated by processing them first.

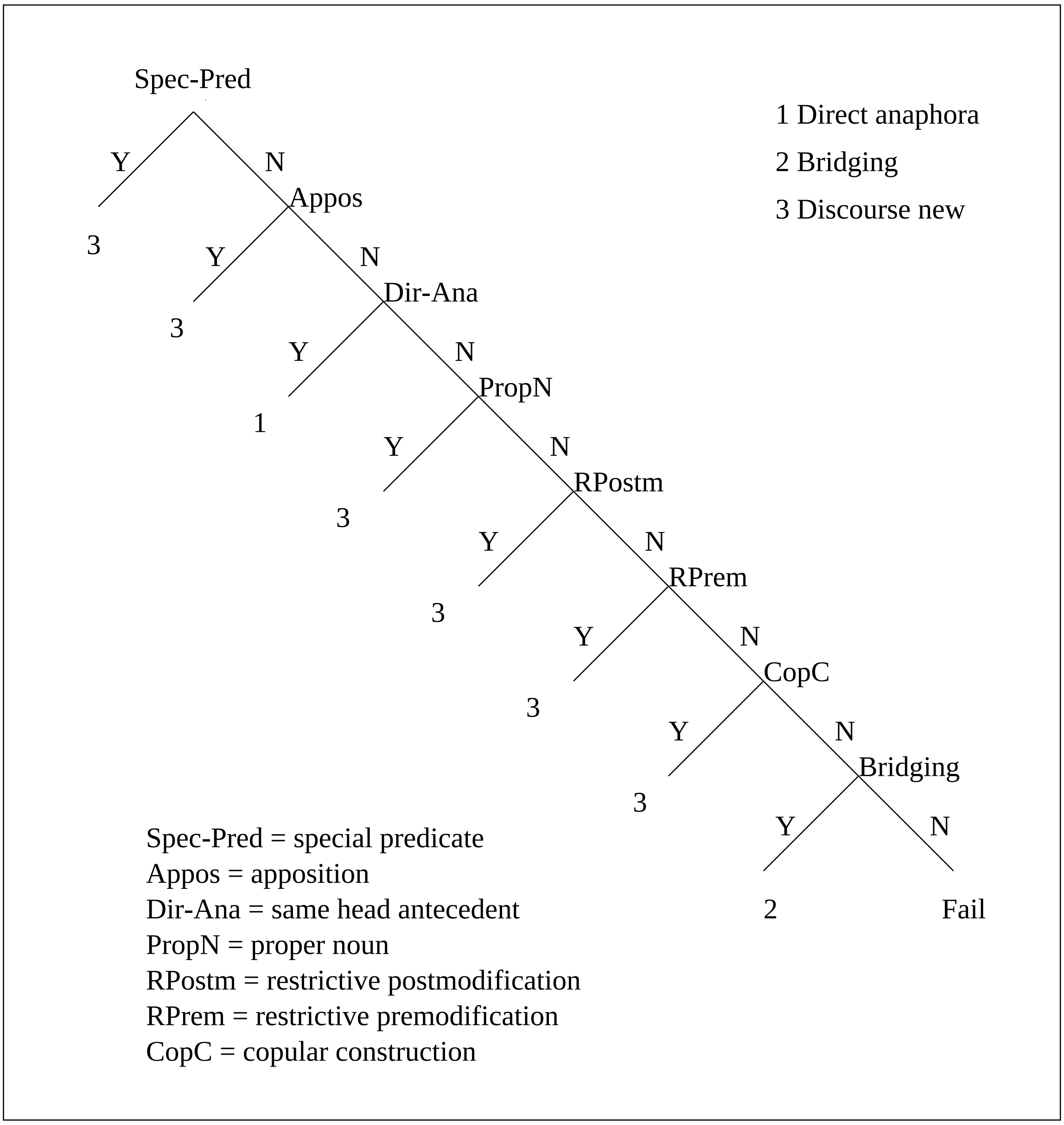

Figure 3
Hand-designed Decision Tree for Version 1 and 2

- only then try to interpret the definite description as a bridge (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. The system does not classify all occurrences of definite descriptions: when none of the tests succeeds, the definite description is not classified. We observed in our first tests that definite descriptions which are not resolved as direct anaphora and are not identified as discourse new by our heuristics were mostly classified in the standardized annotation as bridging descriptions or discourse new. 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.

6.2 Results for Anaphora and Discourse-New Descriptions

6.2.1 Overall results of version 1. We now discuss the overall results of the version of our system dealing with direct anaphora and discourse new descriptions only.

Training data. The output of the optimal configuration of version 1 for the training data is shown in Figure 4. A total of 20 texts were processed, containing 6831 NPs. Almost half of these NPs (2911) were considered as potential antecedents; 1040 descriptions were processed by the system. An antecedent was identified for 270

```
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
```

Figure 4
Global results of version 1 on the training data

of them; for 212 out of the 270 definite descriptions classified as anaphoric same-head by the system, the antecedent was a definite NP. According to the annotation of one of our coders (not the system's output), the 312 anaphoric descriptions were grouped in 164 coreference chains and 86 of these chains were initiated by definite descriptions.

In Figure 5, the results reported by the system are compared with the standard annotation. The figure also shows how descriptions which were not resolved by the system were classified in the standard annotation. Most of the descriptions not classified by the system were bridging descriptions.

The overall precision and recall results of version 1 of the system are shown in Table 18. 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

```
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
```

Figure 5
Summary of the results of version 1 on training data

discourse new descriptions is much higher.

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 18
Global results of version 1 on training data

Test data. Next, the system was evaluated using the test data, corpus 2, which had not been used to develop the heuristics. The results are shown in Figures 6 and 7. The recall and precision figures of the system's performance over the test data are presented in Table 19.

This corpus consisted of 14 texts, containing 2990 NPs. Again, almost half of the NPs were considered as potential antecedents. 464 definite descriptions were processed by the system; of these, the system could classify 324: 115 as direct anaphora, 209 as discourse-new. 88 antecedents were definites themselves. The system incorrectly resolved 77 definite descriptions: 19 anaphoric definites and 58 discourse-new. As before, there were just a few more errors in anaphora resolution than in anaphora classification. The overall recall for the test data was 53% (247/464); precision was 76% (247/324).

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 19
Evaluation of version 2 on the test data

One difference between the results on the two data sets is the distribution into classes of those descriptions that the system fails to classify. In the first corpus, the

```
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
```

Figure 6
Global results of version 2 on 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
```

Figure 7
Summary of the results of version 2 on test data

Bridging Class	Found by System	False Posit.
Names	12	14
C. Nouns	15	10
WN Rel.	34	76
Total	61	100

Table 20
Evaluation of the bridging heuristics all together

largest number of cases not classified are bridging descriptions. By contrast, the largest number of cases not classified by the system in corpus 2 are discourse new.

6.3 Results for bridging descriptions

As discussed in Section §5.4, the results of the heuristics for bridging descriptions presented in Section §4.3 were not very good. We nevertheless included these heuristics into version 2 of the system, which, as discussed above, applied them to those descriptions which failed to be recognized as direct anaphora or discourse new. The heuristics were applied in the following order:

1.proper names,

2.compound nouns,

3.WordNet,

Training data. The manual evaluation of the results of version 2 on the training data is presented in Table 20. The table lists the number of acceptable anchors and the number of false positives found by each heuristic. Note that the system sometimes finds anchors which are not those identified manually, but are nevertheless acceptable.

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 number of right answers.

Test data. Version 2 was tested over the test data using the automatic evaluation–i.e., the system was only evaluated as a classifier, and the anchors found were not analysed. A total of 57 bridging relations were found, but only 19 of the definite descriptions classified as bridges by the system had been classified as bridging descriptions in the standard annotation. Compared to version 1 of the system, which does not resolve bridging descriptions, version 2 has higher recall but lower precision, as shown in Table 21.

System's versions	R	P	F
V.1 Overall	53%	76%	62%
V.2 Overall	57%	70%	62%

Table 21
Comparative evaluation of the system's versions (test data)

6.4 Agreement among system and annotators for version 1 and version 2

As a second form of evaluation of the performance of the system, we measured its agreement with the annotators on the test data using the K statistic.

Version 1 of the system finds a classification for 318 out of 464 definite descriptions in corpus 2 (the test data). If all the definite descriptions that the system cannot classify are treated as discourse-new, the agreement between system and the three subjects that annotated this corpus on the two classes first mention (= discourse-old) and subsequent mention (= discourse-new or bridges) is K=.7; this should be compared with an agreement of K=.77 between the three annotators themselves. If instead of counting these definite descriptions as discourse-new we simply do not include them in our measure of agreement, then the agreement between the system and the annotators is K=.78, as opposed to K=.81 between the annotators. (Notice that the fact that the agreement between annotators goes up, as well, indicates that the definite descriptions that the system can't handle are 'harder' than the rest.)

Version 2 finds a classification for 355 out of 464 definite descriptions; however, its agreement figures are worse. If we count the cases that the system can't classify as discourse-new, the agreement between the system and the three annotators for three classes is K=.57; if we count them as bridges, K=.63; if we just discard those cases, K=.63 again. (By comparison, the agreement among annotators on the three classes was K=.68 overall, K=.70 on just the cases that the system was able to classify.) As said above, the cases that the system can't handle are mainly discourse-new descriptions (see Figure 7).

6.5 Deriving the order of application of the heuristics automatically

6.5.1 Inducing a decision tree. The decision tree discussed in Section §6.1 was derived manually, by trial and error. We also tried to derive the order of application of the heuristics automatically. To do this, we used a modified version of the system to assign boolean feature values to each definite description in the training corpus (i.e., the system checked if the features applied to a definite description instance or not). The following features were used:

1. Special predicates (**Spec-Pred**): this feature has value yes if a special predicate occurs in the definite description (as specified in Section §4.2), and if a complement is there when needed.
2. Direct anaphora (**Dir-Ana**): this feature has value 'yes' if the system can find an antecedent with same head noun for that description (respecting the constraints discussed in Section §4.1).
3. Apposition (**Appos**): 'yes' when the description is in appositive construction.

4.Proper noun (**PropN**): 'yes' when the description has a capitalized initial.

5.Restrictive postmodification (**RPostm**): 'yes' if the definite description is modified by relative or associative clauses.

This list of features, together with the classification assigned to each description in the standard annotation (DDUse), was used to train an implementation of the Quinlan's learning algorithm ID3 (Quinlan, 1993). We excluded the verification of restrictive premodification and copula constructions, since these parameters had given the poorest results before (see Section §6.2). An example of the samples used to train ID3 is shown in example (59).

(59)

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

The algorithm generates a decision tree on the basis of the samples given. The resulting decision tree is presented in Figure 8.

The main difference from the algorithm we arrived at by hand is that the first feature checked by the decision tree generated by ID3 is the presence of an antecedent with same head noun. The presence of special predicates, which we adopted as the first test in our decision tree, is only the fourth test in the tree in Figure 8.

6.5.2 Evaluation of the automatically learned decision tree. The performance of the learned decision tree was compared with that of the algorithm we arrived at by trial and error as follows. The first fourteen texts of corpus 1 (845 descriptions)

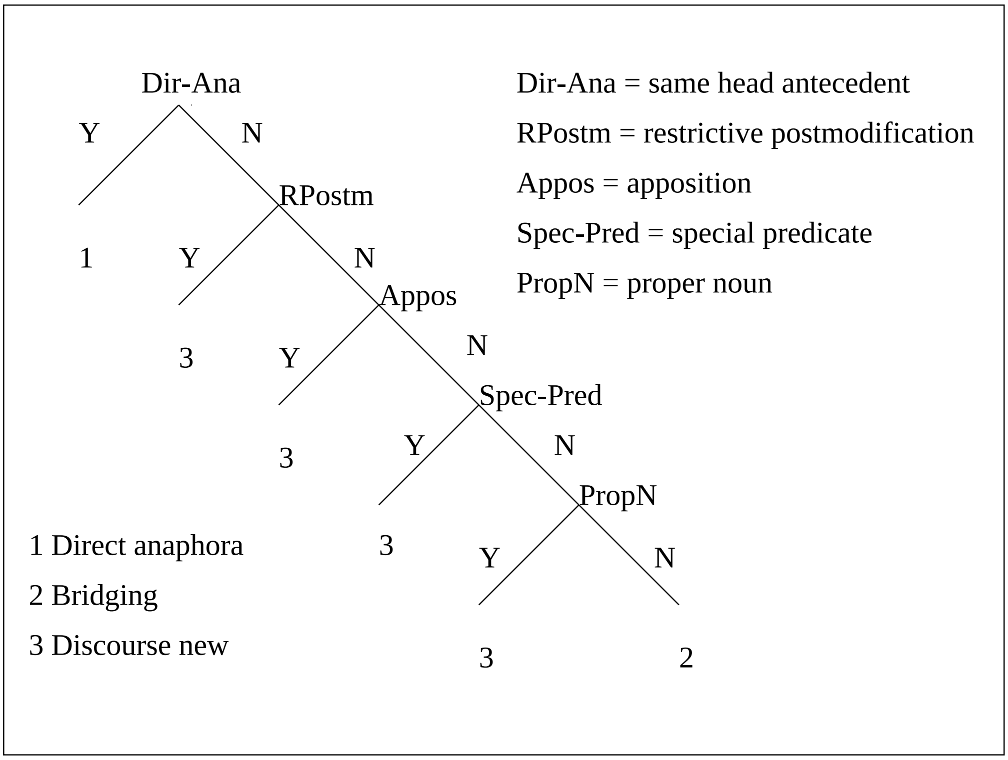

Figure 8
Generated Decision Tree

were used as training data to generate the decision tree. 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 (this test produces the decision tree presented in the previous section);
- 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. The resulting decision tree classifies descriptions with same head antecedent as anaphoric; all the rest as discourse new.

Here we present the results evaluated all together considering the system as a classifier only, i.e., without considering the tasks of anaphora resolution and of identification of discourse new descriptions separately. The output produced by the learned algorithm is compared to the standard annotation. Since the learned algorithm classifies all cases, the number of responses is equal to the number of cases, as a consequence, recall is the same as precision and so is the F measure.

The tests over 6 texts with 195 definite descriptions gave the following results:

- $R = P = F = 69\%$ when the algorithm was trained with three classes;
- $R = P = F = 75\%$, when training with two classes only.

The best results were achieved by the algorithm trained for two classes only. This is not surprising, especially considering how difficult it was for our subjects to distinguish between discourse-new and bridging descriptions.

The hand-crafted decision tree (version 2) achieved 62% recall and 85% precision ($F = 71.70\%$) on those same texts: i.e., a higher precision, but a lower F measure, due to a lower recall, since - unlike the learned algorithm-it does not classify all instances of definite descriptions. If, however, we take the class discourse new as a default for all cases of definite descriptions not resolved by the system, recall, precision and F value go to 77%, slightly higher than that achieved by the decision tree produced by ID3.

As the learned decision tree has the search for a same head antecedent as the first test, we modified our algorithm to work in the same way, and tested it again 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 $R = 0.62$, $P = 0.83$, $F = 0.71$).

In other words, the results were about the same, although a slightly better performance was obtained when the tests to identify discourse new descriptions were tried first.

7. Other computational models of definite description processing

A major difference between our proposal and pretty much all other proposals (theoretical and implemented) is that we concentrate on definite descriptions; most of the systems we discuss below attempt to resolve all types of anaphoric expressions. It's much more common for a system that does anaphoric reference to concentrate on pronouns; but besides the fact that now several systems with this type of concentration already exist, focusing on definite descriptions allowed us to investigate what types of lexical knowledge and commonsense inference are actually used in natural language comprehension.

From an architectural standpoint, the main difference between our work and other proposals in the literature is that we paid considerably more attention to the problem of identifying discourse new definite descriptions.[34]

Previous work on computational methods for definite description resolution can be divided in two camps: proposals which rely on commonsense reasoning (and are therefore either mainly theoretical or domain-dependent), and systems that can be quantitatively evaluated, such as those competing on the coreference task in the Sixth and Seventh Message Understanding Conference (Sundheim, 1995). We discuss these two types of work in turn.

7.1 Models Based On Commonsense Reasoning

The crucial characteristic of these proposals is that they exploit hand-coded commonsense knowledge, and cannot therefore be tested on just any arbitrary text. Some of them are simply tested on texts that were especially built for the purpose

34 This problem is also a central concern in the more recent work by Bean and Riloff (1999).

of testing the system (Carter, 1987; Carbonell and Brown, 1988); systems like the Core Language Engine are more robust, but they have to be applied to a domain restricted enough that all relevant knowledge can be encoded by hand.

Sidner's theory of definite anaphora comprehension.

In her dissertation, Sidner (1979) proposed a complete theory of definite NP resolution, including detailed algorithms for resolving pronouns, anaphoric definite descriptions, and bridging descriptions. She also proposed methods for resolving larger situation uses; the one class her methods do not handle are those definite description that, following Hawkins, we have called 'unfamiliar' uses.

The main contribution of Sidner's dissertation is her theory of focus and its role in resolving definite NPs; to this day, her focus-tracking algorithms are arguably the most detailed account of the phenomenon. The main problem with Sidner's work from our perspective is that her algorithms heavily depend on the availability of a semantic network and causal reasoner; furthermore, some of the inference mechanisms are left relatively underspecified (this latter problem was in part corrected in subsequent work by Carter –see below). Lexical and commonsense knowledge play three important roles in Sidner's system: they are used to track the focus, to resolve bridging descriptions and larger situation uses, and finally to evaluate interpretive hypotheses, discarding those that seem implausible. It is only recently that robust knowledge-based methods for some of these tasks have begun to appear, and their performance is still not very good, as seen above in our discussion of using WordNet as a semantic network;[35] as for checking the plausibility of an

35 An implementation of a (simplified) version of Sidner's focus-tracking algorithms capable of being used by a system like ours was presented in (Azzam, Humphreys, and Gaizauskas, 1998);.

hypothesis on the basis of causal knowledge about the world, we do have now a much better theoretical grasp of how such inferences could be made (see, e.g., (Hobbs et al., 1993; Lascarides and Asher, 1993)), but we are still quite some way off from a general inference engine.

As far as Sidner's resolution rules are concerned, the main problem that we found is that sometimes they are too restrictive. For example, her Co-specification rule 1 prescribes that definite description and focus must have have the same head, and no new information can be introduced by the definite; but this happens fairly frequently in our corpus. Already in (Grosz, Joshi, and Weinstein, 1983) it is recognised that an anaphoric full noun phrase may include some new and unshared information about a previously mentioned entity, and Carter (1987) weakened some of the restrictions proposed by Sidner in his system.

Carter's shallow processing anaphor resolver.

Carter (1987) implemented a modified version of Sidner's algorithm, and integrated it with an implemented version of Wilks' theory of commonsense reasoning. This work is interesting for two reasons: first of all, because Carter, unlike Sidner, attempted to evaluate the performance of his system; and because, in doing this, he addressed the commonsense reasoning problem in some detail.

Carter's system, SPAR, is based on the *Shallow Processing Hypothesis*: that in resolving anaphors, reasoning should be avoided as much as possible. This is, of course, the same approach taken in our own work, which could be seen as pushing Carter's approach to the extreme. The difference is that when it becomes necessary, SPAR does use two commonsense knowledge sources: a semantic network based on Alshawi's theory of memory for text interpretation (Alshawi, 1987) and a causal reasoner based on Wilks' work (Wilks, 1975). In both cases, the necessary information was encoded by hand.

Carter's system was tested over short stories specifically designed for the testing of the system: about 40 written by Carter himself, and 23 written by others. These latter contain about 80 definite descriptions. SPAR correctly resolved all anaphors in the stories written by Carter, and 66 out of 80 of the descriptions in the 23 'test' stories. (Carter himself points out that these results are 'of limited significance because of the simplicity of the texts processed compared to "real" texts' (p. 238).)

The Core Language Engine.

The Core Language Engine (CLE) (Alshawi, 1992) is a domain-independent system developed at SRI Cambridge which translates English sentences into formal representations. The system was used by SRI for a variety of applications, including spoken language translation and airline reservation. The CLE makes use of a core lexicon (to which new entries can be added) and uses an abductive commonsense reasoner to produce an interpretation and to verify the plausibility of choice of referents from an ordered list; the required world knowledge has to be added by hand for each domain, together with whatever lexical knowledge is needed.

The construction of the formal representation goes through an intermediate stage called Quasi Logical Form (QLF). The QLF may contain unresolved terms corresponding to anaphoric NPs including, among others, definite descriptions. The resolution process which transforms QLFs into resolved logical form representation of sentences is described in (Alshawi, 1990). Definite descriptions are represented as quantified terms. The referential readings of definite descriptions are handled by proposing referents from the external application context (larger situation uses) as well as the CLE context model (anaphoric uses). Attributive readings may also be proposed during QLF resolution; some of these seem to correspond to our unfamiliar uses. Thus, the CLE seems to account for discourse new descriptions, although they are not explicitly mentioned, and the methods used for choosing a referential or an attributive interpretation are not discussed. To our knowledge, no analysis of the performance of the system has been published.

7.2 The systems involved in the MUC-6 coreference task

An example of systems that can be quantitatively evaluated are the systems that participated in the MUC-6 competition. There were seven such systems; they achieved recall scores ranging from 35.69% to 62.78% and precision scores ranging from 44.23% to 71.88% on nominal coreference.

It is important to note that the evaluation in MUC-6 differed from ours in three important aspects. First of all, these systems have to parse the texts, which often introduces errors; furthermore, these systems often cannot get complete parses for the sentences they are processing. Secondly, the evaluation in MUC-6 considers the co-referential chain as a whole, and not only one correct antecedent. The third difference is that these systems process a wider range of referring expressions, including pronouns and bare nouns, while our system only processes definite NPs. On the other hand, not all definite descriptions are marked in the MUC-6 coreference task: these systems are only required to identify identity relations; and then again only if the antecedent was introduced by a noun phrase (not if it was a clause or a conjoined NP). This leaves out discourse-new descriptions and, especially, bridging descriptions which, as we have seen, are by far the most difficult cases.

Kameyama (1997) analyses in detail the coreference module of the SRI system that participated in MUC-6 (Appelt, 1995). This system achieved one of the top scores for the coreference task: a recall of 59% and a precision of 72%. The SRI system uses a sort hierarchy claimed to be sparse and incomplete. For definite descriptions, Kameyama reports the results of a test on five articles, containing 61 definite descriptions in total; recall was 46% (28/61), and for proper names 69% (22/32). 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.

7.3 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.

8. Conclusions and future work

8.1 Contributions

We have presented a domain independent system for definite description interpretation whose development was based on an empirical study of definite description use that included multi-annotator experiments. Our system does not only attempt to find an antecedent for a definite description; it also uses methods for recognizing discourse-new descriptions, which our previous studies revealed to be the largest class of definite descriptions in our corpus. Our algorithms for segmentation, matching, and identification of discourse-new descriptions only rely on syntax-based heuristics and on on-line lexical sources such as WordNet; the final configuration of these heuristics, as well as their order of application, was arrived at on the basis of extensive experiments using our 'training' corpus. Because our system only relies on 'shallow' information, it encounters problems when commonsense reasoning is actually needed; on the other hand, it can be tested on any domain without extensive hand-coding.

As far as direct anaphora is concerned, we evaluated heuristic algorithms for segmentation and matching. Our system achieved 62% recall and 83% precision for direct anaphora resolution on our test data. For identifying discourse new descriptions we exploited the correlation between certain types of syntactic constructions and type of use noted by Hawkins (1978) and semantically explained by Löbner (1987). Our system achieved 69% recall and 72% precision for this class on the test data. Overall, the version of the system that only attempts to recognize first mention and subsequent mention definite descriptions achieved a recall of 53% and a precision of 76% on the test corpus if we count the definite descriptions the system can't handle as errors; if we count them as discourse-new, get get P = R = 66

The class of bridging descriptions is the most difficult to process: this is in part because humans themselves do not agree much on which definites count as bridges and what their anchors are, in part because lexical knowledge and commonsense reasoning are necessary to solve them. Our results for this class are, therefore, still very tentative; this did not affect much the performance of the system, however, since in the texts we tried bridging descriptions are a relatively small class. Non co-referent bridging descriptions were around 8% of the total definite descriptions in the corpus, and the class of bridging descriptions including those with a co-referent antecedent with different head noun were about 15% of the total. We tried techniques which do not involve heavy axiomatisation of commonsense knowledge, and only used an existing lexical source, WordNet.

In other text genres the distribution of definite descriptions into classes might change in detail: e.g., in spoken dialogue the number of deictic definite descriptions is going to increase. However, other researchers (Fraurud, 1990) found a similar distribution of first-mention and subsequent mention definites in text corpora; we believe therefore that the heuristics we propose here and their ordering 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.

8.2 What's needed next

We would like to emphasize again that we are not trying to suggest that shallow methods will be sufficient for processing definite descriptions in the long run. What we do believe is that hypotheses about processing should be evaluated; unfortunately only fairly simple techniques can be tested in this way at the moment, but this work can serve to motivate more clearly the use of more complex methods.

We highlighted throughout the paper, and particularly in Section §5, some of the points where shallow methods break down, and better lexical sources or commonsense knowledge are needed. By far the worse results are obtained for bridging descriptions; in this area, the most urgent needs are better sources of lexical knowledge,[36] and some robust focusing mechanism. Of the problems faced in resolving anaphoric descriptions, that of finding better ways of segmenting the text is perhaps the one in which the most progress has been made since we started this project; robust methods for text segmentation are now available (Hearst, 1997; Richmond, Smith, and Amitay, 1997). A proper treatment of modification seems harder; as discussed in Section §4.1, considerable amounts of reasoning may be needed in some cases. In order to improve our treatment of discourse-new descriptions it will be necessary, on the one hand, to find way of automatically acquiring lexical information about the functionality of nouns and adjectives; on the other hand, to have sources of encyclopedic knowledge available.

8.3 Future work

8.3.1 Simple Extensions. In this project we were more interested in clearly identifying the subtasks of the definite description process that in achieving optimum

36 As said above, we have done some preliminary work on acquiring this information automatically (Poesio, Schulte im Walde, and Brew, 1998; Ishikawa, 1998).

performance; as a consequence, there are a number of fairly simple ways in which the final version of the system could be improved. The next step to make our system truly testable on any type of text would be to make it work off the output of a robust parser: we are currently testing for this purpose Abney's CASS parser (Abney, 1991). (See (Ishikawa, 1998).) We are also experimenting with existing software that performs in a more sophisticated way some of the tasks that our system currently implements in a fairly crude fashion, including lemmatization, proper name recognition, and named entity typing.

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 and the annotation would have to be extended to cover all cases of anaphoric expressions.

8.3.2 The role of focus in definite descriptions processing. Our tests with bridging descriptions resulted in a great number of false positives. Our analysis of these data, as well as of other corpora (Hitzeman and Poesio, 1998), suggests that a local focusing mechanism as proposed in (Grosz, 1977; Sidner, 1979; Grosz, Joshi, and Weinstein, 1983; Grosz and Sidner, 1986; Grosz, Joshi, and Weinstein, 1995) would improve the results obtained by our system.

There are several reasons why our system doesn't include such a mechanism yet. One problem already mentioned is that Sidner's algorithms as stated, and even as implemented by Carter, are difficult to implement, since considerably more lexical information is needed than we have available (e.g., about the thematic roles of verbs), a rich knowledge base is needed both to resolve bridging descriptions and larger situation uses, and commonsense inference is needed to evaluate the plausibility of hypotheses. A second problem with Sidner's theory of local focus, as well as others such as Centering Theory (Grosz, Joshi, and Weinstein, 1995), is the lack of a precise characterization of how to deal with complex sentences. Revisions and extensions of Sidner's proposal related to these problems have been proposed in (Suri and McCoy, 1994), that include algorithms for focus updating in complex sentences containing adjunct clauses such as *before*- and *after*-clauses.

We plan to incorporate simpler focus-tracking mechanisms in future versions of the system, possibly along the lines of (Azzam, Humphreys, and Gaizauskas, 1998) or (Tetreault, 1999).

8.3.3 Theoretical Developments. We defended the importance of developing methods for identifying discourse new descriptions, and we believe that there is still need for research into the semantics of this class: i.e., what exactly licenses the use of a definite description to refer to a discourse-new entity. 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 us a better understanding of the problem. The postmodification of a description often acts as an explicit anchor (what Löbner (1987) calls 'disambiguating arguments and attributes'); finding out how the head noun of a postmodified description relates 'semantically' with its complement is a problem similar to that of identifying the semantic relation between a bridging description and its anaphoric anchor, but to date there hasn't been much research on this topic (while there has been a lot of work on finding out the relations that hold between the premodifiers, especially in noun-noun compounds). An NP's head noun may also corefer with its complement, as seen in the examples in (60):

(60) 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.

Acknowledgments

We wish to thank Ellen Bard, Rafael Bordini, Jean Carletta, Miriam Eckert, Kari Fraurud, Rob Gaizauskas, Janet Hitzeman, Chris Mellish, and our anonymous reviewers for comments, help, and suggestions. Renata Vieira was supported in part by a fellowship from CNPq, Brazil; Massimo Poesio is supported by an EPSRC Advanced Research Fellowship.

References

Abney, Steve. 1991. Parsing by chunks. In R. Berwick, S. Abney, and C.Tenny, editors, *Principle-based Parsing*. Kluwer, Dordrecht.

Alshawi, Hiyan. 1987. *Memory and Context for Language Interpretation*. Cambridge University Press, Cambridge.

Alshawi, Hiyan. 1990. Resolving quasi-logical forms. *Computational Linguistics*, 16(3):133–144.

Alshawi, Hiyan, editor. 1992. *The Core Language Engine*. The MIT Press.

Aone, Chinatsu and Scott W. Bennett. 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, Douglas. 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.

Azzam, Saliha, Kevin Humphreys, and Robert Gaizauskas. 1998. Evaluating a focus-based approach to anaphora resolution. In *Proc. of the 36th ACL and 17th COLING*, pages 74–78, Montreal, CA. ACL.

Bean, David L. and Ellen Riloff. 1999. Corpus-based identification of non-anaphoric noun phrases. In *Proc. of the 37th ACL*, pages 373–380, University of Maryland. ACL.

Bikel, Daniel, Scott Miller, Richard Schwartz, and Ralph Weischedel. 1997. Nymble: a high-performance learning name finder. In *Proc. of the 5th Conference on Applied Natural Language Processing*, pages 194–201, Washington, D.C. ACL.

Bosch, Peter and Bart Geurts. 1989. Processing definite NPs. IWBS Report 78, IBM Germany, July.

Burger, John D. and Dennis Connolly. 1992. Probabilistic resolution of anaphoric reference. In *Proc. AAAI Fall Symposium on Probabilistic Approaches to Natural Language*, pages 17–24, Cambridge, Massachusetts.

Carbonell, Jamie and Ralf D. Brown. 1988. Anaphora resolution: A multi-strategy approach. In *Proceedings of the 12th International Conference on Computational Linguistics (COLING-88)*, pages 96–101, Budapest, Hungary.

Carletta, Jean. 1996. Assessing agreement on classification tasks: the kappa statistic. *Computational Linguistics*, 22(2):249–254.

Carletta, Jean et al. 1997. The Reliability of a Dialogue Structure Coding Scheme. *Computational Linguistics*, 23(1):13–32.

Carter, David M. 1987. *Interpreting Anaphors in Natural Language Texts*. Ellis Horwood, Chichester, UK.

Chinchor, Nancy 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.

Chinchor, Nancy A. 1997. Overview of MUC-7/MET-2. In *Proc. of the Seventh Message Understanding Conference (MUC-7).* Available at `http://www.muc.saic.com/proceedings/muc_7_proceedings/overview.html`.

Cho, Sehyeong and Anthony 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.

Clark, Herbert H. 1977. Inferences in comprehension. In D. Laberge and S. J. Samuels, editors, *Basic Process in Reading: Perception and Comprehension*. Lawrence Erlbaum.

Clark, Herbert H. and Catherine R. Marshall. 1981. Definite reference and mutual knowledge. In A. Joshi, B. Webber, and I. Sag, editors, *Elements of Discourse Understanding*. Cambridge University Press, New York.

Fellbaum, Christiane, editor. 1998. *WordNet: An electronic lexical database.* The MIT Press.

Fox, Barbara A. 1987. *Discourse Structure and Anaphora.* Cambridge University Press, Cambridge, UK.

Fraurud, Keri. 1990. Definiteness and the processing of NPs in natural discourse. *Journal of Semantics*, 7:395–433.

Gaizauskas, Robert, Takahiro Wakao, Kevin Humphreys, Hamish Cunningham, and Yorick Wilks. 1995. University of Sheffield: description of the LaSIE System as used for MUC-6. In *Proceedings of the Sixth Message Understanding Conference (MUC-6)*, pages 207–220. Morgan Kauffmann.

Grosz, Barbara J. 1977. *The Representation and Use of Focus in Dialogue Understanding.* Ph.D. thesis, Stanford University.

Grosz, Barbara. J., Aravind. K. Joshi, and Scott Weinstein. 1995. Centering: A framework for modeling the local coherence of discourse. *Computational Linguistics*, 21(2):202–225. (The paper originally appeared as an unpublished manuscript in 1986.).

Grosz, Barbara J. and Candace L. Sidner. 1986. Attention, intention, and the structure of discourse. *Computational Linguistics*, 12(3):175–204.

Grosz, Barbara.J., Aravind K. Joshi, and Scott Weinstein. 1983. Providing a unified account of definite noun phrases in discourse. In *Proc. ACL-83*, pages 44–50.

Hahn, Udo; Strube, Michael; and Markert, Katja. 1996. Bridging textual ellipsis In: COLING '96: *Proc. of 16th International Conference on Computational Linguistics.* Kopenhagen, Aug 5-9 1996, pp.496-501.

Hardt, Daniel. 1997. An empirical approach to VP ellipsis. *Computational Linguistics*, 23(4).

Hatzivassiloglou, Vasileios and Kathleen McKeown. 1993. Towards the automatic identification of adjectival scales: clustering adjectives according to meaning. In *Proc. of the 31st Annual Meeting of the Association for Computational Linguistics*, pages 172–182, Ohio State University.

Hawkins, John A. 1978. *Definiteness and Indefiniteness*. Croom Helm, London.

Hearst, Marti A. 1997. TextTiling: Segmenting text into multi-paragraph subtopic passages. *Computational Linguistics*, 23(1):33–64.

Heim, Irene. 1982. *The Semantics of Definite and Indefinite Noun Phrases*. Ph.D. thesis, University of Massachusetts at Amherst.

Hitzeman, Janet and Massimo Poesio. 1998. Long-distance pronominalisation and global focus. In *Proc. of ACL/COLING, vol. 1*, pages 550–556, Montreal.

Hobbs, Jerry R., Mark E. Stickel, Douglas A. Appelt, and Paul Martin. 1993. Interpretation as abduction. *Artificial Intelligence Journal*, 63:69–142.

Humphreys, Kevin, Robert Gaizauskas, Saliha Azzam, Chris Huyck, B. Mitchell, and Hamish Cunningham. 1998. University of Sheffield: description of the LaSIE-II System as used for MUC-7. In *Proceedings of the Seventh Message Understanding Conference (MUC-6)*.

Ishikawa, Tomonori. 1998. Acquisition of associative information and resolution of bridging descriptions. Master's thesis, University of Edinburgh, Department of Linguistics, Edinburgh, Scotland.

Kameyama, Megumi. 1997. Recognizing referential links: an information extraction perspective. In *Proc. of the ACL Workshop on Operational Factors in Practical, Robust Anaphora Resolution for Unrestricted Texts*, pages 46–53, Madrid, Spain, July. ACL.

Krippendorff, Klaus. 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.

Lappin, Shalom and H. J. Leass. 1994. An algorithm for pronominal anaphora resolution. *Computational Linguistics*, 20(4):535–562.

Lascarides, Alex and Nicholas Asher. 1993. Temporal interpretation, discourse relations and commonsense entailment. *Linguistics and Philosophy*, 16(5):437–493.

Loebner, Sebastian. 1987. Natural language and generalised quantifier theory. In P. Gärdenfors, editor, *Generalized Quantifiers*. D. Reidel, Dordrecht, The Netherlands, pages 93–108.

Mani, Inderjeet and T. Richard MacMillan. 1996. Identifying unknown proper names in newswire text. In B. Boguraev and J. Pustejovsky, editors, *Corpus processing for lexical acquisition*. MIT Press, pages 41–59.

Marcu, Daniel. 1999. A decision-based approach to rhetorical parsing. In *Proc. of the 37th ACL*, pages 365–372, University of Maryland, June. ACL.

Marcus, Mitchell P., Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of english: the Penn Treebank. *Computational Linguistics*, 19(2):313–330.

McDonald, David. 1996. Internal and external evidence in the identification and semantic categorization of proper names. In Bran Boguraev and James Pustejovsky, editors, *Corpus processing for lexical acquisition*. The MIT Press, Cambridge, MA, pages 21–39.

Mikheev, Andrei, Marc Moens, and Claire Grover. 1999. Named Entity recognition without gazetteers. In *Proc. of EACL*, Bergen, Norway. EACL.

Mitkov, Ruslan. 2000. Towards more comprehensive evaluation in anaphora resolution. In *Proc. of the 2nd International Conference on Language Resources and Evaluation* , pages 1309–1314, Athens.

Paik, Woojin, Elizabeth D. Liddy, Edmund Yu, and Mary McKenna. 1996. Categorizing and standardizing proper nouns for efficient information retrieval. In Bran Boguraev and James Pustejovsky, editors, *Corpus processing for lexical acquisition*. The MIT Press, Cambridge, MA, pages 61–73.

Palmer, David D. and David S. Day. 1997. A statistical profile of the named entity task. In *Proc. of the 5th Conference on Applied Natural Language Processing*, pages 190–193, Washington, D.C., March. ACL.

Poesio, Massimo. 1993. A situation-theoretic formalization of definite description interpretation in plan elaboration dialogues. In Peter Aczel, David Israel, Yasuhiro Katagiri, and Stanley Peters, editors, *Situation Theory and its Applications, vol.3*. CSLI, Stanford, chapter 12, pages 339–374.

Poesio, Massimo, Sabine Schulte im Walde, and Chris Brew. 1998. Lexical clustering and definite description interpretation. In *Proc. of the AAAI Spring Symposium on Learning for Discourse*, pages 82–89, Stanford, CA, March. AAAI.

Poesio, Massimo and Renata Vieira. 1998. A corpus-based investigation of definite description use. *Computational Linguistics*, 24(2):183–216, June.

Poesio, Massimo, Renata Vieira, and Simone Teufel. 1997. Resolving bridging references in unrestricted text. In R. Mitkov, editor, *Proc. of the ACL Workshop on Operational Factors in Robust Anaphora Resolution*, pages 1–6, Madrid. Also available as HCRC Research Paper HCRC/RP-87, University of Edinburgh.

Postal, Paul M. 1969. Anaphoric islands. In R. I. Binnick et al., editor, *Papers from the Fifth Regional Meeting of the Chicago Linguistic Society*, pages 205–235. University of Chicago.

Prince, Ellen F. 1981. Toward a taxonomy of given-new information. In Peter Cole, editor, *Radical Pragmatics*. Academic Press, New York, pages 223–256.

Prince, Ellen 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. Ross. 1993. *C4.5: programs for machine learning*. Morgan Kaufmann, San Mateo, CA.

Quirk, Randolph, Sydney Greenbaum, Geoffrey Leech, and Jan Svartvik. 1985. *A Comprehensive Grammar of the English Language*. Longman, London.

Reichman, Rachel. 1985. *Getting Computers to Talk Like You and Me*. The MIT Press, Cambridge, MA.

Richmond, Kevin, Andrew Smith, and Einat Amitay. 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.

Russell, Bertrand. 1905. On denoting. *Mind*, 14:479–493. Reprinted in *Logic and Knowledge*, ed. R. C. Marsh. London: George Allen and Unwin.

Sidner, Candace L. 1979. *Towards a computational theory of definite anaphora comprehension in English discourse*. Ph.D. thesis, MIT.

Siegel, Sydney. and N. John Castellan. 1988. *Nonparametric statistics for the Behavioral Sciences*. McGraw-Hill, 2nd edition.

Strand, Kjetil. 1996. A taxonomy of linking relations. Manuscript. A preliminary version presented at the Workshop on Indirect Anaphora, Lancaster University, 1996.

Sundheim, Beth 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, Linda Z. and Kathleen F. McCoy. 1994. RAFT/RAPR and centering: A comparison and discussion of problems related to processing complex sentences. *Computational Linguistics*, 20(2):301–317.

Tetreault, Joel R. 1999. Analysis of syntax-based pronoun resolution methods. In *Proc. of the 37th ACL*, pages 602–605, University of Marylan, June. ACL.

Vieira, Renata. 1998. *Definite Description Resolution in Unrestricted Texts*. Ph.D. thesis, University of Edinburgh, Centre for Cognitive Science, February.

Vieira, Renata and Massimo Poesio. 1996. Processing definite descriptions in corpora. In Simon Botley, J. Glass, Tony McEnery, and Andrew Wilson, editors, *Proc. of the Discourse Anaphora and Resolution Colloquium (DAARC)*, Lancaster, UK. Lancaster University, UCREL. Also available as Research Paper HCRC/RP-86, University of Edinburgh, Human Communication Research Centre.

Vieira, Renata and Simone Teufel. 1997. Towards resolution of bridging descriptions. In *Proc. of the 35th Joint Meeting of the Association for Computational Linguistics*, Madrid.

Vilain, Marc et al. 1995. A model-theoretic coreference scoring scheme. In *Proc. of the Sixth Message Understanding Conference*, pages 45–52.

Wacholder, Nina and Yael. Ravin. 1997. Disambiguation of proper names in text. In *Proc. of the 5th Conference on Applied Natural Language Processing*, pages 202–208, Washington, D.C., March. ACL.

Ward, Gregory, Richard Sproat, and Gail McKoon. 1991. A pragmatic analysis of so-called anaphoric islands. *Language*, 67:439–474.

Webber, Bonnie L. 1979. *A Formal Approach to Discourse Anaphora*. Garland, New York.

Wilks, Yorick A. 1975. A preferential pattern-matching semantics for natural language. *Artificial Intelligence*, 6:53–74.