Processing definite descriptions in corpora

Renata Vieira and Massimo Poesio

Abstract

We discuss in this paper a system that resolves definite descriptions in written texts. A preliminary study of definite descriptions in a collection of 20 texts revealed that about 30% of the 1040 definites in the collection were cases of anaphoric definites whose antecedents had the same head noun, and 50% introduced novel discourse referents. An algorithm which resolves anaphoric definite descriptions and also identifies novel ones is proposed. We evaluated the algorithm by comparing its results with an annotation produced by human subjects. The analysis of the corpus, the implemented algorithm, and the evaluation of the results are presented in this paper.

1 Introduction

We are interested in definite description resolution techniques and their quantitative evaluation. We discuss in this paper a system that resolves definite descriptions (DDs) (noun phrases headed by the definite article, such as *the car*) in written texts. Our algorithms use linguistic information but do not make use of inference-based mechanisms that would require hand-coded knowledge. The corpus used in this work is a set of 20 parsed articles of the Wall Street Journal selected at random from the Penn Treebank Corpus (Marcus et al, 1993). A preliminary study of definite descriptions in the corpus revealed that about 30% of the cases were anaphoric definites whose antecedents had the same head noun and 50% introduced novel discourse referents. In Section 2 of this paper, we present a taxonomy of uses of DDs and our analysis of the corpus using this taxonomy. We describe the strategies adopted for the implementation of the system and present our algorithm in Section 3. In Section 4, we illustrate the results of the system on one text of the corpus; the text itself is given in the appendix. We evaluated our algorithm by comparing its classification with the classification produced by human subjects, and obtained a recall of 72% and a precision of 82% for anaphoric resolution of DDs whose antecedents had the same head noun; a recall of 74% and a precision of 85% for novel ones. The results of the system for the whole corpus are presented in Section 5, together with their evaluation.

2 Definite Descriptions in Corpora

The use of definite descriptions most commonly discussed in the linguistic literature is the case in which a definite description picks up a referent introduced in a previous discourse, called the ‘anaphoric’ use. Less attention has been devoted to definites whose interpretation depends on knowledge about the existence of certain referents, such as *the nation,* or *the government* for speakers of the same country. Even more neglected are DDs which introduce a novel referent in the discourse but whose interpretation is not dependent on previous knowledge about the referent - for instance, *the long-cherished dream of home ownership* (sentence 6 in the appendix). Clearly, the rate of success of a system intended to resolve DDs depends on the relative frequency of different kinds of definite descriptions, and on the success of the heuristics used to identify those definites which do not need to be resolved with a previously introduced discourse referent. A detailed classification of the uses of definite descriptions is therefore crucial both in the development of such a system and in the assessment of the likelihood of its success.

2.1 Uses of Definite Descriptions

We based our study of DDs on the classification proposed by (Hawkins, 1978). Hawkins' taxonomy was modified for the task of classifying the occurrences of definite descriptions in written natural language texts, which, for example, do not include 'immediate' or 'visible situation' uses. A brief description and exemplification of the four classes which we considered relevant for our purposes follows. Some of the examples of DDs given in the paper are extracted from the text in the appendix; when this happens, we indicate the number of the sentence in which the DD appears.

Anaphoric same head: for this kind of DD an antecedent is given explicitly in the text and by means of a same head noun, as in *a government panel ... the panel* (s. 11)[1] and *a report on the extent and causes of the problem ... the report* (s. 9,10,14).

Associative: these DDs are based on an associated antecedent (trigger) which is explicitly given in the text. The description may refer to the same entity as the antecedent or to an associated one. The antecedent may be a noun phrase (NP) as well as an event represented by a verb phrase, a sentence or even a larger sequence of text. The identification of the pair trigger-associate requires some form of reasoning. Examples are *Y. J. Park and her family ... the 33-year-old housewife* (s. 3) and *to buy a tiny apartment ... the price* (s.1).

Larger Situation/Unfamiliar: no antecedent is provided by the text.
Larger Situation: the description refers to an entity or event whose existence is of common knowledge. Examples are *the National Assembly* (first occurrence) (s. 19); *the Iran-Iraq war; the past year* (s. 20)*; the nation.*
Unfamiliar: the interpretation of the description is based on additional information attached to the definite NP, such as relative clauses, *the average realized for other similar-sized property in an area* (s. 27); associative clauses, *the popular stand of President Roh* (s. 29); NP complements, *the fact that ...*; unexplanatory modifiers, *the first ..., (the best, the highest, the tallest)*; appositive clauses, *the Citizens Coalition for Economic Justice, a public-interest group leading the charge for radical reform* (s. 32); or copula constructions, *the chief culprits are big companies and business groups that buy huge amounts of land* (s. 40).

Idiom: idiomatic expressions as in *It went back into the soup.*

2.2 Data Analysis

The results of our preliminary analysis of the 20 texts are summarised in Table 1.

---

[1] This indicates the number of the sentence of the text in the appendix where the example is extracted from. In this case it indicates sentence 11 .

Class	Total	Percentage
Anaphoric s. h.	305	29%
Associative	192	19%
Larger sit./Unfamiliar	503	48%
Idiom	26	3%
Doubt	14	1%
Total	1040	100%

Table 1: Data Analysis

These results show the distribution of definite descriptions in the kind of text we studied, and are rather encouraging: if simple forms of lexical information and heuristics based on syntactic information can be used to resolve anaphoric same head DDs and to recognise larger situation/unfamiliar uses (DDs that introduce new referents), then about 77% of the total number of definites can be treated. Adding a treatment of associative relations to the system would bring us to deal with 96% of definite descriptions. Our first task was therefore to develop treatments for the two largest classes of definite descriptions. This work is described in the next section.

3 A heuristic-based system for resolving definite descriptions

Our system is implemented in Prolog and was designed to run over texts annotated with their parse tree, according to the format adopted for the Penn Treebank of the University of Pennsylvania (first version) (Marcus et al, 1993). The main tasks of the current version of the system are to collect potential antecedents, to resolve anaphoric definite descriptions with those potential antecedents, and to identify larger situation and unfamiliar uses of DDs. Some NPs are treated as novel in the sense of Heim (Heim, 1982); they are assumed to introduce a new discourse referent to the discourse model, which is stored with information about its properties (head and premodifiers). When the system encounters a definite description, it tries to determine whether it is anaphoric - in which case it is linked to a previous introduced referent - or larger situation/unfamiliar, in which case it is treated as novel.

The system is based on heuristics which were motivated by Hawkins' discussion and by our empirical study of the corpus. The proposed heuristics as well as their interaction were tested on the basis of a manual annotation of the corpus.

Anaphoric Descriptions

The key problems to be dealt with in order to resolve anaphoric definite descriptions are to identify the potential antecedents, and to match the definite descriptions with the most likely antecedents. In our implementation, indefinite NPs (those headed by the indefinite articles *a, an,* and *some*), possessives[2], bare plural and plural NPs with cardinal determiners (as in *three cars*) always introduce a new referent to the discourse model. Definite descriptions themselves may be novel in the discourse, therefore they, as well, may result in a new referent being added to the discourse model. For each new referent the system stores a set of prolog assertions, which may be seen as encoding Heim's 'file cards'. Each 'file card' contains information about the NP index, the whole

---

[2] A possessive NP such as *the government's Office of Bank Supervision and Examination* (s. 42) will be considered as a potential antecedent whose head noun is *Office*. The system, however, will try to resolve only the definite NP *the government*, as our prototype does not provide an account for the interpretation of possessive descriptions.

NP structure, the NP head noun, the NP type (definite, indefinite, bare plural/possessive), the list of premodifiers and the number of the sentence which it belongs to, as shown below:

```
potential_antecedent(NP_index,np(Syntax),head(NP_head),type(NP_type))
premodifiers(NP_index,NP_premodifiers_list)
np_sentence_table(NP_index,Sentence_index)
```

The only form of anaphoric resolution currently carried out by the system involves definite descriptions which have the same head noun as a previously introduced discourse referent. One problem that we considered is the information provided by the modifiers (adjectives and/or complements) of the noun phrase. Simply matching heads would incorrectly suggest antecedent/DD pairs such as *the business community - the younger, more activist black political community*; or, *the population - the voting population*. Various heuristics for treating premodifiers were considered (Section 5.1). The best results were obtained by the following heuristic matching algorithm:

> 1) Allow a pre-modified antecedent to match with a definite whose set of pre-modifiers is a subset of the set of modifiers of the antecedent. This first heuristic deals with definites which contain less information than the antecedent, examples are *an old Victorian house - the house*; *a retired couple in Oakland - the couple*; and *the San Francisco earthquake - the earthquake* . This prevents matches such as *the business community - the younger, more activist black political community*.
>
> 2) Allow a non-premodified antecedent to match with any same head definite. This second part of the algorithm deals with definites that provide additional information. Examples from our corpus of pairs that match thanks to this heuristic are *a check - the lost check*, or *the campaign - the Dinkins campaign*..

Finally, a very simple heuristic for discourse segmentation was taken into account. The resolution process considers only antecedents that appear at a distance not greater than 5 sentences. This constraint is relaxed when the potential antecedent has been already used as an antecedent in another resolution, and when both the definite description and antecedent have identical forms.

Larger situation and unfamiliar descriptions

Whenever the system fails in finding an antecedent for a DD, another set of algorithms is used in order to identify definite descriptions that introduce new referents to the discourse. As discussed above, we concentrated on two such classes: 'unfamiliar' and 'larger situation' definites. The classification performed by the system depends on syntactic and lexical features of the noun phrase.

<u>Proper nouns</u> - If the system fails to find an antecedent for the definite being processed, it checks whether the head is a proper noun by looking if it is capitalised. Examples include *the 1988 Seoul Olympics* (s.11), and *the Federation of Korean Industries* (s.34). If that is the case, the definite phrase is classified as a larger situation use.[3] We

---

[3]Note that because this test is performed after trying to find an anaphoric antecedent, anaphoric resolution takes priority, and the subsequently occurrences of the same proper noun is classified as an anaphoric use.

also considered as larger situation the cases of DDs with proper nouns in premodifier position, as in *the Iran-Iraq war*.

Special nouns - Some cases of definite descriptions are identified by comparing the head noun of the DD with a list of nouns that indicate larger situation or unfamiliar uses. The list of nouns related to larger situation uses includes terms indicating time reference such as *year, day, week, month, hour, time, morning, afternoon, night, period, quarter* and their respective plurals. We use another list of nouns that may take NP complements and indicate unfamiliar uses; this list currently includes *fact, result,* and *conclusion* . We also have a list of what Hawkins calls ‘unexplanatory modifiers’: *first, last, best, most, maximum, minimum, only, closest, greatest, biggest,* and superlatives in general (these modifiers are compared with the premodification of the DD and not with its head noun). Finally, a list of relatives such as *more, closer, greater,* and *bigger* is compared with the DD’s head noun, since comparatives like *the closer they got to saving ... the more the price rose* (s.1) do not require the identification of an antecedent for their interpretation either.

Appositions and Copular Constructions - Definite descriptions occurring in appositive construction, such as *The Citizens Coalition for Economic Justice, a public-interest group leading the charges for radical reform,* often do not need to be resolved. Appositive constructions are treated in the Treebank as NP modification. The system recognises an apposition by checking whether the definite is inserted in a complex noun phrase with structure

[NP,[NP,...],[NP,...]]

consisting of a sequence of noun phrases in which one is a name, as in the example above:

```
[NP,[NP,The,Citizens,Coalition,[PP,for,[NP,Economic,Justice]]],,,
        [NP,[NP,a,public-interest,group]   ...]]
```

If definites occur in certain copular constructions they may not have an antecedent, as in *The chief culprits are big companies...* (s.40) or *The result is that those rich enough to own any real estate at all...* (s. 12).

Restrictive postmodification - The last feature of the structure of the definite noun phrase verified by the system is the presence of restrictive postmodification. Hawkins mentioned referent establishing relative clauses and associative clauses as two constructions that licensed an unfamiliar definite, but also warned that not all relative clauses are referent establishing. It turned out that a large number of definite descriptions with restrictive post-modifiers are unfamiliar in the corpus. Examples include *the full scope of the penalties* (s.28) and *the popular standing of President Roh* (s.29).

To summarise, the algorithm currently implemented by the system is as follows.[4]

For each NP of the input:

---

[4] Note that some of the strategies adopted in our prototype still require further evaluation. The text will make clear which strategies have been evaluated and which ones are just preliminary suggestions.

If the NP is an indefinite, a regular plural, or a possessive, the system creates a new file card for it.

If the NP is headed by the definite article, the system applies to it the following sequence of 6 tests. The first test that succeeds for the DD determines its classification. Whenever a test succeeds the resolution process stops and the next NP is processed.

1. Examine the lists of special nouns in order to identify some of the unfamiliar and larger situation uses of definite descriptions.[5]

2. Check whether the definite NP occurs in an appositive construction. If this test succeeds, a new discourse referent is introduced, and the DD is classified as unfamiliar.

3. Try to find an antecedent for the definite description using a matching algorithm modified to deal with pre-modification and respecting segmentation. When this test succeeds the DD is classified as anaphoric.

4. Verify if the head of the NP is a proper noun (by checking whether it is capitalised). If so, the DD is considered a case of larger situation use and it introduces a new discourse referent.

5. Check if the definite is postmodified. Definites which are not anaphoric and have restrictive post modifiers are marked as unfamiliar and are added to the discourse model as new referents.

6. Finally, the system verifies if there is a proper noun in premodifier position. These cases are also treated as cases of restrictive pre-modification signalling a larger situation use and a new file card is created for them.

If the last test fails, the definite will be included as a new referent in the discourse model in order to be available to further resolution, and the next NP is processed.

The strategy adopted by the system is first to eliminate cases which are potentially non anaphoric (first two tests[6]), then try to find an antecedent (third test) and when an antecedent is not found (last three tests) look for an indication that the DD is new in the discourse. The system is not able to classify all occurrences of definite descriptions, but the implemented heuristics produced results for a considerable number of cases (701 in 1040). We discuss the overall results in Section 5, before that we present an example.

4 An Example

The results displayed by the system after execution on the text in the Appendix are shown in Figure 1. The system counts and displays the number of sentences, the number of potential antecedents, and the number of definite descriptions. The system also counts its own classification of DDs, as seen in the Figure.

---

[5]These are the only cases of non-resolved descriptions which are not considered as potential antecedents by the system, as further occurrences of DDs with these same heads will not be resolved, they will also match with the lists of special nouns.

[6] Considering that there is a large number of descriptions which do not require a textual antecedent for their interpretation, and a large number of potential antecedents with a comparatively small number of actual antecedents, it is interesting that the resolution algorithm avoids unnecessary search. The first two tests proposed are, however, preliminary suggestions.

NUMBER OF SENTENCES: 48
NUMBER OF NOUN PHRASES: 368
NUMBER OF POTENTIAL ANTECEDENTS: 121

- Indefinites: 25
- Plurals and Possessives: 53
- Definites: 43

NUMBER OF DEFINITE DESCRIPTIONS: 78

- ANAPHORA RESOLUTIONS (SAME HEAD): 27
  - Multiply resolved: 0
  - Actual antecedents: 10
    - Indefinites: 2
    - Plurals and Possessives: 2
    - Definites: 6
- LARGER SITUATION AND UNFAMILIAR USES: 30
  - Larger Situation uses: 8
    - Names (first occurrences): 5
    - Time references: 3
    - Restrictive premodifications: 0
  - Unfamiliar uses: 22
    - NP Complements and Unexplanatory modifiers: 4
    - Appositive clauses: 1
    - Restrictive postmodifications: 15
    - Copula constructions: 2
- NON-IDENTIFIED DDS: 21

Figure 1: Results of the system on the text in the Appendix

The system can show the list of definites, their classification and the co-referential chains (discourse referents referring to the same entity) obtained by the resolution of anaphoric descriptions. The user can also check what was found for each class of DDs or subclass (names, time references, etc.). Finally, as discussed below, the system compares its own classification to another; given an external classification, the system can compute the agreement among its own results with those provided, displays disagreements and calculates a coefficient of agreement between subjects (Kappa statistic, Section 5). Figure 2 shows the co-referential chains found for the text in the resolution process. Each DD is preceded by its NP index and followed by the sentence number (s.) in which it appears.

INDEX - ANTECEDENT - SENTENCE INDEX - ANAPHORIC DD - SENTENCE	INDEX - ANTECEDENT - SENTENCE INDEX - ANAPHORIC DD - SENTENCE
39 the government s.7 66 the government s.11 102 the government s.15 127 the government s.19 137 the government s.21 158 The government s.24 173 the government s.26 178 The government s.27 197 the government s.29 243 the government s.33 248 the government s.34 264 the government s.36 273 the government s.37 313 The government s.42 43 a government panel s.8 59 The panel s.11 45 the problem s.8 112 the problem s.17	48 a report on the extent and causes of the problem s.8 50 the report s.9 52 The report s.10 95 the report s.14 89 the population s.14 96 the population s.14 91 the nation s.14 151 the nation s.23 320 the nation s.42 98 the land s.14 279 The land s.38 103 the government's Land Bureau s.15 123 the Land Bureau s.19 125 the National Assembly s.19 139 the National Assembly s.21 316 the National Assembly s.42 138 three bills s.21 146 the bills s.23 267 the bills s.36

Figure 2: Anaphora resolution (co-referential chains)

INDEX - LARGER SIT/UNF DD - SENTENCE	INDEX - LARGER SIT/UNF DD - SENTENCE
12 the closer they got to saving the $40,000 they originally needed s.1 14 the more the price rose s.1 35 the Parks s.6 37 the long-cherished dream of home ownership s.6 49 the past 15 years s.9 65 the 1988 Seoul Olympics s.11 73 The result s.12 87 the prospects of buying a home s.13 98 the land devoted to housing s.14 120 the past three months s.19 122 the office complex where the Land Bureau is housed s.19 125 the National Assembly s.19 133 the past year s.20 143 the inequities in the current land-ownership system s.22 156 the amount of real estate one family can own, to 660 square meters in the nation's six largest cities ... s.23 171 the resale of property s.26	174 the sale of idle land to the government s.26 180 the average realized for other similar-sized property... s.27 190 the full scope of the penalties s.28 204 the popular standing of President Roh s.29 225 The Citizens Coalition for Economic Justice s.32 235 the value-assessment system on which property taxes are based s.32 246 the Federation of Korean Industries s.34 255 the arguments of business leaders s.35 259 the capitalistic principle of private property s.36 278 the shortage of land s.37 295 The chief culprits s.40 319 the first half of 1989 s.42 326 The Ministry of Finance s.43 354 The maximum allowable property holdings for insurance companies s.46

Figure 3: Larger Situation/Unfamiliar DDs

The larger situation and unfamiliar cases found by the system for the text in the Appendix are given in Figure 3. The first two DDs in Figure 3 (indexed by 12 and 14) do not require the identification of an antecedent for their interpretation, they form together a comparative construction, which is identified by the system when comparing the DD's head noun with the list of special nouns. Other cases of DDs whose classification is due to the list of special nouns are time references indexed by 49, 120, 133, and the unexplanatory modifiers in 319 and 354[7]. The heuristic used to classify non-resolved proper names as larger situation uses gives us the DDs indexed by 35, 65, 125, 246, and 326. Note, however, that for the DD index 35, *the Parks*, the heuristic result is not correct, the DD has as antecedent the NP *Y. J. Park and her family* (s.1). The DDs indexed by 37, 87, 98, 122, 143, 156, 171, 174, 180, 190, 204, 255, 259, 278 and 354 are results of the restrictive postmodification heuristic. DDs indexed by 73 and 295 are examples of copula construction. It is not clear, however, to which type (or types) of use the DDs 73 and 295 belong to, they may also be classified as associative uses. One case of appositive construction is recognised for the DD 225, *The Citizens Coalition for Economic Justice, a public-interest group leading the charges for radical reform* (s.32).

INDEX - NON IDENTIFIED DD - SENTENCE	INDEX - NON IDENTIFIED DD - SENTENCE
13 the price s.1	165 the government set ceiling s.25
18 the 33-year-old housewife s.3	189 the penalties s.28
39 the government s.7	193 The administration s.29
44 the extent s.8	194 the measures s.29
45 the problem s.8	213 the proposed changes s.30
53 the blame s.10	228 the charge s.32
89 the population s.14	247 the critics s.34
91 the nation s.14	272 the constitution s.37
132 the real-estate crisis s.20	351 the proportion s.45
140 The proposed legislation s.22	359 the policies s.47
142 the current land-ownership system s.22	

Figure 4: Non-identified DDs

Figure 4 shows definite descriptions which are not identified by the system. These are often cases of associative uses (e.g., 13, 18, 45, 53, 189, 194) and sometimes larger situation uses (e.g., 39, 91, 142). Note that the syntactic annotation presents some problems as for DD index 44, where the coordination was not properly annotated in the corpus for the NP *the extent and causes of the problem* (s.14).[8]

5 Evaluation of the system

The system has been evaluated by comparing its results with a manual classification of the definite descriptions found in the corpus. At the time when this paper was being written we were working on the annotation of a new collection of texts, that we have subsequently used to evaluate our system with respect to a corpus consisting of unseen data.

---

[7] Note that DDs 12 and 14 are also counted as NP Complements and Unexplanatory modifiers.

[8] Such cases are marked as doubt in the standard manual annotation.

The classification initially made by the authors were compared with the classification made by two external subjects, the three were compiled into one classification which we will call the standard annotation. The two external subjects (named A and B) were instructed to say for each description if it referred (or was related) to an antecedent encountered before in the text or if it was introducing a new discourse referent (larger situation/unfamiliar). In case it had an antecedent, we asked them to tell if that antecedent was introduced by a NP with the same head noun (anaphoric same head) or an associated phrase (associative). The subjects also had the option of classifying the definites as ‘idiom’ or ‘doubt’. We implemented an annotation tool which presents the text to the subjects, shows the descriptions one by one, and asks them to input a classification number for each of them.

The results are shown in Table 2. It is hard to achieve perfect agreement in this exercise. It is not always clear to which of the classes a description belongs to, it may belong to more than one class, and the task of finding the antecedents requires a reasonable amount of concentration. Nevertheless, the two subjects distributed the definite descriptions among the classes I-V in a fairly similar way, and also in a way that is fairly similar to the one we obtained ourselves.

Class	**Total A**	**Percentage A**	**Total B**	**Percentage B**
Anaphoric s. h.	294	28%	332	32%
Associative	160	16%	150	14%
Larg.Sit./Unf.	546	52%	549	53%
Idiom	39	4%	2	0%
Doubt	1	0%	7	1%
Total	1040	100%	1040	100%

Table 2: Classification of descriptions according to Annotators A and B.

In Figure 5, the results of the system for the whole corpus are presented.
There is a large number of NPs considered as potential antecedents comparatively to the number of actual antecedents. Indefinite NPs makes approximately 1/5 of the total of antecedents. Fraurud (Fraurud, 1990) has also observed, in a study of Swedish texts, that from the total number of indefinite NPs in her corpus just a small proportion (1/10) were introductions subsequently referred to, and that indefinites represented only 1/3 of all initial mentions. She also observed a large number of initial mention definites.

The system’s results were verified against the standard classification, presented in Table 3. The results of the comparison are presented in Figure 6. We also identified manually the antecedents for anaphoric descriptions and verified whether the antecedents found by the system were correct. The system obtained for its report of 273 anaphoric descriptions, 244 correct classifications against 29 errors. For the resolution process properly (identification of the antecedent), 225 correct results were computed against 36 errors and 12 partially correct resolutions[9] - the errors in this class are mostly due to noun phrase premodification and discourse segmentation.

[9] Cases of multiple resolution where some of the antecedents are correct and some are not.

NUMBER OF TEXTS: 20
NUMBER OF NOUN PHRASES: 6831
NUMBER OF POTENTIAL ANTECEDENTS: 2067
  Indefinites: 608
  Plurals and Possessives: 773
  Definites: 686

NUMBER OF DEFINITE DESCRIPTIONS: 1040

- ANAPHORA RESOLUTIONS (SAME HEAD) : 273
  Multiply resolved: 17
  Actual antecedents: 140
    Indefinites: 31
    Plurals and Possessives: 16
    Definites: 93

- LARGER SITUATION AND UNFAMILIAR USES: 428
  Larger Situation uses: 154
    Names (first occurrences) : 76
    Time references: 41
    Restrictive premodifications (proper nouns): 37
  Unfamiliar uses: 274
    NP Complements and Unexplanatory modifiers: 40
    Appositive clauses: 28
    Copular constructions: 18
    Restrictive postmodifications: 188

- NON-IDENTIFIED DDS: 339

Figure 5: Global Results (Basic Version)

Class	Total	Percentage of total
Anaphoric s. h.	312	30%
Associative	204	20%
Larger sit./Unfamiliar	492	47%
Idiom	22	2%
Doubt	10	1%
Total	1040	100%

Table 3: Standard Classification

TOTAL DISAGREEMENTS FOR ANAPHORA CLASSIFICATION: 29
TOTAL DISAGREEMENTS FOR ANAPHORA RESOLUTION: 36 correct resolutions: 225 (partially correct: 12)
TOTAL DISAGREEMENTS FOR LARGER SITUATION/UNFAMILIAR: 64 time references: 5; NP complement/unexplanatory modifiers: 7; apposition: 2; name: 12; premodification: 15; copula: 7; postmodification: 16
TOTAL NON CLASSIFIED: 339 anaphoric: 39; associative: 161; larger sit./unfamiliar: 114; idiom: 20; doubt: 5

Figure 6: Evaluation of results (Basic Version)

The heuristics for identifying larger situation and unfamiliar uses resulted in 364 correct results against 64 errors; the errors in these cases are related to the fact that the syntactic features of descriptions do not correspond 100% to a specific semantic/pragmatic class (for instance, a definite description which is postmodified may be anaphoric or associated to some other NP, as in a text about an earthquake: *the earthquake - the suffering that people is coming through*, where the second description is related to the first one). Figure 6 also shows the distribution of errors for each subclass of larger situation/unfamiliar uses of DDs. This gives us an idea of the results separately for each of our assumptions. For instance, the rate of errors for copula construction is 7 errors in 18 cases (38%) while our restrictive postmodification heuristic results in 16 errors in 188 cases (8.5%).

Recall and precision in relation to the identification of anaphoric uses, anaphora resolution and identification of larger situation/unfamiliar uses were estimated. Recall and precision figures for the Basic Version are present in Table 4. Recall is the percentage of correct identification reported by the system in relation to the number of anaphora and larger situation/unfamiliar uses recognised by human evaluation. Precision is the percentage of correct reported results in relation to the total reported.

System's tasks	**Recall**	**Precision**
Anaphora classification	78% (244/312)*100	89% (244/273)*100
Anaphora resolution	72% (225/312)*100	82% (225/273)*100
Larger situation/Unf	74% (364/492)*100	85% (364/428)*100

Table 4: Recall and precision (Basic Version)

Note that anaphora resolution has recall and precision figures slightly lower than anaphora classification, this refers to the cases where an incorrect antecedent is identified by the system but it is classified correctly as anaphoric use.

We measured agreement between subjects A and B to test reliability of the coding system using the Kappa statistics (Carletta, 96), which estimates the non-chance agreement among coders. The coefficient of agreement between the two subjects was 0.68 for the whole corpus and 0.77 when considering only the data identified by the system, with 0.8 established as a satisfactory level. Considering our first analysis together with the other two the coefficient of agreement was 0.75 for both the whole corpus and the data handled by the system. The overall agreement among the three analysis and the results produced by the system was 0.77. This shows us that the number of agreement is not affected when the system is considered, and, if the system's output is not perfect, at least its performance produce an annotation which is as different from the manual annotation as the annotations produced by the subjects are different among themselves. The disagreements among subjects may be due to a failure in finding an antecedent which is present in the text but is not evident[10]; sometimes they match a DD with an antecedent which has a similar but different head noun (as in *lawsuit - the suit*). The subjects confuse the classes larger situation/unfamiliar and associative: descriptions such as *the markets, the seed companies, the male-fertile*

[10] For instance, when an antecedent is available for a DD which is a proper noun but the antecedent is not observed. Note that there is a difference between being anaphoric and co-referential: anaphoric DDs would be entirely dependent on the identification of an antecedent for its interpretation while co-referential DDs refer to a discourse-old entity but its interpretation may be independent of it (as in *the West Coast - the West Coast*). Although we use the term anaphora we are in fact dealing with co-referential DDs - although we have made it clear to the subjects, this still may be the cause of some problems in their agreement.

plants, the ear, the security business, are examples of descriptions classified both as larger situation/unfamiliar and associative, they are related to other references in the text but at the same time they are understood as independent references. Of the 339 definites currently not treated by the system, associative uses are the largest class (161), followed by larger situation/unfamiliar (114). Our next goal is the improvement of our treatment of larger situation/unfamiliar and associative DDs.

5.1 Evaluation of Segmentation and Premodifiers Heuristics

We compared the results above with those obtained by versions of the algorithm that did not use some of the proposed heuristics. The results presented by the system <u>without implementing the segmentation heuristic</u> (Version 2) are given in Figure 7.

ANAPHORA (same head): 322 (Multiply resolved : 47)
LARGER SITUATION/UNFAMILIAR: 412
NON-IDENTIFIED: 306

DISAGREEMENTS FOR ANAPHORA CLASSIFICATION: 57
DISAGREEMENTS FOR ANAPHORA RESOLUTION: 72
correct resolutions: 216 (partially correct: 34)
DISAGREEMENTS FOR LARGER SIT/UNFAMILIAR: 56

Figure 7: Results and evaluation (Version 2)

The difference between the number of errors obtained by the versions of the system with and without a segmentation heuristic (approximately 30) gives us a rough evaluation of the impact of discourse segmentation on interpreting anaphoric definite descriptions. About 10% of anaphoric definite descriptions in a text may result in multiple interpretations when segmentation is not taken into account. The results of Version 3 <u>with no heuristics for pre-modifiers and segmentation</u> are given in Figure 8.

ANAPHORA (same head): 366 (Multiply resolved: 47)
LARGER SITUATION/UNFAMILIAR: 397
NON-IDENTIFIED: 277

DISAGREEMENTS FOR ANAPHORA CLASSIFICATION: 84
DISAGREEMENTS FOR ANAPHORA RESOLUTION: 106
correct resolutions: 224 (partially correct: 36)
DISAGREEMENTS FOR LARGER SIT/UNFAMILIAR: 50

Figure 8: Results and evaluation (Version 3)

Recall and precision for versions 2 and 3 are presented in Table 5.

System's tasks	Recall V.2	Prec. V.2	Recall V.3	Prec. V.3
Anaphora classification	84%	82%	90%	77%
Anaphora resolution	69%	67%	71%	61%
Larger situation/Unf	72%	86%	70%	87%

Table 5: Recall and precision figures (Versions 2 and 3)

Recall and precision figures for the 3 versions of the system are graphically summarised in Figure 9.

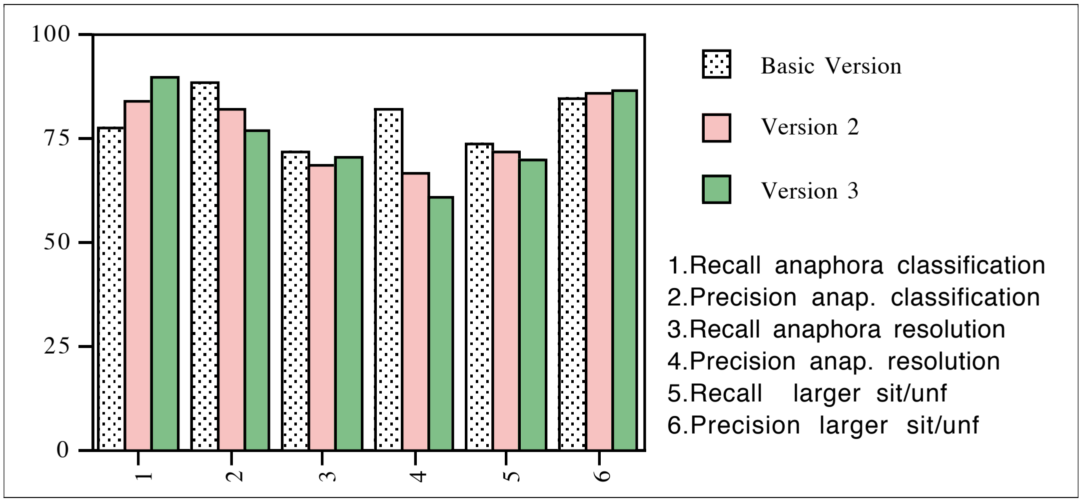

Figure 9: Comparison of different versions of the system

Note how an increased recall for anaphora classification correspond to a lower recall for larger situation and unfamiliar ones.

Errors of the system in anaphora resolution are exemplified by cases such as *rules - the new rules* (where the premodifier modifies the noun in such a way that a different set of 'rules' is referred to); or, the sequence *a house - the house, a couple living in a motorhome - the house itself* (in a text which introduces a house, refers to that house with a definite description and later on refers to a different house whose introduction is made by associative means: a couple living in a motor home because their house -*the house itself* - was demolished by an earthquake).The system's errors in the classification of larger situation/unfamiliar uses as discussed earlier are due to the fact that the presence of certain syntactic structures do not guarantee a specific type of use of DDs, appositive and copula construction or postmodified DDs, for instance, may be cases of anaphoric or associative uses.

The heuristics proposed for the treatment of premodifiers in the resolution process were also tested. We summarise their recall (R) and precision (P) figures in Table 6. All the following versions considered our heuristic for segmentation: Version A is the Basic Version; Version B did not consider the role of premodifiers at all; Version C matches only DDs whose premodifiers are a subset of the modifiers in the antecedent; and finally, Version D matches any form of DDs only with antecedents with no premodifiers. We expect to improve these heuristics. One possibility is by making them work in connection with a treatment of postmodifiers.

System's task	**R/P** A	**R/P** B	**R/P** C	**R/P** D
Anaphora resolution	72%/82%	73%/77%	68%/83%	61%/87%

Table 6: Recall and precision figures (Versions A, B, C and D)

6 Conclusions

Our aim is to develop a system for resolving definite descriptions whose performance on arbitrary texts can be evaluated in a quantitative fashion. We hope to gain from this exercise both a better idea about the uses of definite descriptions and indications about which sort of common-sense knowledge is actually needed.

The activities we have been involved in include: a) the implementation of a tool that retrieves data from a treebank, searches for determined linguistic structures (definite descriptions), shows them to an analyst, and produces an annotated corpus based on the judgement of the analyst; b) a study of the uses of definite descriptions in a corpus, that showed that about 70% of definites in our corpus are potentially interpretable with the techniques employed here, and identified the most common forms of definite descriptions; c) the implementation of a system that resolves anaphoric definites and identifies novel ones; and d) an evaluation of the results.

Anaphoric resolution based on same-head noun might be thought of as straightforward, but in fact we encountered several difficulties, such as how to deal with premodifiers and postmodifiers. Our current heuristics on premodifiers and segmentation seem to work satisfactorily: the system achieved a recall of about 72% and a precision of about 82% for anaphoric definite descriptions.

Our analysis of the corpus made us notice that a great number of definite descriptions introduce a novel discourse referent; most computational approaches to date have considered mainly their anaphoric role. This result suggests that any general treatment of definites should include methods for recognising such definites. We proposed tests based on lexical and syntactic knowledge for identifying novel definites, including looking for an embedding apposition or copula, checking whether the head noun calls for a complement, and inspecting premodifiers and postmodifiers. The evaluation of an implementation of these tests showed a recall of about 74% and a precision of about 85% for larger situation and unfamiliar uses of DDs. Considering all the occurrences of DDs in the corpus the system presented a recall of 56% and a precision on 84%.

The annotation exercise resulted in a fairly good agreement. The agreement is necessary to show reliability of the classification proposed and also to draw conclusions on the performance of the system.

Further work will include improving our heuristics for novel definites and using a source of lexical information to address the problem of associative definites which in the kind of texts we have worked on represents about 20% of the cases. Also, we want to consider the syntactic annotation produced by other parsers, and different corpora are to be analysed, to verify other genres such as fiction, instructional texts and also other languages.

Comparing our work to previous computational approaches to discourse anaphora - (Carter, 1987) and (Sidner, 1979) are examples - we note that differently from the others we have dealt specifically with definite descriptions headed by the definite article; we propose a robust system which runs over domain independent and unlimited data (our only limitation at present being the syntactic annotation); we propose heuristics for the identification of descriptions which are novel in the discourse; and we present a quantitative evaluation of the results produced by our system.

Resolution of definite descriptions as a feature of discourse analysis may serve as a basis for many different applications. Also important is the development of software tools to retrieve and process corpora linguistically, as well as the use of resulting annotated corpora for further research development. In Natural Language Processing and Computational Linguistics we still have to learn about the respective roles of syntactic knowledge, lexical knowledge, and common sense reasoning in natural language interpretation, the work presented is a step in this direction.

Appendix

The following text from our corpus is used to illustrate the system (in Section 4 of the paper). It is text wsj_0761.par from Penn Treebank, CDROM1, Preliminary Release, Version 0.5, Dec. 1992. Each sentence in the text is preceded by a sequential number.

1 Y.J. Park and her family scrimped for four years to buy a tiny apartment here, but found that **the** closer they got to saving **the** $ 40,000 they originally needed, **the** more **the** price rose. **2** By this month, it had more than doubled.

3 Now **the** 33-year-old housewife, whose husband earns a modest salary as an assistant professor of economics, is saving harder than ever. **4** "I am determined to get an apartment in three years”, she says. **5** "It’s all I think about or talk about”.

6 For **the** Parks and millions of other young Koreans, **the** long-cherished dream of home ownership has become a cruel illusion. **7** For **the** government, it has become a highly volatile political issue.

8 Last May, a government panel released a report on **the** extent and causes of **the** problem. **9** During **the** past 15 years, **the** report showed, housing prices increased nearly fivefold. **10 The** report laid **the** blame on speculators, who it said had pushed land prices up nine fold.

11 The panel found that since 1987, real-estate prices rose nearly 50% in a speculative fever fueled by economic prosperity, **the** 1988 Seoul Olympics and **the** government’s pledge to rapidly develop Korea’s southwest. **12 The** result is that those rich enough to own any real estate at all have boosted their holdings substantially. **13** For those with no holdings, **the** prospects of buying a home are ever slimmer.

14 In 1987, a quarter of **the** population owned 91% of **the** nation’s 71,895 square kilometers of private land, **the** report said, and 10% of **the** population owned 65% of **the** land devoted to housing **15** Meanwhile, **the** government’s Land Bureau reports that only about a third of Korean families own their own homes.

16 Rents have soared along with house prices. **17** Former National Assembly man Hong Sa-Duk, now a radio commentator, says **the** problem is intolerable for many people. **18** "I’m afraid of a popular revolt if this situation isn’t corrected”, he adds. **19** In fact, during **the** past three months there have been several demonstrations at **the** office complex where **the** Land Bureau is housed, and at **the** National Assembly, demanding **the** government put a stop to real-estate speculation.

20 President Roh Tae Woo’s administration has been studying **the** real-estate crisis for **the** past year with an eye to partial land redistribution. **21** Last week, **the** government took three bills to **the** National Assembly. **22 The** proposed legislation is aimed at rectifying some of **the** inequities in **the** current land-ownership system. **23** Highlights of **the** bills, as currently framed, are : - A restriction on **the** amount of real estate one family can own, to 660 square meters in **the** nation’s six largest cities, but more in smaller cities and rural areas. **24 The** government will penalize offenders, but won’t confiscate property.

25 - A tax of between 3% and 6% on property holdings that exceed **the** government set ceiling.

26 - Taxes of between 15% and 50% a year on "excessive" profits from **the** resale of property, or **the** sale of idle land to **the** government. **27 The** government defines excessive profits as those above **the** average realized for other similar-sized properties in an area.

28 - Grace periods ranging from two to five years before **the** full scope of **the** penalties takes effect.

29 The administration says **the** measures would stem rampant property speculation, free more land for **the** government’s ambitious housing-construction program, designed to build two million apartments by 1992 - and, perhaps, boost **the** popular standing of President Roh.

30 But opposition legislators and others calling for help for South Korea’s renters say **the** proposed changes don’t go far enough to make it possible for ordinary people to buy a home. **31** Some want lower limits on house sizes others insist on progressively higher taxation for larger homes and lots.

32 The Citizens Coalition for Economic Justice, a public-interest group leading **the** charge for radical reform, wants restrictions on landholdings, high taxation of capital gains, and drastic revamping of **the** value-assessment system on which property taxes are based.

33 But others, large landowners, real-estate developers and business leaders, say **the** government's proposals are intolerable. **34** Led by **the** Federation of Korean Industries, **the** critics are lobbying for **the** government to weaken its proposed restrictions and penalties.

35 Government officials who are urging real-estate reforms balk at **the** arguments of business leaders and chafe at their pressure. **36** "There is no violation of **the** capitalistic principle of private property in what we are doing", says Lee Kyu Hwang, director of **the** government's Land Bureau, which drafted **the** bills. **37** But, he adds, **the** constitution empowers **the** government to impose some controls, to mitigate **the** shortage of land.

38 The land available for housing construction stands at about 46.2 square meters a person - 18% lower than in Taiwan and only about half that of Japan.

39 Mr. Lee estimates that about 10,000 property speculators are operating in South Korea.

40 "**The** chief culprits", he says, "are big companies and business groups that buy huge amounts of land not for their corporate use, but for resale at huge profit".

41 One research institute calculated that as much as 67% of corporate-owned land is held by 403 companies - and that as little as 1.5% of that is used for business. **42 The** government's Office of Bank Supervision and Examination told **the** National Assembly this month that in **the** first half of 1989, **the** nation's 30 largest business groups bought real estate valued at $ 1.5 billion.

43 The Ministry of Finance, as a result, has proposed a series of measures that would restrict business investment in real estate even more tightly than restrictions aimed at individuals.

44 Under those measures, financial institutions would be restricted from owning any more real estate than they need for their business operations. **45** Banks, investment and credit firms would be permitted to own land equivalent in value to 50% of their capital - currently **the** proportion is 75%.

46 The maximum allowable property holdings for insurance companies would be reduced to 10% of their total asset value, down from 15% currently. **47** But Mrs. Park acknowledges that even if **the** policies work to slow or stop speculation, apartment prices are unlikely to go down. **48** At best, she realizes, they will rise more slowly - more slowly, she hopes, than her family's income.

References

Carletta, J. 1996. Assessing agreement on classification tasks: the kappa statistic. *Computational Linguistics*. Research Squib, Forthcoming.

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

Chinchor, N. A, and B, Sundheim 1995. Message understanding conference MUC tests of discourse processing. In Proc. *AAAI Spring Symposium on Empirical Methods in Discourse Interpretation and Generation,* pages 21-26, Stanford.

Clark, H.H. 1977. Bridging. In P. N. Johnson-Laird and P. C. Wason, editors, *Thinking Readings in Cognitive Science*. Cambridge University Press, London and New York.

Clark, H.H. and C.R. Marshall 1981. Definite reference and mutual knowledge. In *Elements of Discourse Understanding*. Cambridge University Press, New York.

Fligelstone, S. 1992. Developing a scheme for annotating text to show anaphoric relations. In G. Leitner, editor, *New Directions in English Language Corpora: Methodology, Results, Software Developments*. TiEL, Mouton de Gruyter Berlin.

Fraurud, K. 1990. Definiteness and the Processing of Noun Phrases in Natural Discourse. *Journal of Semantics,* 7: 395-433.

Garside, R., G. Leech and G. Sampson 1987. Eds. *The Computational Analysis of English, a Corpus Based Approach*. Bath Press, Avon, England.

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

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

Leech G. 1991. The state of art in corpus linguistics In K. Aijmer and B. Altenberg editors *English Corpus Linguistics* Longman London

Marcus, M. P., B. Santorini and M. A. Marcinkiewicz 1993. Building a large annotated corpus of English: the Penn Treebank. *Computational Linguistics*, 19(2):313-330.

Poesio, M. 1993. A situation theoretic formalization of definite description interpretation in plan elaboration dialogues. In P. Aczel, D. Israel, Y. Katagiri and S. Peters, editors, *Situation Theory and its Applications* vol. 3. CSLI, Stanford, chapter 12, pages 339-374.

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