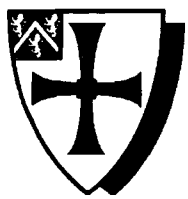
Interpretation of anaphoric expressions in the Lolita system

Agnieszka Joanna Urbanowicz

University of Durham



Interpretation of Anaphoric Expressions in the LOLITA system

Agnieszka Joanna Urbanowicz

Laboratory for Natural Language Engineering,

Department of Computer Science.

Submitted in partial fulfilment of the

requirements for the degree of

Doctor of Philosophy

Abstract

This thesis addresses the issue of anaphora resolution in the large scale natural language system, LOLITA. The work described here involved a thorough analysis of the system's initial performance, the collection of evidence for and the design of the new anaphora resolution algorithm, and subsequent implementation and evaluation of the system.

Anaphoric expressions are elements of a discourse whose *resolution* depends on other elements of the preceding discourse. The processes involved in anaphora resolution have long been the subject of research in a variety of fields.

The changes carried out to LOLITA first involved substantial improvements to the core, lower level modules which form the basis of the system. A major change specific to the interpretation of anaphoric expressions was then introduced. A system of filters, in which potential candidates for resolution are filtered according to a set of heuristics, has been changed to a system of penalties, where candidates accumulate points throughout the application of the heuristics. At the end of the process, the candidate with the smallest penalty is chosen as a referent. New heuristics, motivated by evidence drawn from research in linguistics, psycholinguistics and AI, have been added to the system.

The system was evaluated using a procedure similar to that defined by MUC6 (DARPA 1995). Blind and open tests were used. The first evaluation was carried out after the general improvements to the lower level modules; the second after the introduction of the new anaphora algorithm.

It was found that the general improvements led to a considerable rise in scores in both the blind and the open test sets. As a result of the anaphora specific improvements, on the other hand, the rise in scores on the open set was larger than the rise on the blind set. In the open set the category of pronouns showed the most marked improvement.

It was concluded that it is the work carried out to the basic, lower level modules of a large scale system which leads to biggest gains.

It was also concluded that considerable extra advantage can be gained by using the new weights-based algorithm together with the generally improved system.

Acknowledgements

My thanks go first and foremost to my supervisor, Professor Roberto Garigliano, for his continued guidance and support throughout this project and for always encouraging me to believe that it would succeed. Needless to say, without him this thesis would not exist.

I would also like to thank 3F Ltd for employing me part-time during the final year of the project, and particularly Rick Morgan and Russell Collingham for being very supportive during the final weeks of writing up.

I would probably never have gone down this path were it not for an interview with Professor Phil Mars of Durham University, after which Professor Mars put me in touch with the Laboratory of Natural Language Engineering. I will always be grateful for his advice.

Several people have helped with the preparation of the final copy of this thesis through their constructive comments, proof reading and sorting out the copying, so a BIG THANK YOU to Oliver, Dave, Dominika and Debbie. Thanks also to Sanjay, Kevin, Chris and Paul for putting up with me during the times when they had to share an office with me, and thank you to my family and friends (Mum, Dad, Buba, Jean, Harry, Laura, Debbie, Matthew and Giles) who have been patient, understanding, kind and affectionate during many difficult times throughout the last five years.

Finally, I would like to thank Oliver for his love and support and for never failing to see the optimistic side of things.

Declaration

The material contained within this thesis has not previously been submitted for a degree at the University of Durham or any other university. The research reported within this thesis has been conducted by the author unless indicated otherwise.

Contents

List of Figures

Chapter 1

Methodological Introduction

1.1 AI: modelling of human behaviour

In this project an attempt is made to model an aspect of human behaviour which concerns a certain linguistic task. The following assumptions are made:

- although the hope is to achieve something which is fairly general, only a small subset of the behaviour is chosen as the object to be modelled
- an attempt is made to model first and foremost the behaviour on which human subjects agree
- in cases where human behaviour differs between people, i.e. there is disagreement between human subjects, it is considered a success if the model performs the same as the majority of humans
- it is not considered a failure if the model predicts a solution used by a minority; this is considered acceptable, though a smaller success.

1.1.1 Definition of the task

The aspects of human behaviour which are dealt with here have been chosen and described by the organising committee of the 6th Message Understanding Conference (DARPA 1995), sponsored by the Defense Advanced Research Projects Agency of the United States.

The particular task to be modelled in this project is designed specifically for the purpose of evaluating current natural language technology and for promoting further advancements in this technology DARPA (1995, p. 7).

The performance of humans on this task has been shown to be fairly consistent (Sundheim 1995) and it will be assumed here that the majority of human subjects fully agree on at least the core elements of the task.

1.2 Natural Language Engineering

The framework of Natural Language Engineering (or NLE) is the general context of this project. Insights from the fields of artificial intelligence, computational linguistics, theoretical linguistics, psycholinguistics and cognitive science are used as ideas for the development of the model. However, the aim here is not to provide (and/or test) a theory of behaviour that would be true of humans, though an attempt is made to avoid implementing strategies which appear to be cognitively implausible. Instead, the aim is to develop a system which produces behaviour that is acceptable to humans.

"The principal, defining characteristic of NLE work is its objective: to engineer products which deal with NL and which satisfy the constraints in which they have to operate." (Garigliano 1995). It's not enough to design a model which can be tested with a pen and paper. Such an aim might be interesting too, in its own right, but it would fall short of the objectives of NLE.

1.2.1 Why work with LOLITA?

LOLITA[1] is a natural language system which is currently under development in the Laboratory for Natural Language Engineering, at the University of Durham. It is a large scale system designed around a core of natural language capabilities (a detailed description follows in chapter 5).

The key features of the system are, on the one hand, its large semantic network, which can store all kinds of knowledge and support various forms of reasoning; and on the other, its approach to the analysis of natural language. The system attempts a full, 'deep' analysis of the input (including a full parse) and aims to produce a semantic representation of the text. These key features are believed to be extremely important for the performance of the task addressed in this project.

Furthermore, if the ultimate aim is to develop a working and useful product, it is much more worthwhile to integrate the new model within an existing, large scale system, rather than try to develop an independent system that performs only the given task. This is because, on its own, the performance of the task is not hugely beneficial. However, when taken together with a big, general purpose, large-scale system such as LOLITA, the potential for developing practical applications increases dramatically.

The whole task of processing natural language is vast in computational terms. In order to offer the hope of achieving a realistic level of competence in such a task it seems necessary to work within a framework of the ambitious scope of LOLITA.

1.2.2 Aspects of working with LOLITA

The task that this project aims to model relies on other tasks being carried out by several other components of the LOLITA system. This poses two potential

[1] Large-scale Object-based Linguistic Interactor, Translator and Analyser

problems.

The first problem arises from the fact that many of the core components are currently being developed by other researchers and are therefore constantly changing. This makes the evaluation of the proposed model less than straight-forward. It is not always easy to ascertain whether the behaviour resulting at the end of the development are due to the newly developed model or to the changes or improvements in other aspects of the system.

The second problem is that the proposed model has to rely on a fairly successful performance of other components of the system. If some aspects of that performance are found to be less than satisfactory, it may not always be possible to redress the situation, without introducing fundamental changes to the system as a whole. An attempt is made therefore to either limit the task at hand to only those phenomena which do not pose major problems to LOLITA's subcomponents or, whenever possible, to make sure that all the necessary subcomponents are working well.

The final evaluation of the system proposed in this project is performed in such a way that allows us to assume that all the relevant components are working fairly well, though perfection cannot be expected.

1.3 The aims of the project

The primary aim of the project is:

- to improve the performance of the LOLITA system on the chosen task.

The performance is tested and evaluated using the procedures outlined in section 1.4, below.

Other aims of the project are:

- to discover what levels of performance are achievable using the LOLITA system
- to improve the interpretative components of the LOLITA system
- to gain insights into the issue of what techniques achieve better results on the chosen task.

1.4 Stages of the project

The steps taken in this project are as follows:

1. Select a set of 15 articles from the training corpus provided by the Linguistic Data Consortium (via the MUC6 organisers), to constitute the so called "training set"; prepare answer keys in accordance with the task definition; additionally select one long article on a similar topic from the British press. Prepare an answer key for the single article likewise;
2. Select two sets of 10 articles from the MUC6 training corpus, tag them according to the task description and put aside to use as semi-blind tests during development and as an aid in the final assessment of the evaluation results;
3. Select the 30 articles used in the MUC6 formal evaluation as a formal blind test;
4. Debug the LOLITA system which was used at MUC6 ("the old system") using the articles in the training set and the single article from the British press. The objective at this stage is to eliminate obvious bugs and make the old system perform in the way that it is expected to, without trivial errors;
5. Improve all those aspects of the analysis of the training set and the single article that are independent from the task to be modelled;
6. Evaluate the "improved old system" using the single article, the training set, the semi-blind sets and the formal blind test set;

7. Develop and implement the new algorithm to model the task at hand, while keeping all other aspects of LOLITA unchanged;

8. When the scores on the single article, the training set and the semi-blind sets improve, freeze the "new system" and evaluate it using the formal blind test;

9. Compare and assess the results obtained from the "improved old system" with respect to the "new system";

10. Analyse selected examples demonstrating the effects of the new algorithm;

The sequence of steps 6 and 7 could be seen as an 'ideal world' scenario. In practice, however, developing and implementing a new algorithm is a slow process and other changes and improvements to the whole system are likely to be made during this time. It is important, therefore, to ensure that it is possible to switch between the existing and the new algorithm for the purposes of the evaluation. As a result, when considering real time-scales, the evaluations of the existing and the new systems are carried out simultaneously.

1.5 The progression of the thesis

Chapter 2 describes in detail the kind of task which the project aims to deal with.

Chapter 3 reviews other work broadly related to the task.

Chapter 4 describes how other systems competing at MUC6 deal with the same task.

Chapter 5 provides a general introduction to the LOLITA system.

Chapter 6 presents the way LOLITA deals with the task at hand, before the introduction of any improvements (this is a description of the "old system").

Chapter 7 describes general improvements carried out to the system since the time of MUC6, resulting in an "improved old system".

Chapter 8 deals with issues relating to the evaluation of large scale systems and their sub-components. The chapter also discusses two methods of scoring.

Chapter 9 presents the results of the evaluation carried out after the general improvements to the system and before the introduction of the new algorithm.

Chapter 10 describes the new way of dealing with the task at hand.

Chapter 11 presents the results obtained using the system with the new algorithm (the "new system").

Chapter 12 contains the conclusions and discusses future work.

Chapter 2

Interpreting anaphora

2.1 Introduction

The aspect of human behaviour which is modelled in this project is the so called *anaphora resolution process*, or, in other words, the task of identifying co-referential natural language expressions.

Many surveys of the 'anaphora problem' have been published to-date. Comprehensive surveys can be found in: Webber (1979), Sidner (1979), Hirst (1981) or Carter (1987), among others.

In this chapter the problem is reviewed briefly, with most emphasis given to the types of anaphora that are central to the project. The use of terminology is explained. Section 2.4 gives a flavour of what are considered to be core and peripheral cases of anaphora. In sections 2.5 and 2.6 examples of selected types of anaphoric dependencies are discussed.

The chapter addresses the following issues:

- the notions of *utterance* and *utterance interpretation*

- what is understood by the term *anaphoric expression*
- what is involved in finding a referent for an anaphoric expression
- why this process is a problem for natural language systems.

2.2 Utterance Interpretation

The definition of the term *utterance* adopted in this project will be that of Smith & Wilson (1979, p. 45): an *utterance* is a string of words produced by a speaker or writer on a given occasion and in some context.

To successfully *interpret* an utterance is, for the purposes of this project, to arrive at some representation of the proposition (or message or meaning) that the speaker or writer intended to communicate through the utterance. This process is commonly thought to involve, in some way, at least the following steps: parsing, semantic analysis, resolution of ambiguities and integration of the message with prior knowledge.

2.3 Discourse

Discourse is taken to be a sequence of one or more utterances. In the context of this project, any utterance or a sequence of utterances (for example, a newspaper article or a conversation), written or spoken with an intention to communicate and as such, constituting a coherent whole, will be considered to constitute discourse.

2.4 Anaphoric expressions and co-reference

The terms *anaphor* or *anaphoric expression* are used interchangeably throughout this thesis to describe any linguistic expression whose referent has already been mentioned in the preceding discourse. The term *referent* is taken to be some object (entity, event, etc.) in the world or in a representation of the world. The preceding mention of the referent is usually called the *antecedent* of the anaphoric expression.

The relationship between the anaphoric expression and its antecedent is sometimes called "co-reference", a term which will be adopted here. To assert that a co-reference relation holds between two expressions means that the two expressions refer to one and the same object (entity, event, etc.) in the world or in a representation of the world.

The process of finding the antecedent for an anaphoric expression is commonly termed *anaphora resolution*[1].

Typically, anaphoric expressions include **personal pronouns** (such as *he*, *she*, *they* etc.), **reflexive pronouns** (e.g. *himself, herself*), **definite noun phrases** (i.e. nouns preceded by a definite article, e.g. *the woman*, *the car* as well as nouns specified by some other determiner such as *Britain's* or *Michael's*) and **proper names** (e.g. *Michael*, *Britain*). These types of anaphora could be considered as 'core' cases.

Other types of anaphoric expressions are the so-called **verb phrase anaphora** (*did* in *Michael slipped and Paul did too*), or **one anaphora** (as in *Michael saw a tiger and Paul saw one too*). These types are different from the core cases above, because here the relationship between the anaphoric expression and the antecedent is not that of identity (e.g. the tiger that Paul saw is not the same as the one that Michael saw).

Sometimes expressions such as predicative nominatives, e.g. *John Smith, president*

[1]NB: the term *anaphora* is understood to be the plural of *anaphor*.

of ABC Corp., ..., could be said to enter into a co-reference relation. In this case, the expression *president* would be co-referential with *John Smith* or possibly with the whole phrase *John Smith, president of ABC Corp.* These types of co-references are referred to in this project as **is_a based co-references**.

Also expressions such as *the sales rose to $5,000,000* might be seen to contain co-referential links. In this case the noun phrase *the sales* could be seen co-referring with *$5,000,000.* This sort of co-reference is labelled **function and value** co-reference.

The "is_a" based co-references and "function and value" co-reference could be argued to be on the periphery of the anaphora problem. However, they form part of the MUC evaluation, along with the core cases.

Anaphora of all kinds are ubiquitous in both written and spoken discourse. Despite usually being resolved by human hearers or readers with great ease, they have proven a big challenge to account for in a theoretical way.

2.5 More examples of core anaphora cases

The two sentences in (1) illustrate a simple case of an anaphoric dependency, or "co-reference":

(1) *Jenny left the room. She was smiling.*

In the absence of any other contextual information the pronoun *she* is immediately understood to co-refer with *Jenny*, or, to put it another way, *she* and *Jenny* refer to the same entity/person in the world.

The example in (1) is straight-forward, since there is one reference to a person and one pronoun in the sentences, and so in the absence of information to the contrary, any human reader simply assumes that the two co-refer.

More complex examples involve cases where several possible referents are present in the discourse, therefore the choice of an antecedent for any given anaphoric expression is not so straight-forward. For example:

(2) *Susan sold the car to Jenny because she decided to take up cycling.*

A human reader/hearer would probably decide that *she* in the second clause of (2) is co-referential with Susan, presumably on the basis of what is known about cycling, driving and selling. The way the knowledge is used and the exact computational steps involved in arriving at this decision are of interest here and would have to be accounted for in detail if we were to model the process in a computer system.

As mentioned above, proper names and noun phrases also enter into co-reference relations, for example in:

(3) *BMW is Germany's biggest auto-maker. The company employs more than a thousand workers. BMW's spokesman gave a press conference on Friday.*

Germany's biggest auto-maker, *the company* and the two occurrences of *BMW* all co-refer. The link between *BMW* and *Germany's biggest auto-maker* may be easy to resolve: it could be derived from the use of the copula verb *to be*, which asserts identity here. However, the link between *the company* and the first two entities requires more effort and possibly presupposes that the hearer or the NL system knows that *company* can be a general term for *auto-maker*. By contrast, the link between the two occurrences of *BMW* might be arrived at through simple pattern matching.

Entities such as organisations (companies, or businesses) often prove tricky as antecedents. Quite frequently they can be referred to, often within one and the same discourse, with the use of either a third or a first person pronoun or, with either a plural or a singular pronoun. For example:

(4) *BMW is Germany's biggest auto-maker. "We employ more than a thousand workers", said the company spokesman.*

Arguably, the pronoun *we* can be interpreted co-referentially with *BMW*, regardless of the violation of the person and number agreement. Similarly:

(5) *BMW is Germany's biggest auto-maker. On Friday, the company announced their intention to bid for Rover.*

Here the plural possessive determiner *their* is clearly a reference to *BMW*, despite the latter being singular; such use presumably implies that the company is seen as some sort of collective body.

Other cases to consider include this one from Webber (1983):

(6) a. *John gave Mary five dollars. It was more than he gave Sue.*
b. *John gave Mary five dollars. One of them was counterfeit.*

The examples in (5) and (6) show an interaction between the processes of co-reference resolution and the disambiguation of senses.

In the 'BMW' case, the phrases *the company* or *BMW* could be seen as ambiguous between the meaning of a single human organisation and the meaning which designates all the people that work for the company (or even one which really refers to only those people who make decisions in it). In the case of *five dollars*, the expression 'dollars' seems to be ambiguous between the meaning 'currency name' and 'banknotes'. In both examples, it would appear that only when the pronoun is encountered can a firm decision be made as to which of the two meanings was intended in each case. This shows that sense disambiguation might have to work in conjunction with pronoun resolution.

2.6 Anaphora without explicit antecedents

In most of the examples used so far, the antecedent for the anaphoric expression is explicitly mentioned in the preceding utterances. The problem lies in choosing the appropriate one from a list of possibilities. But this is not always the case.

For example, in the so called *cataphoric* uses, the mention of the referent for the pronoun is found in the discourse following the pronoun:

(7) *When he returned home, John found his front door was open.* (Neale 1990)

In some examples, the antecedent may not be explicitly present in the discourse at all:[2]

(8) a. *No male driver admits that he is incompetent.*

b. *A man who gave his pay-check to his wife was wiser than the man who gave it to his mistress.*

c. *Jenny was frightened by her neighbour's Doberman. They are dangerous beasts.*

d. *Jack went to New York. The trip changed his life.*

The (8a) example is problematic because the pronoun *he*, somewhat paradoxically, appears to refer to an entity which doesn't exist. The interpretation of such a pronoun will thus call for a more complicated procedure than that of just identifying a likely referent from some existing list of possibilities.

In the next example the pronoun *it* clearly refers to a 'pay-cheque' but not to the one which is explicitly mentioned.

[2]Examples (8a-c) are sometimes referred to as *sloppy identity anaphora*

In (8c) the referent of *they* is understood to be a universal set of Dobermans, even though only one member of this set is explicitly mentioned.

In the final example, (8d), *the trip* is, most likely, Jack's trip to New York — an event which has to be inferred from the knowledge that 'going to places' somehow involves 'trips'.

2.7 What co-reference links are possible

2.7.1 Links within sentence boundaries

Depending on their syntactic position within a sentence, some types of anaphoric expressions allow references to some antecedents but not others. Syntactic conditions, therefore, can be said to constrain interpretation of anaphoric expressions to some extent. The following set of sentences illustrate the issues involved:

(9) a. *Michael said that he blamed Paul.*
b. *Michael said that the man blamed Paul.*
c. *He said that Michael blamed Paul.*
d. *Michael blamed him.*
e. *Michael blamed himself.*

In (9a) it is possible to interpret *he* to mean *Michael*, but when the pronoun is replaced by a definite NP, such an anaphoric link is no longer licensed, i.e. *the man* in (9b) cannot be *Michael*, or at least it's not very easy to make such a link.

Similarly, it is natural to assume that *He* and *Michael* in (9c) or *Michael* and *him* in (9d) refer two different people, respectively.

The contrast between (9c) and (9d) shows that where a pronoun cannot be understood to refer to a subject, a reflexive pronoun can.

These sort of rules appear, at first, to be very strong, almost unbreakable. However it is possible to find a situation where even they can be broken. If the example (10) were to be uttered soon after John Major's loss of the general election:

(10) *He lost the election, because John Major had mismanaged the campaign.*

it might have sounded slightly unnatural, but there would have been no difficulty in linking the *He* with *John Major.* (More discussion on this and related issues can be found in Reinhart (1986) or Lasnik (1989)).

2.7.2 Co-reference outside the sentence

The conditions mentioned in section 2.7.1, above, apply only to intra-sentential anaphora. When it comes to reference outside the sentence — the so called inter-sentential anaphora — co-reference links seem to be influenced by a variety of other factors, about which no unified theory currently exists. The factors often put forward as playing an important part are: some notion of discourse focus, a notion of salience of referents, distance between the anaphoric expression and its antecedent, and a 'first mention effect', among others.

Various accounts of such factors have been proposed and a selection of these is reviewed in chapter 3.

2.8 When do we search for an antecedent?

Do all types of anaphoric expressions always trigger a search for a possible antecedent? It seems that pronouns nearly always require an antecedent. Exceptions

here are the so called pleonastic uses of the pronoun *it*, such as:

(11) *It seems that Jenny fainted.*

Similarly, some uses of *there* are non-referential:

(12) *There is a lot of work to do.*

as opposed to the referential use, for example:

(13) *Jenny emigrated to America. She settled there.*

where *there* functions as a referential adverbial.

Definite noun phrases very often require antecedents, though corpus research has demonstrated that possibly as much as 50% of the time, definite noun phrases introduce a new entity into the discourse, or designate some generic concept (Fraurud 1990, Poesio & Vieira 1997), rather than refer to something that has already been mentioned. Typical cases of where a definite noun phrase does not require an antecedent are when it refers to a generic entity, e.g.:

(14) *The horse is a beautiful animal.*

and, when it's the head of a relative clause, e.g.:

(15) *The man who came to visit yesterday brought some flowers.*

A range of other examples can be found in Christopherson (1939) or Hawkins (1978), among others.

Proper names typically introduce new entities into the discourse, though they might also easily be used anaphorically, usually to re-introduce a referent into a more prominent position.

Indefinite noun phrases pose different problems. They too are used to introduce new entities into the discourse, and according to standard grammar books this is their typical role. For example, in (16):

(16) *Michael wanted to buy a new car.*

the indefinite noun phrase *a new car* does not refer back to any particular *car*, it simply introduces this concept into the discourse. But if the utterance in (16) were followed by two others, as in (17):

(17) a. *Michael wanted to buy a new car.*
b. *So last week he bought a new car.*
c. *Now he has a car that will last many years.*

it could be argued that the second and the third occurrences of the *car* (in (17a) and (17b), respectively) refer to the same entity, i.e. the *car* that Michael actually bought. In other words, the third occurrence of the indefinite noun phrase *a car* appears to be anaphoric.

More complex indefinite noun phrases can be used anaphorically. Consider this excerpt from an article in the *Wall Street Journal*:

(18) *...Blockbuster's recent purchase of a controlling interest in Spelling...*
(...) American Financial Corp. had held a 48.2% stake in Spelling that Blockbuster acquired last month in a $140 million stock swap.

It is likely that *a controlling interest in Spelling* and *a 48.2% stake in Spelling that Blockbuster acquired* refer to the same thing, yet the second NP is not a definite one.

It seems that when indefinite noun phrases are used "non-specifically", (as in example (16)) where the communicator hasn't got a specific referent in mind, they are not anaphoric (see Lyons 1977, p. 187ff and Brown & Yule 1983, p. 208-209 for an exploration of this view). The problem then lies in identifying which uses are 'specific' and which are not.

This generalisation does not seem to apply to the so called generic type of indefinite noun phrases, which too can be anaphoric. For example in (19):

(19) *I like books. Books are fun.*

the two references to some generic set of books can easily be seen as co-referential, though, neither could be taken to refer to a specific referent.

2.9 Anaphora from the NLE perspective

2.9.1 Anaphora resolution: a problem for NLE

Interpretation of utterances in general and the anaphora resolution process in particular appear to involve all kinds of knowledge: syntactic knowledge, semantic knowledge, knowledge of discourse rules and finally common sense and world knowledge as well the ability to use inference with the latter two.

Some types of knowledge can be made available to the natural language system with a lesser or greater degree of success, however some types, e.g. the common sense knowledge, prove particularly challenging.

It could be argued that in a natural language system such as LOLITA it will be fairly easy to make grammatical and syntactic information available to the anaphora resolution process. Likewise, a great deal of semantic information (e.g. a semantic hierarchy of concepts) can be provided. It seems, also, that certain aspects of discourse knowledge, to the extent that the latter is tractable, can be modelled and made use of.

Some levels of common sense or world knowledge and an advanced form of inference, on the other hand, would be very useful, but it is not expected that they can be used in this project (or at least, it won't be possible to rely on them).

The problem of designing an anaphora resolution algorithm can be viewed from the perspective of what sources of knowledge are needed for successful resolution and how that knowledge should be combined.

If all the necessary sources of knowledge were available to an NL system, and if that system were able to use them, the anaphora problem would be solved. The fact is, however, that not only are some kinds of knowledge presently inaccessible to an NL system, but also, we don't necessarily know how best to use the knowledge we do have. In view of this, the current project aims to achieve as much as possible with as much knowledge as is currently accessible in a system such as LOLITA and to assess how successful that is.

2.9.2 The importance of anaphora in NLE

Recent trends in natural language technology, particularly in information retrieval and information extraction, have indicated that the development of shallow techniques (usually based on pattern matching) may be good for only a limited number of applications. These methods have been shown to succeed on certain tasks, for example on the Named Entity recognition task (ARPA 1993, DARPA 1995). But the ability to successfully mark proper names of organisations, people and geographical locations, despite being useful for some purposes, seems to have only a

limited potential.

Resolving anaphoric expressions successfully could be seen as a vital step to many applications requiring a 'deeper understanding' of the text, e.g. extraction of some core information from the text in the form of templates, text summarisation, etc. The example in Fig 2.1 illustrates this point.

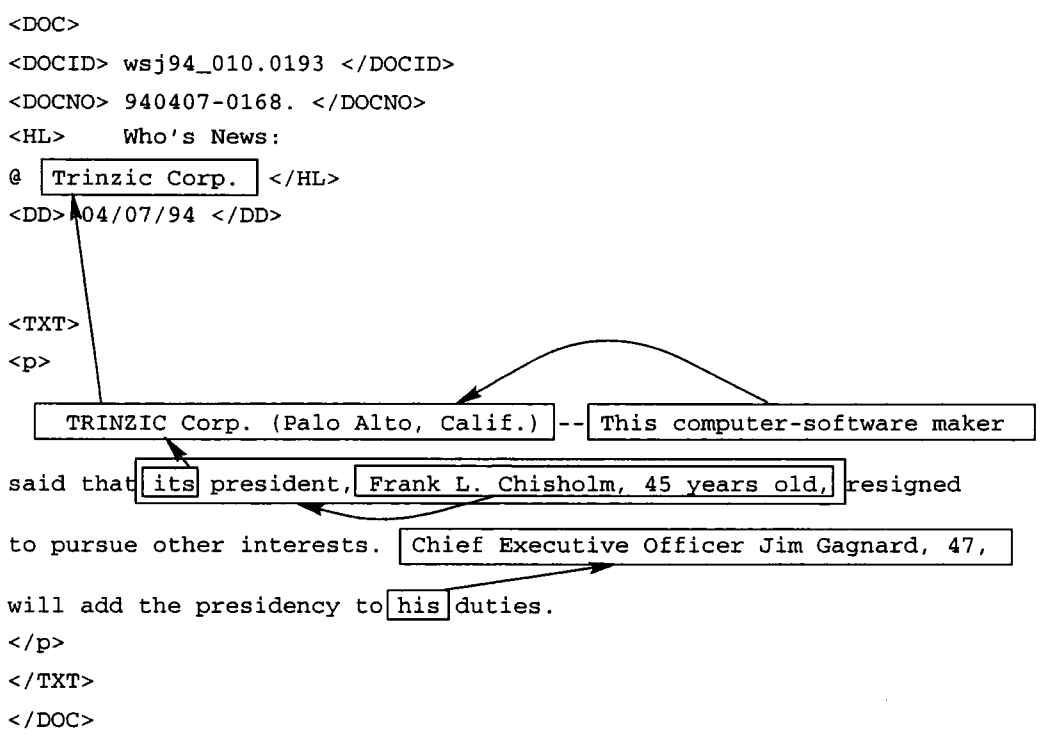


Figure 2.1: Example of a MUC article

The figure contains an article which constituted part of the training data used in the MUC6 evaluation (and which also belongs to the set of 15 training articles used in this project). The figure shows the co-reference links (marked with boxes and arrows) which are expressed in this text.

It could be argued that making the above connections is a necessary step to further analysis. For example, to discover that *Trinzic Corp* is a maker of computer software we need to make a co-reference link between *TRINZIC Corp* and the noun

phrase *This computer-software maker.* Similarly, to conclude that it was the president of *Trinzic Corp* who resigned, it is necessary first to connect the *its* with *Trinzic Corp*, as well as make a link between *Frank L. Chisholm* and the noun phrase *its president* and so on.

2.10 Scope of the project

2.10.1 Which types of anaphora?

The anaphora types that this project focuses on are, first of all, the core cases exemplified in section 2.4. The peripheral cases mentioned in 2.4 (as required by MUC), as well as the phenomena outlined in section 2.7.1 are also addressed.

However, anaphoric links such as those shown in section 2.6 are beyond the scope of the project. Likewise, the following, mentioned in section 2.4 are **not** included:

- anaphora with quantified or negated antecedents
- 'verb phrase' anaphora
- 'one' anaphora.

The aim is to account for all those co-reference links that are specified in the MUC6 co-reference task definition (see appendix A). Thus, the co-reference links involving two or more of the following expressions will be dealt with:

1. **full noun phrase anaphora**
    - definite and indefinite NPs
    - proper names (including proper name abbreviations and shortened versions of proper names);

2. **pronominal anaphora**
    - personal pronouns
    - possessive pronouns used as determiners
    - demonstrative pronouns;
3. **adverbials "here" and "there"**;
4. **certain time expressions**;
5. **certain currency expressions**.

2.10.2 Final analysis and sub-division of the problem

In order to design a new algorithm for resolving anaphoric expressions in LOLITA, the anaphora problem will be divided into the following main sub-problems:

1. **Recognizing the anaphoric expression**
    - how to distinguish anaphoric from non-anaphoric expressions
    - how to recognize that the existence of a specific entity is implied by the discourse
    - how to recognize non-referential uses of *it* and *there*.
2. **Choosing a set of possible antecedents**
    - where in the discourse can candidates for referents be found
    - which objects are ruled out as referents and which are allowed.
3. **Choosing the intended antecedent from among possible ones**
    - what factors influence which antecedent is chosen
    - how do these factors interact.

2.11 Criteria for success

The current project will be regarded as successful if at its conclusion a rise in scores in the evaluation tests is achieved.

Improvement in the interpretation of pronouns, proper names and noun phrases will be considered particularly important, the assumption being that the latter require the most general and least domain dependent analysis to work successfully.

By contrast, slightly less emphasis will be placed on the success in the area of "is_a" based co-references (cf. section 2.4), because in the majority of cases such co-references appear to rely most heavily on correct parsing rather than on a successful resolution algorithm.

Finally, the least emphasis will be placed on the performance on the 'function and value' co-references, as these are quite domain dependent and in many cases can be made on the basis of pattern matching.

Chapter 3

Background work

3.1 Introduction

This chapter reviews selected general work on anaphora. In recent decades, the phenomenon has been studied from a wide variety of perspectives and consequently literature on anaphora is vast.

As shown in chapter 2, many, often diverse, phenomena fall under the label of "anaphoric dependencies". Because this project concentrates on a subset of these, the current chapter will review only that work which seems relevant to the chosen subset.

A lot of researchers concentrate on very small areas of the problem and analyse those areas in great depth. Their insights might be useful, but it's impossible to take them all into account. In general, the aim here is, instead, to review work which would cover a large area of the core of the anaphora problem and which would be useful from the NLE perspective.

Approaches from different backgrounds deal with different aspects of the problem, often concentrating on different types of knowledge that might be involved in the

anaphora resolution process. In the following sections these approaches will be classified in terms of what knowledge they focus on or in terms of what factors are seen by them as important to anaphora resolution. Their relevance to the current project will also be considered.

3.2 Focus on syntactic factors

3.2.1 Generative Linguistics

Work in theoretical linguistics has concentrated on the description of the structures which license co-reference as compared to those which do not and, ultimately, on the search for the principles which underlie such structures. This work typically concentrates on problems exemplified by the sentences in (9a–d), p. 15.

3.2.1.1 Chomskyan generative grammar

The study of anaphora in theoretical generative linguistics is concerned mainly with the search for grammatical constraints on possible co-reference, usually within the boundaries of one sentence. A lot of interesting generalisations have been made by linguists regarding this issue and it would benefit an NL system to implement some of them.

For instance, the treatment of anaphora in Chomsky's Government-Binding (GB) framework (Chomsky 1986), extended and developed by several other researchers (Reinhart (1986), Grodzinsky & Reinhart (1993), Fiengo & May (1994)) predicts that certain types of co-reference cannot occur. To take a simple example:

(20) *Mary likes her.*

In this sentence 'her' and 'Mary' cannot be co-referential. In the GB analysis, this

restriction can be expressed in terms of the sentence's syntactic structure. According to this theory, the pronoun in (20) is found to be within the same 'governing category' as the name 'Mary', a situation defined to preclude co-reference.

This may be a trivial example, and it could be argued that co-reference could be ruled out on other grounds, for example the relative closeness of the pronoun and the antecedent, or the fact that if X is a subject of the sentence and Y is its direct object then the two cannot co-refer (unless Y is a reflexive pronoun, as in *Mary likes herself*). The example, is, however, illustrative.

Other, maybe less trivial cases, which clearly need to invoke grammatical principles are:

(21) a. *Paul was late for work because he got up late.*

b. *He was late for work because Paul got up late.*

In the above examples (modified after Hirst (1981)), *Paul* and *he* can be co-referential in (21a), but cannot in (21b). This is accounted for by one of the binding principles which states that a class of expressions to which proper names belong cannot co-refer with any preceding objects found in certain structural positions, within the same sentence. This structural position is usually described as a *c-commanding* position, where the relation of *c-command* is defined as follows (this definition is due to Reinhart (1976)):

> α c-commands β if and only if α does not dominate β and the first branching node dominating α also dominates β.

3.2.1.2 Relevance to the current project

The GB theory of syntax neatly predicts some co-referential links to be impossible, however it relies crucially on a complex analysis of the structure of the sentence.

This analysis differs from that of LOLITA's in many respects, hence it is difficult to implement insights from this theory straight-forwardly. However, this doesn't mean it would be impossible and a model of many of the phenomena handled by the GB theory does form a crucial part of the new algorithm implemented in this project.

The generalisations expressed by binding theory are useful to the extent that they are able to rule out some co-reference links or, with respect to reflexive pronouns, in the way they limit the search space within which an antecedent has to be found. However, the theory makes no predictions with regard to anaphoric dependencies which cross sentence boundaries. If anything, it shows that syntactic constraints do not hold across sentence boundaries, an important fact.

Also, the theory does not have anything to say about which antecedent should be chosen if more than one are present within the same sentence (cf. example (2), p. 12).

3.3 Focus on semantic factors

3.3.1 Preference Semantics

This framework, proposed by Wilks (1975) and later used and improved on by Boguraev (1979), assumed a 'decomposition' approach to meaning, partly inspired by generative semantics of Katz & Fodor (1963) and Katz (1971) popular at the time. It was postulated that 60 semantic primitives could be used to express "the semantic entities, states, qualities and actions about which humans speak and write" Wilks (1975)331. These primitives were used to make up formulae which represented the senses of English words.

For example, the formula expressing the meaning of the verb 'to drink' would consist of the primitives ANIMATE, FLOW, STUFF, THRU and MOVE and would

encode (in some specified way) the fact that the verb should be read as an action, which is performed *preferably* by animate entities, that the object of the action should be a liquid, etc.

The idea is related to the so called 'selectional restrictions' of Chomskyan generative grammar (1965) which postulates that each predicate imposes semantic restrictions on the choice of its arguments (cf. McCawley (1968), Jackendoff (1972)). However, in Wilks' system the concept of *restriction* is weakened to that of *preference*. The notion of a preferred rather than 'required' agent or object is important in the theory — the system of preferences does not aim to rule out cases where semantic restrictions are not met: if in a sentence the type of the agent (for example) does not match the preferred one, an interpretation can still be assigned, however it will be a metaphorical one.

The resolution of some types of anaphora may be possible (or at least aided) with this technique: candidates for pronoun assignment, for example, might be scanned and chosen as possible antecedents only when they satisfy the relevant preferences. For instance, given the two sentences:

(22) *Joe picked up the tea. He drank it.*

on the basis of semantic preferences alone, it would be possible to determine that the *it* in the second sentence can only refer to *the tea* and not, for example to *Joe*. This would follow from the semantic specification of *to drink*, which prefers a non-animate, liquid object. Similarly, the pronoun *He* is predicted to be *Joe* rather than anything else because *to drink* requires an animate subject.

The scheme would not, however, help us choose between referents which all happened to satisfy the semantic preferences, e.g.:

(23) *Joe picked up the tea which was standing near the coffee. He drank it.*

Here both *tea* and *coffee* are equally good candidates for a referent for *it*, however it is other knowledge (e.g. that one is likely to drink something one happens to be holding, having just picked it up, rather than drink something else) which helps us decide on the referent. That sort of knowledge is not encoded in the semantic knowledge of predicates.

While it could be argued that some co-reference links even in those limited cases (as in (22)) could be made on the basis of other knowledge (for example by the matching of the gender features inherent in the pronouns with those of the antecedents, which would bypass any information we have about the predicate), the knowledge employed by preference semantics is an important step towards modelling and employing knowledge of the world in the interpretation process.

Also, preference semantics can be seen to provide an account of some discourse phenomena about which the structure of the sentences (or grammar rules) don't have anything to say.

The criticism that could be made of this approach in general is that for the system to work reliably, a very rich database would have to be provided — one in which all concepts were broken down into a set of primitives. It is not clear how easy it would be to define such a set.

3.3.1.1 Relevance to the current project

For a large scale system, a knowledge representation scheme employing detailed breakdown of predicates in terms of semantic features might be difficult to achieve efficiently. Some researchers regard it an unrealistic prospect Jacobs & Rau (1993, p. 157).

However, despite its limitations, the central idea behind the preference semantics approach — the proposal that predicates impose restrictions/preferences onto their arguments — has a lot to offer and many (almost all, according to Allen (1994,

p. 320)), natural language systems, including LOLITA (cf. sections 5.1.1.4 and 5.1.6.1, in chapter 5), have employed it in some way. In most cases, what these systems have done is used this idea in conjunction with other methods, as clearly it cannot provide a satisfactory mechanism on its own (cf. for example Dyer 1983, Hirst 1987).

3.3.2 DRT and formal semantics

In the field of formal semantics work on anaphora often concentrated on sentences whose quantificational issues posed particular problems for pronoun resolution. Often, reseachers in this area would deal with examples such as (8a) and (8b), page 14.

Similarly, research into anaphora from the perspective of Discourse Representation Theory or DRT (Kamp 1981) focuses on the issues of representation of a subset of linguistic phenomena, often involving cases of complex quantification, and their relationship with pronominal reference. Typical examples that this theory aims to deal with include the so called bound-variable anaphora:

(24) *Every man thought he was ill.*

where the pronoun *he* could be seen as a *variable*, bound by the universally quantified phrase *every man.* The analysis of such cases would not pose problems, unless the pronoun were to be found outside the scope of the quantifier, for example in a separate clause or even a separate sentence:

(25) a. *Every farmer who owns a donkey beats it.*

b. *Each girl was very harshly treated. She had to be up by 5am and do all the housework before she could go to school.*

(Example (25b) is due to Kempson (1990)). The objective of DRT is to capture

main properties of both inter- and intra-sentential anaphora within one, unified theory. This involves postulating a certain type of representation which is capable of handling both the generalisations to do with anaphora within a single sentence as well as those that span a larger discourse.

3.3.2.1 Relevance to the current project

The theory doesn't offer any ideas as to how to choose between possible antecedents, when more than one is available, and from this point of view, it is not useful for the current project. On the other hand, its insights could be very useful in the future, when more effort can be spent developing the way LOLITA handles quantified antecedents.

3.4 Focus on discourse factors

3.4.1 AI approaches

3.4.1.1 Early algorithms

After initial pessimistic conclusions by Charniak (1972) that the resolution of anaphoric expressions requires an arbitrarily detailed world knowledge, researchers in AI began to look for more attainable solutions.

More work on anaphora was carried out by Hobbs (1979), Webber (1979) and also Hirst (1981), who provided an overview of solutions proposed up to that time.

Hobbs suggested an algorithm of pronoun resolution within sentence boundaries, roughly based on the principles which were pre-cursors to a fuller treatment of pronouns within the government-binding framework. Any pronoun not resolved by those principles was assigned to the most recent antecedent (stored in some

representation or model of the discourse) whose semantic features didn't clash. Webber (1979) further developed the notion of the discourse model and tried to define the way potential antecedents could enter it.

Hobbs' syntactic algorithm, despite being labelled 'naive' (by Hobbs himself), has been shown to perform well on some texts. For example, Lappin & Leass (1994) report its success rate to be 82% for their test set involving 360 pronouns occurring in sentences taken from computer manuals.[1]

In the late 70s and early 80s the notions of *discourse focus* entered the scene since it was noted (in agreement with intuition) that resolving an anaphoric expression often involves choosing an entity which is the most salient within the reader's or hearer's focus of attention. Grosz (1977) and Sidner (1981, 1983) provided the early attempts to model focus/salience of referents in a discourse.

By far the most influential work which subsequently developed is that of Grosz & Sidner (1986) and Grosz *et al.* (1995).

3.4.1.2 Grosz and Sidner: structural approach

Grosz and Sidner propose that discourse is organised along three distinct but inter-related dimensions: linguistic, intentional and attentional.

The linguistic level is the sequence of utterances which constitute the discourse.

The intentional level divides discourse into segments according to the purposes (or intentions) of the discourse participants in any particular exchange (so that for example a new discourse segment begins when the purpose of the participants changes). The segments are related to one another via a hierarchy: one segment can dominate another if the purpose behind the dominating segment somehow includes the purpose of the dominated one.

[1]The success rate expresses the number of all anaphoric referents which are correctly resolved out of all possible anaphoric referents.

The attentional level is a data structure in the form of a stack which collects entities and events that are being talked about in each segment. For each segment a separate slot is built. If the new segment is dominated by the previous one at the intentional level, its entities and events will be pushed on the focus stack on top of the slot derived from the previous segment. If, however, the new segment is a subordinate of the previous one, the slot belonging to the previous segment will be popped from the stack before the new one is pushed onto it. This system provides a mechanism whereby objects become either more or less accessible to pronominal (or other) reference as the discourse progresses. Salience of objects is thus expressed in terms of accessibility from the focus stack.

Definite noun phrase interpretation is said to use the focus stack to search for possible referents. However, other factors, such as inference using world knowledge (not specified in great detail) are also said to be involved. Pronoun interpretation, on the other hand, as well as using the focus stack, is argued to warrant an augmented mechanism based on the centering theory (Grosz *et al.* 1995, Gordon *et al.* 1993), which is discussed in section 3.4.1.3, below.

The relationship between discourse segments depends directly on the purposes of the participants in that particular segment. Since it is this relationship that crucially affects the contents (ordering of slots) on the focus stack, it follows that anaphora resolution will depend first of all on the success of recognising the intention behind each discourse segment and then on recognising how that intention fits into the existing hierarchy of intentions. This aspect of the theory is not fully developed by Grosz & Sidner (1986), however other researchers have pursued it independently, often as part of the planning paradigm (e.g. Cohen *et al.* 1990). Grosz and Sidner merely propose that heuristics based on cues from the linguistic level of the discourse could be used to determine at least the boundaries between segments, while general knowledge of actions and objects will help participants determine the relationships between them.

3.4.1.3 Centering

As has been mentioned above, a separate theory, 'the centering theory', is posited to account for pronominalisation in discourse. The theory proposes that:

> each utterance in a coherent discourse segment contains a single semantic entity — the backward-looking centre — that provides a link to the previous utterance, and an ordered set of entities — the forward-looking centres — that offer potential links to the next utterance (Gordon *et al.* 1993).

The theory itself has little to say about the ranking of the individual elements in the 'forward-looking centres'. It merely suggest that several factors, such as subject-hood, role in a sentence, being the backward-looking centre might affect this ranking.

The centering theory further proposes rules which are said to characterise coherent discourse. The most interesting one of these (from the anaphora problem point of view) is the rule which stipulates that:

> if in a given utterance a reference is made to one or more elements of the previous utterance's forward-looking centre, then the most highly ranked one of these will be the 'backward-looking centre' of the current utterance and it expected that it will be realized as (referred to by) a pronoun.

This rule should not be confused with a much stronger one which might state that *the most highly ranked element of a forward-looking centre will be referred to by a pronoun in the following utterance*, even though such a situation may occur frequently in a coherent discourse.

The rule predicts that, for example (26a) will seem less natural than (26b):

(26) a. *It was Susan who helped Betsy with the physics homework.*
She knew that Susan was good at science.

b. *It was Susan who helped Betsy with the physics homework.*
She knew that Betsy was poor at science.

The first part of these two examples places high stress on Susan by using the 'cleft' construction ('it is X who....') thereby arguably making Susan the most highly ranked element of the forward-looking centre. When the second utterance refers to Susan by name rather than a pronoun in (26a) this is counter to expectation and doesn't read as well as (26b), in which the reference is made via a pronoun.

Gordon *et al.* (1993) have some experimental (psychological) data to support this rule. The results of a reading-time task experiment reported by them suggest that passages whose utterances do not obey the rule (such as those in 26a) were read significantly more slowly than passages which do.

Other psychological experiments reported by Gordon *et al.* (1993) further suggest that the subject of an utterance (regardless of surface position) is the preferred site for the backward-looking centre.

Grosz and Sidner's theory predicts that pronoun interpretation and the interpretation of anaphoric noun phrases differ. In their model of discourse structure pronouns provide links between adjacent utterances, while noun phrases are expected to provide links between non adjacent utterances (or even non adjacent segments). The focusing mechanism of the 'focus stack' which in general provides referents to noun phrases has a more global role to play than the more local focusing mechanism due to centering.

This is an interesting empirical claim, which could be tested by designing an algorithm for anaphoric resolution that takes account of it. The algorithm would have to distinguish between different types of focus and keep track of the different fo-

cus level for each possible referent and use this information in making co-reference links.

A major drawback of both the discourse structure theory and the centering theory is that they make anaphora resolution depend on other, largely unsolved, problems in NLP, namely the recognition of intentions and the identification of factors involved in making some objects in a discourse more focused than others.

The first (and, to some extent, the second) problem could be avoided by using a focus structure which is independent of the intentional level of the discourse. Such an approach has been suggested by Alshawi (1987).

The principles of centering theory have been used as a basis for a pronoun resolution algorithm proposed by Brennan *et al.* (1987) and Walker *et al.* (1994). The algorithm described by Brennan *et al.* (1987) offers an interesting account of reader preferences in pronoun interpretation (see Kehler (1998) for discussion). A way of ranking the elements of the forward-looking centre is also suggested (this aspect was missing from the centering theory itself). The authors do not, however, offer an evaluation of their algorithm on more than a very small set of examples.

3.4.1.4 Alshawi's salience model

Alshawi proposes a model which assigns salience values to discourse referents according to the following factors: being a major syntactic constituent, being a subject phrase, being a nested term, being a 'relation'. For example, if an entity is the subject of the utterance it will get marked relatively highly — since the marks will come from both the first and the second of the factors listed here. During the processing of the subsequent utterance the marks are revised according to a defined 'decaying' function and according to whether the objects are mentioned again or not. The latter case results in marks being awarded as above and the former in the new marks being added to the existing ones. The objects are in focus (and available as antecedents) if their ranking is greater than 0.

This model has recently been implemented by Huls *et al.* (1995) as part of their EDWARD system (in Dutch). They report to have tested the implementation on 125 sentences containing anaphoric expression such as pronouns and noun phrases used in simple dialogues with the system. All of these dependencies were resolved correctly.

They also compared this model with another, more naive one, in which co-reference links between an anaphoric expression and the most recent, semantically compatible antecedent is made. This method resulted in the correct resolution of 119 cases.

It must be noted that most of the texts used contained simple anaphoric links and only one sentence used more than one pronoun. Interpreted in this light, these results are perhaps not as impressive as it would first appear.

Despite that fact, it can still be said that Alshawi's approach has an important advantage over that of Grosz and Sidner's in that it does not depend on the theory of intentions. However, it does depend on the theory of factors influencing salience. In the absence of such a theory, the model makes assumptions about these factors in a way which results in correct referent resolution.

3.4.1.5 Relevance to the current project

While the account of global focus in the discourse structure model is close to providing an actual algorithm for interpreting noun phrases, the centering theory does not provide an actual algorithm for interpreting pronominal reference. However, by postulating 'backward-' and 'forward-looking' centres and making claims about discourse coherence in terms of the contents of these centres it allows us to see what can be expected in a coherent discourse. Based on such expectations, one can then try to produce heuristics for interpretation.

What happens though, if the expectations are not met? For example the passage in (26a) deviated from the predicted form, yet it hasn't become uninterpretable as

as result. Possibly, separate rules and heuristics would be needed to account for such cases.

One interesting claim made by the centering theory which could be exploited in a practical implementation, is that if indeed the most salient referent from one utterance is realized as a pronoun in the following utterance, the referent is likely to be very salient in the following utterance. This idea could easily be employed in a mechanism which tracks saliency.

The dependence of the discourse model on an intentional level would pose difficulties in the system such as LOLITA, because LOLITA does not as yet have a mechanism for the recognition of intentions. From this point of view, a solution such as Alshawi's (assuming that it could perform well on a larger corpus) would be more advantageous.

3.4.2 Psycholinguistics

Psycholinguists have traditionally concentrated on the study of mental processes, including those involved in anaphora comprehension. Their search is for rules and heuristics which help human readers and listeners assign referents to anaphoric expressions. Typically, they would deal with sentences such as example (2), p. 12. The following sections briefly review the most important aspects of the anaphora resolution processes that psycholinguists have been investigating.

3.4.2.1 Local salience of discourse referents

Some psycholinguistic studies have attempted to discover how certain properties of discourse affect the reader's focus of attention during text comprehension. For example, Gernsbacher & Hargreaves (1988) have suggested, after a series of experiments, that the first mentioned entity in a discourse is much more prominent in the reader's memory than any other entities of the discourse. This effect is said to

be independent of any syntactic factors (such as, for example, subject-hood).

Gernsbacher *et al.* (1989) further investigated the interaction between the first-mention effect and the so called 'clause recency effect', whereby a discourse entity mentioned in a current clause is more accessible to the reader than the one mentioned in a previous clause. Gernsbacher *et al.* conducted experiments involving two-clause discourses, with each clause containing a referent. The results indicate that the clause recency effect holds while the reader is processing the second clause. However, when the information from the two clauses is integrated, the first mention effect returns and makes the first mentioned entity more accessible again.

Other experiments have shown that, in general, recency of mention (so not only 'clause recency') features highly as a factor contributing to the salience ranking of the referent in the reader's memory (Clarke & Sengul 1979).

Some experiments on pronoun comprehension have looked at how factors such as the thematic role of the antecedent (for example, whether it's a goal or a source of the action, agent or recipient etc.) affects its accessibility for reference. For example, in sentences such as:

(27) a. *Paul(Source) passed the comic to Bill(Goal).*
b. *Bill(Goal) grabbed the comic from Paul(Source).*

As the 'Goal', *Bill* tends to appear as the more salient entity, regardless of its surface position (Stevenson *et al.* 1994). Experiments indicate that other thematic roles follow similar patterns.

Other studies have supported the intuition that subjects are often more highly ranked than other constituents in the sentence (Sanford & Garrod 1981).

This last finding appears to be contradictory with respect to the immediately previous one. This may reveal weaknesses behind the findings and the psycholinguistic

experimentation behind it. On the other hand, it might point to the enormous complexity involved in the processes of anaphora resolution, where apparently contradictory tendencies come into play, yet it is not known exactly how the whole system works.

3.4.2.2 The notion of global topic

To date, many experiments have been conducted by psychologists and psycholinguists testing the effects of global topic on discourse comprehension. The notion of global topic could be informally described as the situation that the discourse is about.

Experiments consistently show that if a reference is made in the discourse to the topic or to individuals closely associated with the topic, human readers find comprehension of the discourse easier (reading times are facilitated in such cases). The experiments by Anderson *et al.* (1983), McKoon *et al.* (1993), or Marslen-Wilson *et al.* (1993), are among many conducted in this area.

Sanford *et al.* (1988) suggest further that the global topic is very important in the comprehension of definite descriptions and proper names. This contrasts with the effect of local salience, which does not seem to affect the interpretation of definite descriptions and proper names to the same extent as it affects the interpretation of pronouns.

3.4.2.3 Other factors in comprehension

One of the more important findings of Garnham *et al.* (1992) as well as Stevenson & Urbanowicz (1995) is that pragmatic factors, such as inferences based on world knowledge, are of most influence in the processes of anaphora comprehension and they tend to override any heuristics that might otherwise be used (possibly with the exception of those involving gender cues). This is consistent with the obser-

vation of the ease with which human subjects interpret sentences like the example (2), page 12. However, such inferences are among the most difficult aspects of comprehension to model in a computer program.

There is also some experimental evidence to suggest that human subjects appear to assign an ambiguous pronoun to a preceding noun phrase with the same grammatical function. This is sometimes called the "parallel function" strategy (Caramazza & Gupta 1979, Stevenson *et al.* 1995).

Other experiments provide evidence that implicit causality of a verb tends to focus on the entity which is associated with the causes, rather than the results, of the action described in a sentence (Garnham *et al.* 1992). This effect can be compounded or modified by the use of discourse connectives such as *because* and *so*, which are sometimes seen as devices that direct the hearer's attention towards the causes or consequences of events in the discourse (Stevenson *et al.* 1994).

3.4.2.4 Relevance to the current project

A lot of the insights gained from psycholinguistics research can be usefully employed in a system such as LOLITA. Indeed, this project has attempted to take account of some of them (more details follow in chapter 10). In particular, any ideas regarding what guides the human subject in their choice of antecedent can be used as inspiration for reference resolution heuristics.

However, two facts must be borne in mind. Firstly, research in psycholinguistics concentrates on human comprehension. Furthermore the experiments usually involve artificially constructed materials which exaggerate the particular feature or factor studied. When attempting to build a computer system to process texts such as newspaper articles, it cannot be automatically assumed that those particular features or factors will be present. And if they are, they may not be easy to track down. It is not known how the factors in a real-life discourse interact or how exactly their interplay affects comprehension.

Secondly, even if human readers find some discourse factors helpful and facilitating comprehension, they are still capable of understanding examples of discourse in the absence of such facilitating factors. This is because their use of knowledge of the world and inference appears to over-ride (or possibly even obviate) the need for facilitating factors.

Overall, the meta assumption behind drawing upon the results of psycholinguistic research in this project is that given that human discourse is (generally) easy to comprehend, we assume that it must contain at least some of the appropriate cues. We can therefore tune our computer program to be sensitive to those cues.

3.5 Conclusions from literature review

For the purposes of the current project, the most worthwhile points arising from the literature review could be summarised as follows:

- Inter-sentential co-reference links are not governed by rules which bar some co-references while allowing others. Their resolution appears to be governed by (what can be vaguely termed) discourse rules. The latter cannot be as neatly defined as the rules for intra-sentential anaphora.
- However, the crucial notion that is used in inter-sentential anaphora resolution is salience. All the models for anaphora resolution discussed differ in the way they define and use the notion of salience.
- On the other hand, the use of some aspects of world knowledge, or the ability to rely on advanced forms of inference (when the latter could provide the most convincing information needed for anaphora resolution) do not appear to be possible in the current state of the art.
- Psycholinguistic research offers several interesting ideas for the design of resolution heuristics. Some of those would be easier to model in LOLITA than others.

Chapter 4

Review of Co-reference Resolution Systems

4.1 Introduction

This chapter reviews existing systems which perform a task similar to the task described in this project. The review includes mainly other systems which participated in the MUC6 evaluation. The systems can be classified in terms of the general method they use to perform the task. For example, some are rule-based, others are based on machine learning. Some can be seen as hybrids of the two.

The final section of the chapter also examines the RAP algorithm proposed by Lappin & Leass (1994).

4.2 Systems employing symbolic techniques

This section reviews the systems which are rule based and whose rules are 'manually engineered'.

4.2.1 The PIE system

The PIE system was developed at the University of Manitoba and was used during MUC6 to perform all the tasks of the competition (Lin 1995). Its main component is the parser ("PRINCIPAR"), based on GB theory of syntax. There is no attempt to build a full semantic representation of the processed text. A series of pattern matching rules operates directly on the parser output to perform (or provide material for) the MUC tasks.

The co-reference task is carried out by building chains out of noun phrases found in the parser output and matched according to a set of 'compatibility' rules. Four types of rules are used: semantic compatibility rules (with information derived from WordNet (Miller 1990)), rules based on binding theory, heuristics inspired by centering theory and finally string matching rules.

The compatibility rules are applied to pairs of noun phrases in a chain and result in a series of equality (or inequality) assertions about the pairs. Different rules give different weights to their assertions. Special rules combining all the assertions are applied at the end of the process to result in the final co-reference chains.

Some of the rules used by the system appear to leave room for improvement. For instance, the implementation of the binding theory seems too strong — it rules out co-reference between a pronoun and a c-commanding noun phrase found in the same clause, yet outside the local domain Lin (1995, p. 119). So, in an example like this one:

(28) *John likes his mother.*

the co-reference between *John* and *his* is not allowed, contrary to intuition.

Also, the system doesn't seem to take into account gender information and co-reference links between pronouns of one gender with proper names of another seem to be allowed, for instance, in:

(29) *John likes Susan. She is happy.*

She is resolved to be *John*. The PIE system is available on the World Wide Web and the above examples were submitted to it soon after the MUC6 evaluation. The co-reference results were then as reported here; however, the system has been updated since that time and now (at the time of writing), the example in 28 does allow co-reference, while 29 remains erroneous.

Despite the above shortcomings, this system was one of the best at MUC6. Its scores were: **Recall 63%; Precision 63%.**[1] The overall approach to the task is interesting, as it attempts to combine several different factors and sources of knowledge in order to discover what co-reference links are expressed in a text.

4.2.2 The LaSIE system

The LaSIE system was developed at the University of Sheffield. The system is designed to parse the input text, analyse it semantically and then build a discourse model to represent it (Gaizauskas *et al.* 1995). A *world model* of ontological classes and their properties is used in the building of the discourse model. The model then provides input for the MUC tasks. A novel aspect of this system is that it does not use a purpose built lexicon. The lexical information used for parsing is computed during processing, through part of speech tagging and morphological analysis.

This system is the only system in MUC6 competition which is similar to LOLITA, in that one of its stated aims is to use all possible levels of linguistic analysis, including a full parse, to aid with the co-reference task. It is also similar in that it builds a representation of the text which it can then use in a way suitable to any given application.

Co-references are made between concepts in the model, as the input is being pro-

[1] For the definition of the terms *recall* and *precision* and the related concept of *F-measure* see chapter 8, section 8.1.4, pp. 118ff.

cessed, subject to agreement in properties and ontological class. Specific heuristics for string matching are used for co-reference between proper names. Other, co-reference specific rules, include the following:

- proper names to enter into co-reference over the whole
- pronouns can only refer back within the current paragraph (except for paragraph initial pronouns, which can refer back to the previous paragraph)
- all other noun phrases are allowed to refer back over a span of two paragraphs.

It would be interesting to know whether the system employs parsing information to implement grammatical constraints for intra-sentential co-reference, however no details about this issue are provided.

The official score of the LaSIE system was **Recall 51%** and **Precision 71%**. However, due to a trivial error this score is for the output on 29 out of the 30 texts used in the test. After the correction of the error, the system's score changes to **Recall 54%** and **Precision 70%**. This constitutes a rise in F-measure from 59.4 to 61.

4.2.3 The New York system

The New York System entered at MUC6 was based primarily on pattern matching. In this system, the input text is first tokenized and then tokens from each sentence are looked up in a fairly comprehensive set of databases (including a broad-coverage English dictionary, a gazetteer of geographical names and a company dictionary, among others).

After the dictionary lookup, there are four stages of pattern matching: one for names, one for noun groups, one for verb groups and finally one for some semantically defined patterns, specific to the scenario present in the MUC texts. At the

end of the four stages, a logical representation of the sentence, consisting of entities and events (and relations between them) is built.

Co-reference resolution examines each entity and event, and tries to find a match in the logical forms which were built for the preceding text. Special matching rules exist for definite noun phrase matching and for proper names matching. Indefinite noun phrases don't trigger a search for an antecedent.

The score of this system on the co-reference task was **Recall 53%** and **Precision 62%**.

4.2.4 The University of Pennsylvania system

The co-reference resolution system from the University of Pennsylvania performs the task in three stages, each one dealing with a different type of anaphoric expression (Baldwin *et al.* 1995). Before the actual co-reference resolution comes into play, other processing is employed: tokenization, part of speech tagging, noun phrase detection, non-referential *it* identification and Named Entity recognition.

The first of the co-reference resolution processes aims to link up into appropriate chains the proper names which were identified at the Named Entity recognition phase. Rules which match abbreviations and shortened names with full proper names are used at this stage. A database of abbreviations as well as a database of geographical names are also drawn upon.

The second stage employs a parser whose output is searched for syntactic patterns expressing 'is_a' based co-references (e.g. *John Smith, president of ACME*) and various 'function and value' type co-references (e.g. *$53 or 20 cents a share*).

The third stage deals with definite noun phrase and pronominal anaphora. It employs an extension of the CogNIAC system (Baldwin, 1995) built at the University of Pennsylvania specifically for pronoun resolution. CogNIAC itself is reported to

make co-reference links with high precision. It is conservative in the sense that if, according to the system's rules, more than one referent qualifies as an antecedent for a pronoun, the pronoun is left unresolved.

The rules for pronoun resolution are as follows (Baldwin 1997):

- if there is a single possible antecedent in the read-in portion of the discourse, then pick it as the antecedent
- pick the nearest possible antecedent in the read-in portion of the discourse if the anaphor is a reflexive pronoun
- if there is a single possible antecedent in the prior sentence and the read-in portion of the current sentence, then pick it as the antecedent
- if the anaphor is a possessive, and there is a single exact string match for the possessive in the prior sentence, then pick it as the antecedent
- if there is a single possible antecedent in the read in portion of the current sentence, then pick it as the antecedent
- if the subject of the prior sentence contains a single possible antecedent and the anaphor is the subject of its sentence then pick it as the antecedent.

A possible antecedent for the pronoun is one which matches the pronoun in gender and number and does not violate co-reference restrictions derived from centering theory.

The algorithm has been tested on a set of training data consisting of narrative texts. The test contained 198 pronouns which were resolved with **Precision** of **97%**. The **Recall** was around **60%**.

On a blind set of narrative texts (containing nearly 300 pronouns), the algorithm performed at the level of **92% Precision** and **64% Recall**.

The authors of the Pennsylvania University MUC6 system have extended CogNIAC to deal with definite noun phrases. They report that the noun phrase resolution system is much more 'eager' than the pronoun component and that it posits too many co-reference links. Various heuristics are used to prevent this from happening.

The overall score of the system at MUC6 was: **Recall 55%; Precision 63%**. However, the score is reported to have been affected by formatting errors. Once those are removed, the score goes up to: **Recall 63%; Precision 72%**. This makes it the highest scoring system at the competition.

A detailed analysis of the system's performance on the first half of the official evaluation is said to reveal that three quarters of all errors were due to factors independent from CogNIAC itself. However, it is also reported that some of the errors where due to problems with the system's rules. For example, the rule which chooses a subject of the prior sentence as a referent for an anaphor in the subject position doesn't work when indirect speech is involved. Further errors were due to the fact that gender information for both anaphors and antecedents was not always reliable.

The CogNIAC system is interesting because it provides an insight into the sort of numbers of pronouns that can be resolved when there is no ambiguity in the vicinity of the pronoun. Its precision figures for pronouns are impressive.

The success of this system (in terms of its precision) shows that it's possible to design highly successful rules for a subset of anaphora resolution. The problem of the system is its relatively low recall and not so precise resolution of other types of expressions.

4.2.5 The FASTUS system

The FASTUS system was developed at SRI International and is described as a series of finite state transducers each providing a separate level of analysis of the input

text (Appelt *et al.* 1995). The initial phases of the system are domain independent. The functions they perform are to tokenize the text, to recognize fixed phrases (such as 'because of') and to recognize Named Entities. Then some form of island parsing is performed which finds constructs that can be unambiguously recognized as 'noun groups' and 'verb groups'. This is followed by another stage, in which phrases are combined into larger units to the extent that the phrases' content allows them to be combined fairly unambiguously.

The final stage of the system depends on the domain being processed and is designed to recognize the particular structures that are likely to contain information that the system is meant to extract.

A co-reference resolution module designed to identify co-references between individuals is implemented, however it is not described in detail. It is said to be rule-based, using 'simple algorithms'. Different types of anaphoric expression are said to use slightly different rules.

The FASTUS system achieved **Recall of 59%** and **Precision of 72%**.

4.3 Systems using statistical techniques

4.3.1 RESOLVE

RESOLVE is a co-reference component of a system built by the University of Massachusetts at Amherst (Fisher *et al.* 1995), and further developed by McCarthy (1996). It is fully trained on data, without the use of manual engineering.

The only manual input is in the initial stages of the building of the system. It involves the preparation of a set of labels which classify noun phrases in terms of features, such as being a name, being a pronoun, being the subject of a sentence, being the most recent compatible subject for an anaphor, etc. This information is

(presumably) added to a corpus of texts where co-reference relations are annotated. Such a corpus is then used to train the system.

At the outset, RESOLVE's designers decided to only deal with references to people and organizations, which according to the their count, constituted 66% of the training corpus. This introduces an a priori limitation to the system's performance.

To perform the co-reference task RESOLVE receives as input a text which is tagged with part of speech information and in which the Named Entities have been identified. It examines pairs of noun phrases using 'decision trees' created on the basis of the training texts during the system's learning stage.

The system's score at the MUC6 evaluation was **Recall 44%; Precision 51%**. However, given that by design the system was only aiming at the two thirds of all possible co-reference links, its 'real' Recall measure has been estimated to be around **67%**.

One of the weakness of the system is the limited number of features that are used to classify anaphoric expressions for the purposes of training.

The strength of the system might be its ability to 'discover' rules for resolution. An example of such a rule is given:

> anytime we had two references to the same type of object, neither is a pronoun, the second phrase is not a proper name, both are in the same sentence and the first phrase is a proper name, the two references are classified as co-referent.

This may be an interesting rule, however, it appears to overgeneralise. For example, for a sentence:

(30) *John looked at the man.*

the rule would predict that *John* and *the man* co-refer, counter to intuition (and counter to a strong grammatical constraint).

From a different point of view, though, the rule could be seen as very interesting, because it indicates that the kind of discourse RESOLVE was trained on (i.e. the MUC6 training corpus) does not necessarily always obey the grammatical constraints.

4.4 Other reference resolution algorithms

4.4.1 The RAP algorithm by Lappin and Leass

The system described below focuses on just a subset of the anaphora problem. That is, it deals with just pronoun resolution, while ignoring other types of co-references. However, some aspects of its overall approach are similar to what is proposed in chapter 10 of this thesis. For this reason it is included here.

Lappin & Leass (1994) present an algorithm called RAP (Resolution of Anaphora Procedure) for resolving pronominal anaphora using heuristics based on syntactic structure and on a model of salience. The algorithm handles personal pronouns, possessive determiners, reflexives and reciprocals.

It rules out some anaphor-antecedent pairings on the basis of grammatical rules in the spirit of binding theory, though the latter is not explicitly mentioned (possibly because the syntactic framework of their approach is different). Antecedents for intra-sentential reflexives and reciprocals are assigned first on the basis of syntax alone. If this procedure still leaves ambiguity, the anaphoric expression is resolved later, together with other pronouns.

Lappin & Leass propose a number of salience factors, each of which assigns a weight to a possible antecedent. The factors are the following:

1. sentence recency
2. subject emphasis
3. existential emphasis
4. accusative emphasis
5. indirect object and oblique complement emphasis
6. head noun emphasis
7. non-adverbial emphasis.

Thus, when processing a sentence, all entities introduced by the sentence receive 100 points by way of factor (1). At the same time, all the entities kept on from the previous sentence have their weights halved. Factors (2), (4) and (5) award saliency weights of 80, 50 and 40 points respectively, in accordance with a thematic role hierarchy, where subjects are seen as inherently more salient than direct objects (in 'accusative' case), while direct objects are seen as inherently more salient than indirect objects. The saliency hierarchy is said to originate from Keenan & Comrie (1977) and be similar to Johnson (1977).

Heuristic (3) rewards phrases following existential constructions such as *There is* X, where X is the rewarded entity.

Heuristic (6) favours any noun phrase which is not a constituent of another noun phrase.

Finally, the non-adverbial emphasis heuristic awards points to any noun phrase which is NOT a constituent of an adverbial prepositional phrase.

Apart from the emphasis weights, several other rules are also used during resolution. These include the following:

- gender and number compatibility must be taken into consideration
- if the pronoun and the antecedent occupy the same syntactic role, give this antecedent an extra 35 points ('parallel role reward')
- if the pronoun occurs before the antecedent, the antecedent is penalized by having 175 points taken away ('cataphora penalty')
- if, at the end, there is more than one candidate with the same weights, choose the one which is closer to the anaphor (in terms of word order in the sentence).

The input to the RAP system is first processed by a parser. A Prolog implementation of RAP then produces antecedent-anaphor pairings.

The system was developed and tested on sentence pairs selected from computer manuals. The pairs were chosen in such a way that the first sentence contained no pronouns and the second contained at least one third person pronoun (including reflexives and reciprocals). The total number of pronouns to resolve was 560. The system correctly resolved 85% of them. A blind test was also carried out on a similar set of sentence, this time containing 360 pronouns. The success rate was 85%. In both test sets, the system score was higher on intra-sentential anaphora than on inter-sentential anaphora.

The method described by Lappin & Leass has recently been employed by Kennedy & Boguraev (1996) in their discourse processing system. However, unlike Lappin & Leass, Kennedy & Boguraev do not rely on a full parse as input for anaphora resolution. In the authors' view current parsing technology is not reliable enough. Instead, they attempt to utilize only part of speech tagging for providing the input to the resolution system.

This system is interesting for two reasons. It tries to build a logical representation of the processed text, which is a similar aim to that of LOLITA. Secondly, it has been more thoroughly evaluated on a variety of texts, which included several types of discourse (web pages, magazine articles, new stories). Furthermore, the system

appears to make co-reference links between other types of expressions (i.e. Named Entities) and not only resolves pronominal references.

However, in order to compare their system with that of Lappin and Leass, Kennedy & Boguraev (1996) only report the performance on pronominal resolutions. In their evaluation set there were 306 pronouns to be resolved and 231 (75%) of them were resolved correctly.

Unfortunately, Kennedy & Boguraev do not provide any more details of their system.

Chapter 5

An outline of the LOLITA system

This chapter describes the architecture of the LOLITA system and provides a general description of its main components. Those components which are particularly relevant to the current project (i.e. to the co-reference task) are only outlined here, and are dealt with, in greater detail, in chapter 6.

LOLITA is designed as a general purpose natural language processing system. Its aim is to provide a *core* of natural language capabilities upon which many different applications can be built. The system has been under development at the University of Durham since 1986 and work on improving it still continues.

The general architecture of the system and the core components were designed and implemented by Prof. Garigliano, the head of the LOLITA Group, with contributions from Morgan, Baring-Gould, Smith and Callaghan (see, e.g. Garigliano *et al.* 1992, Long & Garigliano 1994, Morgan *et al.* 1995).

Several prototype applications have also been built to-date, including:

- information extraction applications: summary templates (Garigliano *et al.*, 1993) and user defined templates in the financial domain (Costantino, 1997)
- a dialogue system (Jones, 1994)

- a Chinese tutor (Wang, 1994)
- an English-Italian translator (Morgan *et al.*, 1994)

5.1 LOLITA's core components

The diagram in Fig. 5.1 illustrates the architecture of LOLITA, which in general follows the traditional 'pipeline' model. At the heart of the system lies its knowledge base—the Semantic Network, or *SemNet* (Baring-Gould, forthcoming).

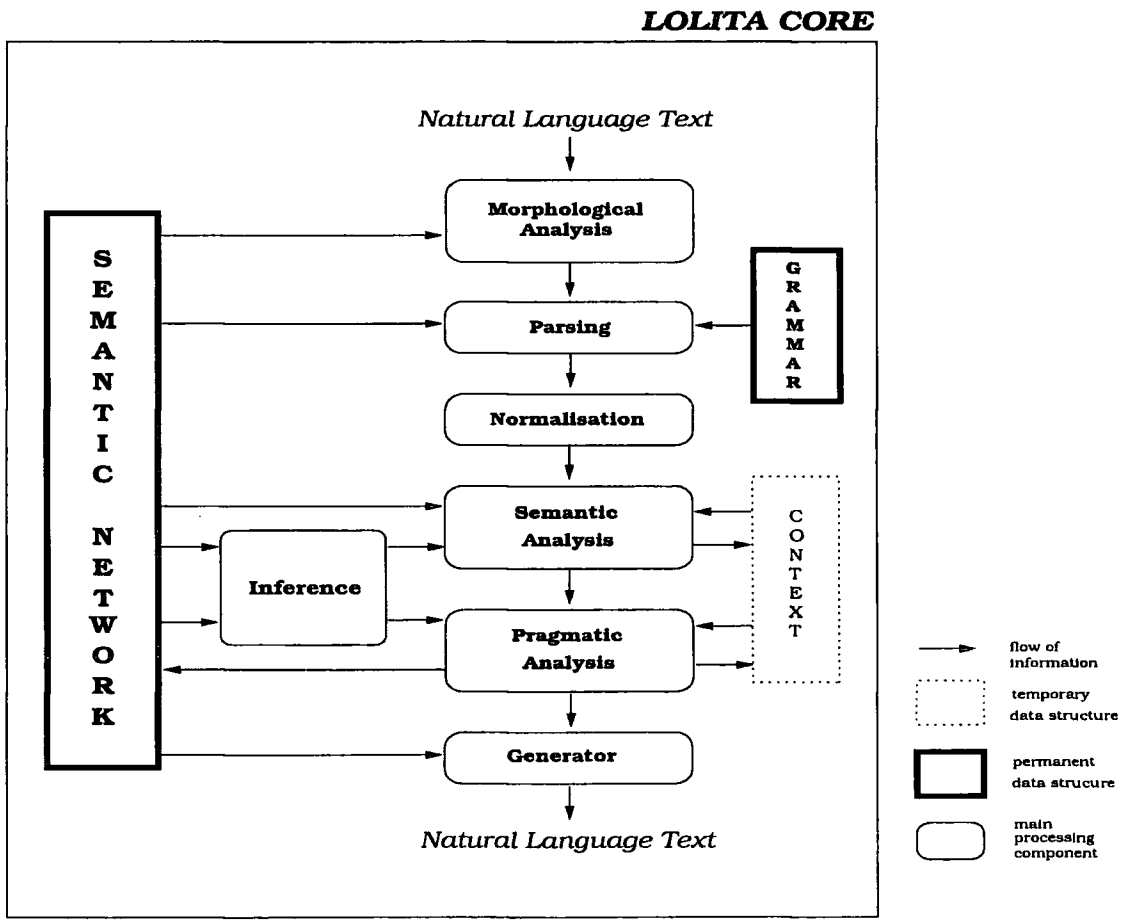


Figure 5.1: The architecture of the core of the LOLITA system

In the subsequent sections, the knowledge base and its representation is described, followed by a discussion of those components which deal directly with the analysis of natural language input.

5.1.1 SemNet and the ideas behind it

SemNet is a knowledge representation system based on the idea of *conceptual graphs* (Sowa, 1984). It consists of hierarchies of nodes connected together with arcs.

The nodes represent various types of information, while arcs represent direct relationships between the nodes. This is illustrated in Fig. 5.2 where a node representing the real world's concept of a `motorbike` is shown. In this example, the arc `generalisation_` connects the concept of `motorbike` to the concept `vehicle`. Other arcs link this concept to words in other languages or to its synonyms or similar concepts. Each node in the SemNet has a unique number, sometimes called a **noderef**, which forms part of LOLITA's internal representation of the node.

```
  * motorbike: 19864 *
generalisation_:
  vehicle - 4978
synonym_:
  motorcycle - 8591
similar_:
  bike - 32958
  minibike - 51389
italian_:
  moto - 28854
  motocicletta - 19102
spanish_:
  motocicleta - 95338
  moto - 95337
chinese_:
  {Not Printable Text} - 28698
  {Not Printable Text} - 28697
```

Figure 5.2: Example of a node and its links in the SemNet

At the time of writing, the SemNet consists of over 100,000 nodes, many of which have been acquired using WordNet (Miller, 1990) as a major source.

It should be noted here that it is not necessary for each concept in the SemNet to map onto some natural language word. In fact, the opposite is often the case. The SemNet is based on the assumption that there can exist many more concepts than words. For instance, it is quite legitimate to build a single node in the network representing "the man who owns a motorbike" for which no single word happens to exist in English or whatever natural language we are dealing with.

Each concept in the SemNet is meaningful to the extent that it is connected to other nodes in the network. That is, it can be interpreted fully only in relation to the rest of the SemNet.

Additionally, however, some kinds of frequently needed information are associated with each node more directly, in the form of the so called control variables. Altogether there exist more than 50 control variables in the system. The most important of them are: **type**, **rank** and **family**.

5.1.1.1 Type

The main **types** of concepts in SemNet are *entities*, (for instance, `airplane`, `Roberto`), *events* (`departure`, `Paul fainted`), *relations* (`to depart`, `to faint`) and *attributes* (`heavy`, `old`). There is some correspondence between linguistic categories and types. For example, the type *relation* usually corresponds to verbs, while the type *attribute* to adjectives. *Entities*, on the other hand, in general correspond to nouns. N.B.: instances such as `Roberto` or `Paul fainted` are differentiated from classes such as `airplane` or `departure` by means of **rank** (see next section).

Several nodes of types other than the four above also exist in the SemNet. These include: *determiner, pronoun, preposition, conjunction, subconjunction, prefix, suffix, punctuation, greeting and particle.* Their presence is motivated purely by lin-

guistic reasons—they are needed to categorize and store information about linguistic units. Items marked with these types do not carry any particular meanings in the SemNet, as they are only interpretable in the context of bigger *linguistic* structures.

5.1.1.2 Rank

The **rank** of a concept provides information about the concept's quantification. The use of ranks obviates the need of creating variables in the network and having to quantify over them. Instead, the LOLITA system uses only constants, marked with different ranks, and employs a multi-sorted logic to reason about them. Thus, depending on its rank, a concept obeys different inferential rules.

The values which a concept's rank can assume are: *universal, individual, named individual, bounded existential, framed universal, existential class, universal class* or *individual class.*

The rank of general classes of concepts such as `airplane` or `departure` would be *universal*; on the other hand the instances such as `Roberto` or `Paul fainted` would be ranked *named individual* and *individual* respectively.

5.1.1.3 Family

The family of a concept provides information as to which semantic class the concept belongs to. The most important semantic classes include: *living, animal, human, human organisation, inanimate, inanimate manmade, location, quantity.* Altogether there are just over 40 semantic classes in the SemNet.

5.1.1.4 Prototypical events

Apart from the knowledge inherent in the concepts hierarchy, LOLITA uses a class of events, labelled 'prototypical', to store other aspects of the knowledge of the world. Currently, the only prototypical events which are used in the system are associated with nodes representing actions. Each such event specifies the types and families of 'actors' that are 'proto-typically' involved when a given action takes place. For example: *Prosecutors prosecute suspects*, *Humans cook food*, *Owners own things*. It is intended to extend this system of representation by adding pre- and post-conditions to the prototypical events, as well by adding other types of prototypical knowledge, for example, ones defining entities.

5.1.1.5 Reasoning with SemNet

Knowledge implicit in the SemNet can be extracted using inheritance. Other forms of reasoning have also been implemented and these include analogy (Long & Garigliano, 1994), epistemic reasoning, reasoning about location (Baring-Gould, forthcoming), as well as standard logical connective reasoning.

5.1.2 Morphological Analysis

Morphological analysis is the first major stage in the processing of natural language input.

Initially, the input is divided into words and sentences ('tokenized') on the basis of spaces and punctuation or indicators such as SGML mark-ups. Contracted forms are expanded into full forms (e.g. *can't* into *cannot*) and some idiomatic phrases are recognized and joined together (e.g. *because of* is processed as one unit: *because_of*).

Next, root forms of words in the input are extracted using knowledge of inflectional rules.

Recognition of linguistic units and their categories then takes place on the basis of the knowledge stored in the SemNet. Standard grammatical categories are used, e.g. *noun* or *verb*, etc., along with some more semantically based categories, e.g *noun of cognition* or *noun of communication.*

Another aspect of morphological analysis is the guessing of the meaning of unknown words. This comes into play particularly when dealing with texts such as newspaper articles which frequently contain many names of objects, people and companies unlikely to be known to LOLITA. A system for guessing such words has been designed. It uses information such as whether the word is capitalised, whether its neighbourhood is capitalised, where in the text the word occurs, whether there are identifiable designators in the neighbourhood—such as "Mr" or "Corp.", etc. The work on improving the correct recognition rate is currently under way.

Brill's tagger (Brill, 1994) is employed in cases where LOLITA's own morphological analysis fails to lead to a parse. The tagger has the advantage of generating a smaller range of morphological possibilities for each word in a sentence. With a reduced input the parser might have a better chance of success second time round.

A list of all morphological possibilities is passed onto the next stage: parsing.

5.1.3 Grammar and Parsing

5.1.3.1 The grammar for English

The grammar for English is written in a formalism invented by LOLITA's designers. It is equivalent to a context free, phrase structure grammar, augmented by a feature system.

The grammar currently consists of nearly 1600 rules and its coverage of English is very wide. The aim is to account for as big a range of naturally occurring English sentences as possible, including those which traditional grammars might regard as

ungrammatical. In Chomskyan terms (Chomsky, 1965), LOLITA's grammar might be described not as a 'competence grammar' but a 'performance' one.

The grammar still has some gaps, however. For example, in a recent test (MUC7 dry-run) which used 25 previously unseen *New York Times* articles, comprising just over 880 sentences (with an average length of 21 words per sentence), approximately 7% of the sentences were not parsed due to missing grammar (though it must be added that over a quarter of the failures occurred in the titles of the articles, which are well known to be challenging).

Currently, the parsing mechanism is based on the Tomita algorithm (Tomita, 1986). The parser is a variant of the shift-reduce parser with a graph-based stack. The Tomita algorithm produces "tree forests" which are potentially very large. Sometimes, this causes problems for the feature unification part of the parsing process, which leads to reduced efficiency.

If the parser fails to produce a parse within a certain predefined time interval, it is by and large on sentences exceeding 40 words. In the test mentioned above, the rate of failures which were induced by time constraints (of not allowing the parser more than 5 seconds per word) was approximately 8%.

5.1.3.2 The parsing mechanism

The parsing stage takes as input the results of the morphological analysis and first tries to determine whether it can find any easily recognizable, unambiguous noun phrases during the so called 'island-parsing' stage. For example, some kinds of proper noun phrases (e.g. *"Mr William Grosvenor"*), date and time expressions (e.g. *"the fiscal 1992"*) or descriptions of locations (e.g. *"Palo Alto, Calif."*) can be analysed as such at this stage. The benefits of this are twofold: firstly, with some subparts of the input already locked into phrases the parsing of the whole input is simplified; secondly, if the overall parse fails, the subparts can still be used in later (though not so reliable) analysis. The drawback of this system is that if the

island-parsing is erroneous, no mechanism of recovery is provided.

Once the island-parsing is complete, the parser tries to find a suitable structure for the whole of the input. Due to structural and word sense ambiguity, it is often the case that many possible structures exist. To deal with this, the grammar and parsing modules incorporate a unique system of penalties, which allows the parser to discard the majority of the structures, thereby greatly limiting the number of analyses produced for each input.

5.1.3.3 The feature system

An important aspect of parsing and morphology is to combine the semantic and syntactic information of the linguistic input and express it in terms of **features** which are then attached directly to phrases as well as the terminal nodes (the so called 'leaves') of the syntactic tree.

```
sen [Sing, Past, TextDoc, Not Passive, Princ]
  full_propernoun
    MICHAEL [Sing, Male, Per3, Human, TextDoc, Princ, GramSubj]
  auxphrase_advprepph [Past, Tensed, TextDoc, Not Passive, Princ]
    verb
      TALK [Past, Tensed, Verb, TextDoc, Not Passive, Princ]
    advprepphs [TextDoc, Princ]
      prepp [TextDoc, Princ]
        prepNormMode TO [Adprep, TextDoc, Princ]
        full_propernoun
          JENNY [Sing, Female, Per3, Human, TextDoc, Princ]
      prepp [TextDoc, Princ]
        prepNormMode
          ABOUT [Adprep, TextDoc, Princ]
        poss_detph [Sing, Neutral, Per3, Term, TextDoc, Princ]
          possessiveDet
            HIS [Sing, Neutral, Per3, Term, TextDoc, Princ]
            comnoun JOB [Sing, Neutral, Per3, Term, TextDoc, Princ]
```

Figure 5.3: Syntactic tree with features

For example, in the parse tree produced for the sentence *Michael talked to Jenny about his job* each phrase and each terminal node of the tree (each *branch* and each *leaf* in LOLITA's terminology) has got features associated with it. A selection of these is illustrated in Fig. 5.3.

Some **features** perform a syntactic role, i.e. they are employed only during parsing, but several also play a crucial role in the later stages of the analysis. These include:

- **Number**. This feature is derived from the morphology of the word or from its syntactic environment and can assume the values: `Plur`, `Sing` or `NoNum`.
- **Gender**. This feature is inherent to the concept and in the majority of cases is derived from the knowledge stored in the SemNet, if available. Possible values are `Male`, `Female`, `Neutral` or `Sexed`.
- **Person**. This feature is usually inherent to the concept and its possible values are `Per1`, `Per2`, `Per3`, and additionally `NoPer3S` for items where the person can be anything but *Per3*.
- **Clause level**. This is a purely syntactically based feature stating whether or not an item occurred in the principle clause, subordinate clause, a relative clause or inside a prepositional phrase. Its values are `Princ`, `Sub Princ`, `RelPrep` or `Prepp`.
- **Semantic category**. Another feature inherent to the concept, this one reflects the family that a concept belongs to; its values can be `Human`, `Temporal`, `Location`, `Inanimate`, etc.
- **Document part**. This feature records, for example, whether the word has occurred in the body of the text or in the title part of the document. Its values are `TextDoc` for the body of the text, or `Headline` for the title part.

A lot of the feature information for the terminal nodes in the parse tree (the so called 'leaves') can be rebuilt by referring to the SemNet. However, at the phrase

level the features are only available via the syntax and so they have to be recorded at this stage.

5.1.4 Normalisation

The normalisation component's role is to transform, much like in early Chomskyan transformational grammar, some types of syntactic parses into other structures. In general, normalisation allows for a reduction in the number of rules needed in the Semantic module.

The output of the parser can be transformed in several ways: through various grammatical transformations, transformations of idiomatic expressions into a different form, filling in of gaps, rearrangement of Prepositional Phrases, etc.

An example of a transformation is a change of dative constructions into 'prepositional' ones (e.g. *"Paul gave Jenny a book"* becomes *"Paul gave a book to Jenny"*), or a fuller spelling out of some idiomatic expressions (e.g. in a sentence *"He went home"*, the expression *"home"* will be substituted with *"to his home"*).

It must be added that many normalisation rules are currently being phased out. This is mainly due to the fact that the grammar often undergoes small changes, which entails changing the patterns that trigger normalisation. Thus, from the software engineering point of view, the benefit of having fewer (though often more complex) semantic rules appears to be outweighed by the benefit of not having to make frequent changes to the normalisation patterns.

5.1.5 Semantic analysis

This stage of the analysis takes a parse tree as input and maps its elements onto a semantic network structure. Semantic rules are, in general, compositional.

To give semantics to a sentence *Michael talked to Jenny* (see the illustration in Fig. 5.1.5 on page 68), the rules apply from left to right, building partial SemNet structures for all the constituents first, before building the complete structure for the whole sentence.

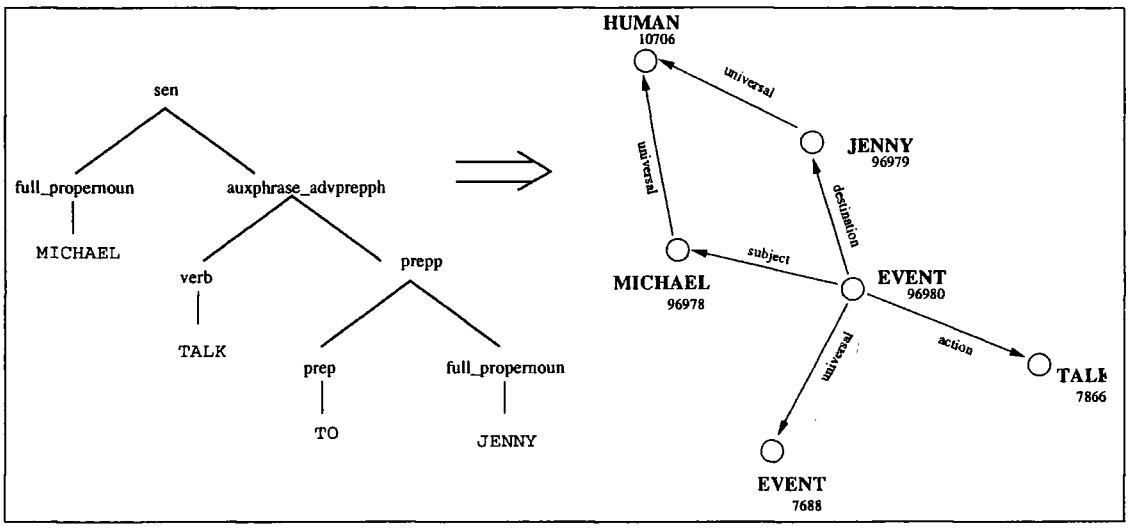


Figure 5.4: A parse tree and a corresponding piece of SemNet for the sentence *Michael talked to Jenny*

Thus, first a node is built to represent *Michael.* It is a node with rank "named individual" and it is connected to a node representing all humans, with a rank "universal". Next, the semantic rule for the verb is applied to the word *talk.* In this case, the rule just gives the verb a label: "Action". Next, the prepositional phrase is analysed whereby the preposition *to* is passed on together with the semantics for *Jenny* to the phrase level analysis. Here the rule for the phrase "prepp" labels *Jenny* as "Destination" (due to the presence of the preposition *to* on the left of the phrase). The verb phrase "auxphrase_advprepph" adds no more semantics to the elements it receives but simply passes them up to the rule for "sen". At this point, all the pieces are finally connected to form elements of a new event of *talking*, in which *Michael* is a subject and *Jenny* is a destination. The newly built event is also connected to a set of universal events of *talking.*

This example is simplified, because in reality there is more than one sense of the verb *to talk*, and so processes of sense disambiguation also come into play.

5.1.6 Pragmatic analysis

The next stage of processing involves analysing the event(s) produced by semantics in the context of LOLITA's knowledge stored in the SemNet. There are two main goals here. First, to disambiguate and enrich the meaning of the input and, second, to evaluate the resulting interpretation with respect to plausibility—that is to assess whether the new event is believable in the world that LOLITA knows about.

In order to achieve these goals a variety of heuristics are used and the knowledge stored in the semantic net is drawn upon.

Given that many aspects of pragmatic analysis are at the centre of the work carried out in this project only an outline is provided in this section. A more detailed discussion follows in chapter 6, with the new developments discussed in chapter 10.

5.1.6.1 Processing single events

LOLITA's knowledge of the world is organized around 'action-centred' prototypes (cf. section 5.1.1.4 above), therefore whenever a new event is being processed, its action has to be disambiguated first. Currently, the choice of verb sense is guided mainly by the word frequency information derived from WordNet or other sources (e.g. from the British National Corpus). Also, a heuristic of preferring those meanings for which there exists a prototype is employed. If that still leaves ambiguity, one of the meanings is chosen at random.

Once the sense of the action is decided, each of the remaining nodes in the event is compared with the corresponding prototypical one. This allows for further enrichment and disambiguation, as well as an assessment of the plausibility of the

event as a whole. For example, given a sentence *He cooked a joint* (and assuming no preceding discourse) the semantics would produce a structure equivalent to: *A male creature cooked a {joint1 or joint2 or joint3...}*[1]. Given that a prototype for `cook` states that the subject of `cook` must be human, pragmatics is able to enrich the meaning of *male creature* into something more specific, i.e. *a male human individual.*

Disambiguation too can be facilitated by the prototype: the object of cooking must belong to the set of all foods, therefore the only plausible meaning of *joint* must be the one synonymous with *roast*, i.e. a piece of meat, as opposed to other senses, such as, for example, the 'body part' sense or the 'shape' sense.

If the original sentence had read *He cooked a table*, no sense of *table* could be found which satisfied the requirement of the prototype, and LOLITA would conclude that the event must be implausible (a `status` of *low belief* would be attached to the final analysis).

On the other hand, if more than one sense of a word turns out to be compatible with the prototype, a series of heuristics is applied in order to try and eliminate the remaining ambiguity. The details of these will be explored in subsequent chapters.

5.1.6.2 Analysing larger fragments of text

LOLITA processes its input clause by clause. Already at the parsing stage, complex sentences are divided into their constituent parts, whenever possible, so that at any one time the system deals with a single event. However, the system also assumes that any larger chunk of text it receives (such as a newspaper article, or a paragraph taken out of a longer text) constitutes a coherent whole and so the information contained in any clause should be integrated with the information obtained from the preceding text. Thus each new event is considered in the context of the events that have arisen from the text so far.

[1]There are currently 7 senses of the noun *joint* in LOLITA

All the information gained from processing a text is available to the system at all times, as it is stored in the SemNet. However, for the purposes of various pragmatic processes (including sense disambiguation and anaphora resolution) it is also important to keep a separate record of the information mentioned recently. This allows the system to keep track of what the text is focusing on at any given time. This particular aspect of the system is described at length in chapter 6, with new developments introduced in chapter 10.

5.1.7 Generator

The generator in LOLITA (Smith, 1995) has been developed without focusing on any specific applications and is therefore flexible and general. It takes as its input a piece of SemNet (at the same time, having access to the whole of the SemNet) and produces as output natural language utterances. It is used as an interface to LOLITA and its applications, as well as a debugging tool.

5.1.8 Implementation

LOLITA is written mostly in Haskell, a non-strict functional programming language (Hudak *et al.*, 1992). Two sections are written in C: the parsing algorithm and the SemNet data structure and its access functions.

Haskell is similar to LISP, in that building programs involves writing functions. It has a garbage-collected heap, uses lists as a basic type and provides for a higher-order use of functions. Two important advantages of the language are the enforcement of referential transparency and laziness. A 'lazy' style ensures that code is not executed unless needed. Thus, even though the system has the external appearance of a pipeline architecture, the evaluation of individual pieces of code need not occur in that strict order.

5.2 Where does the current project fit in?

The work carried out in this project aims to improve certain aspects of pragmatic analysis in the system. It focuses on redesigning the way in which information is used during pragmatic analysis. In particular, it focuses on those pragmatic processes which solve ambiguities resulting from the use of anaphoric expressions. As such, the work does not concern those aspects of pragmatics which assess events with respect to "believability". It deals only with the pragmatic processes of some types of disambiguation.

Chapter 6

Resolving References in LOLITA: the old system

This chapter concerns the LOLITA system as it was used during the MUC-6 evaluation (with only a very small number of added enhancements).

This system will be referred to from now on as **the old system** or **the old MUC6 system**, to distinguish it from the **the improved old system** (cf. chapter 7), or **the new system**, introduced in chapter 10.

The following sections describe how the old system handles ambiguous utterances and in particular anaphoric ambiguities.

6.1 Processing ambiguous utterances

Given an utterance in (31):

(31) *Michael talked to Paul about his script.*

it is not possible to determine with certainty which sense of *script*[1] is intended or whether it is Michael's script or Paul's script. However, when taken together with a short prior context, for example:

(32) *Paul, the TV producer, asked Michael to write a screenplay. At the end of the month, Paul and Michael had a meeting.*
Michael talked to Paul about his script.

it is realistic to assume that the sense of *script* is the one closely related to *screenplay.* The *his* is arguably still ambiguous, though it is rather biased towards *Michael* (of course, it's possible to think of a story which would make it biased towards *Paul*, e.g. just by introducing a possibility that *Paul* had also agreed to have a go at the script himself, perhaps).

Let us suppose that most hearers or readers would go for the first interpretation, i.e. the *script* as related to *screenplay* and *his script* as being *Michael's script.* The following sections describe how the LOLITA system attempts to model a process which would arrive at just such an interpretation of the utterance in (31). They also describe the way the system represents ambiguity.

In the interest of clarity, it is assumed that initial utterances in (32) have been fully disambiguated and their content is available to the system when the final utterance arrives. The problem of how that initial input is analysed falls slightly outside of the general mechanism discussed here and will be considered separately in section 6.2.1, below.

[1]currently there are 3 possible senses in LOLITA which would fit this syntactic position: script(1): a written version of a play or film, etc.; script(2): an orthographic system, e.g. a syllabary; and script(3): type of handwriting, e.g. calligraphy

6.1.1 The Context buffer

As utterances in a discourse are processed the information gained from analysing them is stored not only in the semantic net, but is also placed in a temporary data structure known as "Context".

The Context buffer contains the following information:

- SOURCE(S): who is the author of the utterance
- TONE: the style of the utterance
- LOCATION(S): where the utterance is taking place
- DATE: the date and time of the utterance
- TENSE LIST: a list of tenses used so far in the utterance
- TENSE TREE: a structure used to analyse tense into temporal relations
- TOPIC(S) : the fundamental concepts expected to appear in the text
- REFERENTS : the concepts that have been mentioned so far in the utterance
- INITIAL REFERENTS: referents which have appeared in the first one or two sentences of the discourse (up to a maximum of 20).

The SOURCE, TONE and LOCATION of the utterance are generally more important to LOLITA for dealing with dialogue or reported speech. All three of these can be put to one side for the purpose of this project, given that the texts used in the MUC6 co-reference task contain a negligible amount of reported speech or dialogue.

Also, because it is required that co-reference links be made regardless of whether they hold now or did so in the past, the aspects such as DATE, TENSE LIST and TENSE TREE do not play a part in the anaphora resolution scheme. (It is one of the anomalies of the MUC6 co-reference task that time is to be disregarded;

this sometimes leads to highly counter-intuitive connections in cases where, for example, the discourse describes one post occupied by different people at different times: *John resigned as president of ABC Corp. Paul became president of ABC Corp.* — according to the task definition, *Paul* and *John* would have to be seen as co-referent.)

The TOPIC part contains a static list of concepts that are likely to occur in the processed texts. Such a list is not always available, but, for example for MUC6, it was possible to define a set of concepts (all to do with "management succession") which were frequent in the input. The TOPIC list could then be used to aid the disambiguation of senses.

The most important part of the Context for present purposes is the list of REFERENTS. The REFERENTS are nodes representing entities, events or actions that have occurred in the discourse so far and which can be referred to in the subsequent discourse.

The maximum number of REFERENTS stored in the Context is 30 concepts or all the concepts within the last 5 sentences, whichever is the smallest. When the limit is reached, older nodes are discarded before the new ones are added, so that the limit is never exceeded.

There are three kinds of information that each REFERENT carries:

- information derived from the links that the referent has with the SemNet
- information inherent in the control variables associated with the referent
- feature information carried over from the morphology and parsing stages.

The most important REFERENTS contained in the Context after processing the first two utterances of (32) are shown in Figure 6.1, p. 77.

```
76268: have          - type: relation
                       [Sing,Neutral,Pres,NoPer3S,Verb]
96913: michael       - type: entity
                     - rank: named individual
                     - family: propername human
                       [Sing,Male,Ncont,Per3,Propnoun,Human]
96933: month         - type: entity
                     - rank: individual
                     - family: temporal quantity
                       [Sing,Neutral,Ncont,Per3,Term,Temporal]
96927: meeting       - type: entity
                     - rank: individual
                     - family: human organisation
                       [Sing,Neutral,Ncont,Per3,Term,Inanimate]
96910: paul          - type: entity
                     - rank: named individual
                     - family: propername human
                       [Sing,Male,Ncont,Per3,Propnoun,Human]
78674: write         - type: relation
                       [Sing,Neutral,Pres,Verb]
96911: writing       - type: event
                     - rank: individual
                       [Sing,Neutral,Ncont,Per3,Term,Event]
96914: screenplay    - type: entity
                     - rank: individual
                     - family: communication
                       [Sing,Neutral,Ncont,Per3,Term,Inanimate]
5069:  ask           - type: relation
                       [Sing,Neutral,Pres,Verb]
96921: asking        - type: event
                     - rank: individual
                       [Sing,Neutral,Ncont,Per3,Term,Event]
96915: tv            - type: entity
                     - rank: bounded existential
                     - family: inanimate manmade
                       [Plur,Neutral,Ncont,Per3,Propnoun,Inanimate]
```

Figure 6.1: Main referents in the Context buffer after processing the input: *Paul, the TV producer, asked Michael to write a screenplay. At the end of the month, Paul and Michael had a meeting.*

Note that the `TV producer` has not been placed in the Context buffer. This is because it is assumed that in the predicative nominal structure *Paul, the TV producer* only the head of the structure, i.e. in this case *Paul* introduces a concept which can be referred to in subsequent discourse (see also section 6.1.7.1, below, for more discussion of these structures). On the other hand, the concepts `writing` and `write`, as well as `asking` and `ask` can subsequently be referred to, hence they are stored in the context buffer (section 6.1.7.3 returns to this issue).

6.1.2 Representation of ambiguity

Whenever an utterance containing ambiguity is received by the semantic module, a special SemNet structure is built to express the ambiguity and to allow for later resolution. This section describes the steps involved in building a representation for the example in (31). It is shown how both the ambiguous and the unambiguous elements of the same utterance are dealt with.

First point of disambiguation concerns the action of the main event. In the analysis of 31, the most frequently occurring meaning of the verb *to talk* is chosen.

The noun phrases *Michael* and *Paul* are not regarded as ambiguous: the system checks the Context and finds two male individuals with matching names already present (see figure 6.1, p. 77). It therefore assumes that the current references to *Michael* and *Paul* are anaphoric and there is no need to create new nodes to represent them (section 6.1.6 below provides a more detailed description of how the matching of names is carried out). It simply connects them to a new event of `talking` (with `action_: talk`) as `subject_` and `destination_`, respectively.

When analysing the ambiguous phrase *his script* the system examines each component of the phrase in turn. If the input is a pronoun or a possessive determiner, as in the current case, the system checks the Context to see if any referents match the features. All such matches are collected and a `DUMMY` node is created (node `96932` in figure 6.2) which becomes the subject of a special kind of event: the so called `referring` event, i.e. one whose action is the LOLITA internal concept: `to refer_: 95960` and the objects are the matches found in the Context, in this case `Paul: 96910` and `Michael: 96913`.

For the head noun *script* the system recognizes that there exist three concepts in the SemNet. Therefore, another `DUMMY` node (`96930`) is built, this time one that is linked (via another `referring EVENT`) to the three possible senses of *script*, rather than to any objects in the Context.

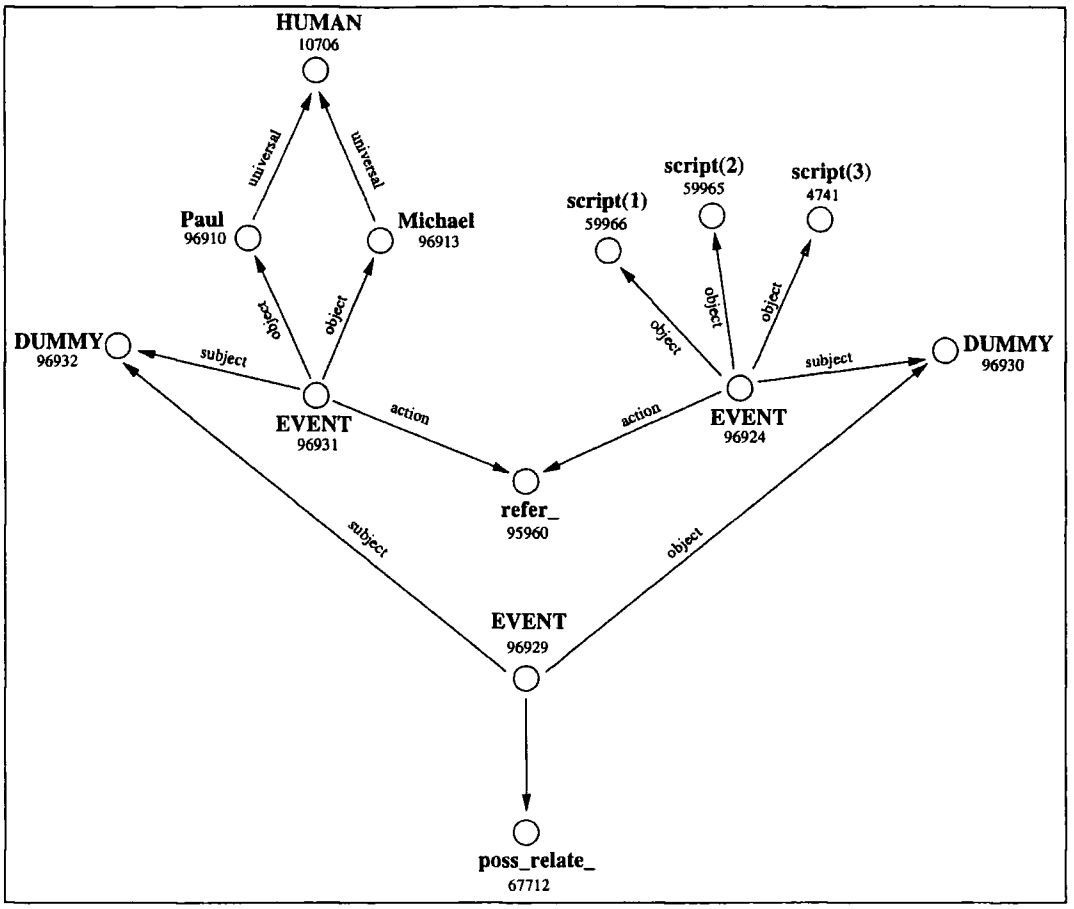


Figure 6.2: A piece of semantic net representing the phrase *his script* in the utterance of (31). NB the example here is simplified, as it disregards arcs recording the time, place and source of the utterance, as well as ignores some matters to do with quantification.

The two dummy nodes `96932` and `96930` enter also into another relation, represented by the `event 96929`. This is a relation deduced from the fact that a possessive pronoun was used, therefore, whatever the dummy nodes are resolved to, it is certain that one of them will be the subject and the other the object of an event whose action is some sort of possessive relation, represented in LOLITA with a system-internal concept `poss_relate 67712`.

The resulting representation of the ambiguous part of the utterance is shown in Figure 6.2.

6.1.3 Disambiguation process

During the next stage of interpretation the `talking` event is examined. The system attempts to check first whether the two unambiguous components: the `subject_` and `destination_` in this event are compatible with the prototype. Currently, however, there is no prototype for *talking* in the SemNet, so the system simply accepts what it is given as plausible.

Were the system's data to contain a prototype for this action, LOLITA would check the **type** and **family** of the candidate nodes to see if they were compatible with the information stored in the prototype. For example, if the prototype for `talking` specified that the subject must be *an entity* of family *human* (as is likely), `Michael` would clearly satisfy the prototype. If the utterance were to contain a node of another type and/or family as subject, the system would report a clash. A type clash carries a big penalty and immediately causes the whole event to be stamped with *low belief*. A family clash would cause the system to check its family hierarchy to see if the input family is compatible with that specified in the prototype. If not, again, the event would be treated with *low belief*.

Having dealt with the `subject_`, the `object_`arc is examined. Here, all three meanings would (presumably) match any putative prototype, as the object of `talking` could be anything. (In other cases, prototypes are likely to be more useful than this, as they are capable of ruling out any senses whose types and/or families are not compatible with the prototypical ones, thus reducing ambiguity.)

In the present example, the system looks at each of the three possible meanings of *script* and finds that one of them matches very closely an existing object in the Context: `screenplay: 96914`. A decision is made that the *script* of the current utterance and the `screenplay` in the Context are one and the same. (The algorithm responsible for this match is described in more detail in section 6.1.6, below.)

Next, the ambiguity of *his* resolved. The two possibilities, *Michael* and *Paul* have the same features as well as are of the same type and family, so no more semantic

rules could decide between the two. Instead, the system uses a series of heuristics to choose between them.

6.1.4 Preference heuristics

The following set of heuristics is applied for choosing between several possible senses, or referents in case of anaphora (for ease of reference in subsequent sections each heuristic is given a short name, shown here in brackets):

- Prefer the meaning that belongs to the appropriate hierarchy, as specified by the current event's prototype (*prefer_proto_child*).
- Prefer a more specific meaning to a more general one (*prefer_less_general*).
- Prefer the meaning which is closest to those in the list of 'topical' meanings given to the system beforehand—such a list is sometimes provided if the general topic of the text is known beforehand (*prefer_topic*).
- Prefer the meaning designating an entity rather than that designating an event or an action (*prefer_entity*).
- If there are several entities mentioned in the context prefer the meaning which refers to one carrying more inherent focus — for example, prefer the subject of a sentence over object over adverbial *etc* (*prefer_main_role*).
- Prefer the meaning that appears in the Context most recently (*prefer_last_mention*).
- Prefer the meaning whose frequency of occurrence (e.g. in the Brown Corpus) is higher (*prefer_common*).
- Prefer the meaning designating a 'named individual' to one with any other rank (*prefer_named_individual*).
- Prefer the meaning designating an 'individual' to any other (*prefer_individual*).

- Prefer a meaning which is more closely related to the family *human* (*prefer_human*).
- If there is a choice between a countable noun and a non-countable one, prefer the former (*prefer_countable*).
- Prefer a meaning about which more is known, i.e. it has most connections in the SemNet (*prefer_more_connections*).

The heuristics apply strictly in the order in which they are listed here. This order was arrived at by the process of trial and error, whereby several ways of ordering the rules were tested on sets of examples until what was deemed to be satisfactory ordering was established. Any non-preferred meanings or referents are rejected, while the preferred ones are kept and subjected to subsequent heuristics. In this sense, the heuristics could be said to act as filters, filtering out any non-preferred items. If any ambiguity remains after the final test, one possibility is chosen at random.

In our example, the heuristic preferring a meaning which was a subject of the most recent, preceding event (*prefer_main_role*) would choose `Michael: 96913`, as `Michael` was the `subject_` of the action `to talk`.

Unless only one possible candidate for a given anaphoric expression is found in the context, the system always uses the set of heuristics listed above to decide between the possibilities.

The same system would come into play too in cases where noun phrase ambiguity could not be resolved as easily as the ambiguity of *script* in the current example.

6.1.5 The ordering of referents in Context

In many cases where an anaphoric expression is resolved to something found in the Context, the resolved item is placed at the top of the list of REFERENTS. This

is to reflect a view that at the point of resolution the object can be regarded as most salient with respect to the reader's/hearer's current focus of attention. In this sense the list of REFERENTS is ordered with respect to this notion of saliency.

After the example utterance, the resulting Context structure is as illustrated in figure 6.3, below.

```
67712: poss_relate - type: relation
                     [Sing,Neutral,Pres,Verb]
78668: talk        - type: relation
                     [Sing,Neutral,Pres,Verb]
96928: talking     - type: event
                   - Michael talked to Paul.
                     [Sing,Neutral,Ncont,Per3,Term,Event]
96914: screenplay  - type: entity
                   - rank: individual
                   - family: communication
                   - The screenplay that Michael controls.
                     [Sing,Neutral,Ncont,Per3,Term,Inanimate]
96910: paul        - type: entity
                   - rank: named individual
                   - family: propername human
                     [Sing,Male,Ncont,Per3,Propnoun,Human]
96913: michael     - type: entity
                   - rank: named individual
                   - family: propername human
                     [Sing,Male,Ncont,Per3,Propnoun,Human]
76268: have        - type: relation
                     [Sing,Neutral,Pres,Verb]
96933: month       - type: entity
                   - family: temporal quantity
                   - The month that something has ends.
                     [Sing,Neutral,Ncont,Per3,Term,Temporal]
96927: meeting     - type: entity
                   - rank: individual
                   - family: human organisation
                     [Sing,Neutral,Ncont,Per3,Term,Inanimate]
78674: write       - type: relation
                     [Sing,Neutral,Pres,Verb]
96911: writing     - type: event.
                     [Sing,Neutral,Ncont,Term,Event]
5069:  ask         - type: relation
                     [Sing,Neutral,Pres,Verb]
96921: asking      - type: event
                     [Sing,Neutral,Ncont,Term,Event]
96915: tv          - type: entity: inanimate manmade
                     [Plur,Neutral,Ncont,Per3,Propnoun,Inanimate]
```

Figure 6.3: Main referents in the Context buffer after processing the input: *Paul, the TV producer, asked Michael to write a screenplay. At the end of the month, Paul and Michael had a meeting. Michael talked to Paul about his script.*

The figure shows the referents that are present in the Context after all three utterances of (32) have been processed. The phrase *his script* has been identified as co-referential with `screenplay: 96914` and moved towards the top of the Context, as it was the last item mentioned. A new event, `talking: 96928`, has been added. The ordering of `Michael` and `Paul` has now changed: `Paul` is seen as more recent than `Michael`. The reference to `Michael` via the possessive determiner *his* has not led to `Michael` being moved above `Paul`. This is because in the old system references made using possessives are not regarded as salient enough to warrant the reordering of the REFERENTS.

6.1.6 Resolving noun phrase anaphora

The preceding sections focused on how LOLITA represents and resolves utterances containing ambiguities, in particular pronominal ambiguities. In the following sections the processes specific only to noun phrase anaphora are described.

6.1.6.1 Proper names

Proper names pose a problem for LOLITA because the system has been designed to operate on constants to which it can refer by their unique names. It doesn't have a level of representation which would handle the concept of the "name" itself. So, an individual called *Paul* is naturally represented as a constant `Paul` and a member of the set of `humans`, for instance. However proper names are not unique enough in the world for LOLITA's scheme to work successfully, as many people can bear the name *Paul*.

In order to represent an entity like `Paul` and to express the idea that this entity "has a name *Paul*", the solution is to create a set of "entities that have a name *Paul*", to which the individual `Paul` would belong. The question is then where in the family hierarchy the set of proper names would belong, given that it has to be compatible with all those families of objects which can have proper names—and the range of

such objects is diverse. A special family of "proper names" has been created which belongs to a generic set of "linguistic units", an abstract concept embracing nodes such as `word`, `discourse`, `morpheme`, among others. It includes the concepts `name`, `forename`, `surname`, all of which are assigned family "communication".

```
94649 : Paul. (=> Forenames. : 94761)
 rank: universal
 family: propername human
 type: entity
 emotional value: indifferent
 level of language: common level
 output data: singular
 fixed form: yes
 gender: male
 language: english
```

Figure 6.4: LOLITA's representation of the set of human entities who are called *Paul*

The system's knowledge base contains representations of common forenames and surnames, for example *Paul, Michael, Jenny* or *Smith, Jones, Woods*, etc. They are assigned family "propername human" to distinguish them from proper names given to other classes of entities (for example, artifacts such as *Ford Escort* — this would have family "propername artifact", or organisations, such as *General Motors* which would have family "propername organisation"). Parts of proper names adopted by organisations are listed in LOLITA's data with family "propername organization". They include words like *Corporation, Associates, Association, Institution.*

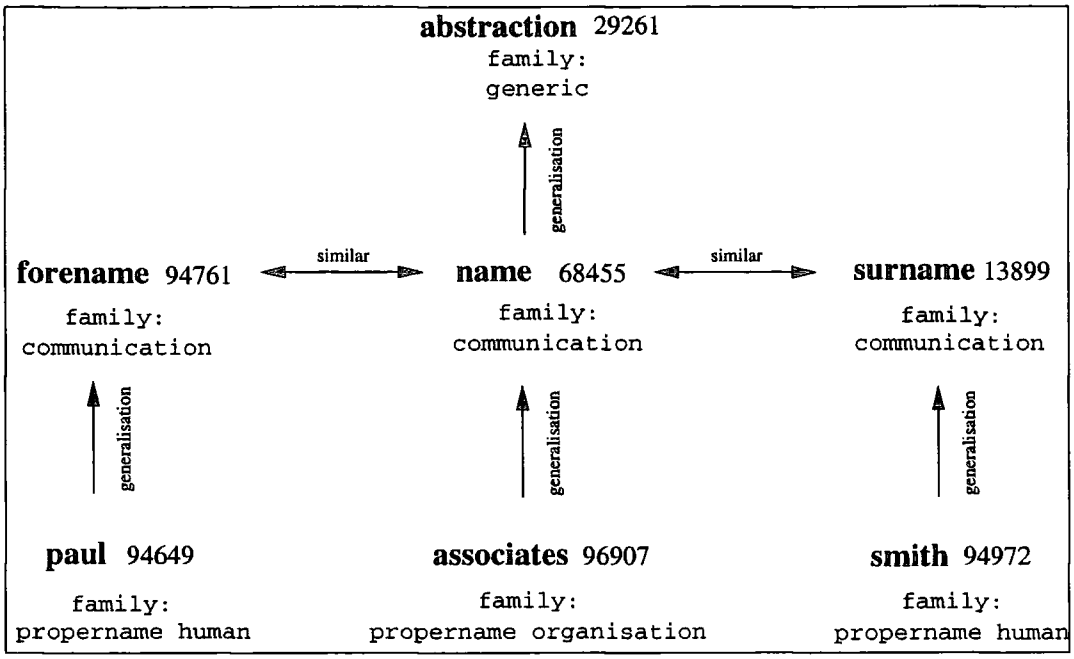


Figure 6.5: A fragment of the hierarchy around the concepts of **name**

The piece of semantic net in figure 6.5 illustrates how the concept of proper name fits into the hierarchy. Figure 6.4 illustrates the node and the control variables of the set of humans who bear the name Paul.

When analysing the example in (32), repeated here for convenience:

(32) *Paul, the TV producer, asked Michael to write a screenplay. At the end of the month, Paul and Michael had a meeting.*
Michael talked to Paul about his script.

the system first builds a new object (96910) to represent *Paul* and makes it a subject of an event (96921) with action **ask**:

```
   * paul: 96910 *
universal_:
  human - 10706 - rank: universal - family: human
  paul - 94649 - rank: universal - family: propername human
  producer - 96918 - rank: universal - family: human
subject_of:
  event - 96921 - rank: individual  (ask)  - suspended_
```

This representation of *Paul* is a hybrid between two different families of entities: `paul: 96910` becomes both an instance of all humans and of all proper names, but this doesn't cause problems, because the system treats these families as compatible. Also, because of being described as a *TV producer*, *Paul* is represented as an instance of all PRODUCERS.

When in the second utterance the system comes across the proper name *Paul* again, it tries to match it against the nodes stored in the REFERENTS list. The following tests are applied to find a match:

1. Is the type of the incoming node the same as the type of the Context node?
2. Is the family of the incoming node the same or compatible with the family of the Context node? (In LOLITA's family tree one family is compatible with another either if it is a subset of it or if the two are known to intersect.)
3. Is the gender of the incoming node compatible with the gender of the Context node?
4. Is the string from the incoming node compatible with the string of the Context node? (Here the string and abbreviation matching takes place.)
5. Is the feature Person of the incoming node compatible with the corresponding feature of the Context node? (Here the nodes are compatible if either their person is the same or one of the two is unspecified.)

If all the above tests return "True" then the system concludes that it has found a match. The node `paul: 96910` is a match in our case, while the other nodes from the Context failed on one or more of the above tests.

The test for assessing the compatibility of the actual strings requires a further comment. For the current example, this test could be quite trivial i.e. it would only have to check if the strings associated with the two nodes are the same. However, proper names often contain more than one string and can frequently be referred to by using just a subset of those strings. For example:

(33) *Madison Group Associates Inc. appointed a new president. A spokesman for Madison talked to the press.*

Also, abbreviations or acronyms can be used to refer to previously introduced individuals, as in:

(34) *The Securities and Exchange Commission contacted the company. The company did not comply with the SEC regulations.*

The string matching algorithm takes into account the above phenomena in the following way:

1. If both the incoming node and the Context node are associated with just one string each and these two strings are identical, then assume a match
2. If the incoming node is associated with a set of strings that constitute a subset of the strings associated with the node in Context, then assume a match
3. If the incoming node is an abbreviation which is compatible with all the strings associated with the node in Context, then assume a match

Rule 1 would be responsible for matching all the occurrences of *Paul* and *Michael*, respectively, in example (32). Rule 2 would allow us to make a co-reference link between *Madison* and *Madison Associates Group Inc* in example (33) and rule 3 would match *SEC* with *Securities and Exchange Commission* in example (34).

6.1.6.2 Definite noun phrases

When analysing definite noun phrases, such as *his script* in the example (32), the system searches the Context to see if it can find a match for any of the senses of the head noun (*script* in this case).

The following matching rules are applied:

1. If the incoming node is a synonym of or is similar[2] to the node in Context, then assume a match

2. If the incoming node is a more general concept than the node in Context, then assume a match

3. If the incoming node is a more specific concept than the node in Context, then assume a match

For node A to be more general than node B, both nodes must belong to the same hierarchy and node A has to be higher than node B in the hierarchy. The nodes have to belong to the same family. In simpler terms, this test asks the question "is B a kind of A?", e.g. "is a cat a kind of mammal?" — if so, it is concluded that B refers to A (i.e. "the mammal" refers to "the cat").

For node A to be more specific than node B, node A has to be below node B in the same hierarchy. The requirement for both nodes to belong to the the same family also holds. This tests, therefore, can be seen as asking the question "is A a kind of B?" and if so, the conclusion drawn is that B refers to A (in this case, "the cat" might be taken to refer to "the mammal").

In example (32), one of the senses of *script* (`59965`) is a more general concept than the *screenplay* found in the Context (in the SemNet, `script: 5996` is a direct generalisation of `screenplay: 9233`), so, according to rule 2, the two may co-refer.

Rule 2 would also be responsible for matching noun phrases such as *the company* and *BMW* in the examples 3–5, in chapter 2 (pp. 12–13). That is because when the system recognizes a string to be a name of an organisation or company, the representation it builds is connected to the appropriate hierarchy, for example, the following would represent *BMW*:

[2]The similarity of nodes is a weaker concept than synonymity. Most "similar" connections between concepts in LOLITA are equivalent to the "synset" connection in WordNet

```
   * BMW: 96913 *
universal_:
  company - 15702 - rank: universal - family: human organisation
  BMW - 96912 - rank: universal - family: propername
```

This sort of representation allows a noun phrase *the company* (or any synonyms) to match the node representing *BMW*.

If the system finds more than one match for any of the senses in the Context, it passes the final disambiguation decision to the preference heuristics described in section 6.1.4, above. If more than one sense of the incoming noun finds a match (or matches), the disambiguation heuristics are then applied to all the candidates rather than to the actual list of possible senses of the noun.

6.1.7 Other types of co-reference

Apart from the co-references discussed so far, the following types of co-reference link were also required in the MUC-6 task: set membership-based co-reference and the so called 'functions and values' based co-reference. On the other hand, references to events and actions, or references with multiple antecedents were not required. Each type is briefly discussed below, with an explanation of how the phenomena are handled by the old LOLITA system.

6.1.7.1 Set membership-based co-reference

This is the type of co-reference found in examples like this:

(35) *Paul, director of ABC Corp, retired.*

A co-reference relation is assumed to hold between *Paul* and the predicative nominal *director of ABC Corp.*

These kinds of links are handled by LOLITA mainly on the basis of parsing, semantics and a special treatment at the point where actual co-reference links are added to the original text, i.e. the "application" level. The system builds so called "is_a" events, like the following:

```
   * event: 96918 *
subject_:
  paul - 96910 - rank: named individual - family: propername human
action_:
  is_a - 19894 -
object_:
  director - 96917 - rank: universal - family: human
time_:
  present_ - 20989 -
```

Next, the co-reference marking application finds such an is_a event and adds an appropriate markup to the `object_` and the `subject_` of the event.

No searches for antecedents in the Context are involved for the predicative nominals in such constructions.

6.1.7.2 'Function and value' co-references

These co-reference relations arise in examples like this:

(36) *The sales of the new product rose to $1,000,000.*

where *The sales* and *$1,000,000* are assumed to refer to the same amount. Or:

(37) *The company announced a 14% jump in profits.*

where *14%* is seen as co-referential with *a 14% jump in profits.*

The treatment of these by-passes LOLITA's pragmatics rules etc. It is handled, again, at the application level, analogously to the "is_a" based co-references, and relies on a correct parsing and semantic analysis, with no searches of the Context involved.

6.1.7.3 References to events and actions

Some types of co-reference links are not considered markable in the MUC-6 task. For example, uses of some expressions as referring to events:

(38) *Michael passed his exams. This pleased everyone. It meant that he could now go to university.*

Both *This* and *It* refer to *Michael's passing of the exams* which is not expressed in the utterance as a noun phrase. It would be interesting to evaluate how the system manages to do such co-references, however within the existing evaluation scheme this is not possible. We are limited to marking co-reference relations which hold only between noun phrases in the input text.

By analysing isolated examples, it seems that the old system doesn't cope with such co-reference too well, because of the early application of the rule which prefers entities over events. So, while it would correctly resolve the items in (38) when processed in isolation (because here the only possible candidates are events), in other cases, where there were singular entities to choose from as well as events, the entities would always win (assuming, of course, that they hadn't got ruled out by the knowledge available in the relevant prototype before reaching this stage).

6.1.7.4 Multiple antecedents

Another type of co-reference not markable in MUC-6 is co-reference to more than one antecedent. For example the pronoun *They* in (39) will not be marked:

(39) *Michael talked to Paul. They discussed the script.*

The old system resolves this co-reference correctly. However, if there was a plural antecedent already in the Context, it would have been preferred, which is not necessarily always desirable. In any case, for the purposes of the MUC evaluation, co-reference links with multiple antecedents have to be filtered out (however, this happens at the peripheral level and not in the system's core).

6.2 Other aspects of interpretation

6.2.1 Disambiguating initial input

It may be justified to assume for the sake of argument that the initial input is already disambiguated when discussing the analysis of non-initial utterances. But in real life, it too has to be disambiguated by the system, in the absence of any Context. If the analysis of this input then serves as an anchor to subsequent analysis, it is clearly very important that it's done as reliably as possible.

This is where a TOPIC list becomes very useful for the disambiguation of senses aspect of the analysis. If the TOPIC does not help, the system uses information such as sense frequency, if available, or knowledge encoded in the prototypes, if present. For anaphora resolution, it is assumed that potential anaphoric expressions will either find antecedents within the initial input or will be regarded as "new mentions", i.e. the system will build new objects to represent them.

6.2.2 Recognition of what is anaphoric

In general, the system searches for a referent for any object categorized at the syntax level as potentially anaphoric. That is, we don't take into account the data from the corpus studies (Poesio & Vieira 1997, Fraurud 1990) which suggest that

anaphoric expressions may vary in this respect. This strategy may lead to loss of precision if we end up with many spurious links.

6.2.3 Non-referential uses of *it* and *there*

It is aimed to recognize the non-referential uses of *it* and *there* at the level of syntax. If this strategy works as intended, the semantic component receives no stimulus to start a search for an antecedent. In the old system the following types of constructions are handled successfully:

(40) a. *It seems that Michael went to the meeting.*
b. *It is certain that Michael went to the meeting.*
c. *It is thanks to Paul that Michael was on time.*
d. *It is expected that Michael will arrive on time.*

However, the following constructions (which differ syntactically from those in (40)) are not handled well:

(41) a. *It is easy for Michael to write a script.*
b. *Michael finds it easy to write.*
c. *There is a script to be written.*

In constructions such as those in (41), the system launches a search for an antecedent for *it* or *there*, which leads to an inevitable loss of precision.

6.3 Drawbacks of the old system

6.3.1 Criticism of the filter-based algorithm

One of the major drawbacks of the current algorithm is that it allows no backtracking. This may often lead to erroneous analyses. Consider the following example:

(42) *To discuss the script, a meeting of all the staff was organized. It was very productive.*

The *It* from the second utterance will be resolved, erroneously in the present system, to the *script*, because the early heuristic of preferring non-events to events filters out the *meeting* as a potential candidate. The two rules which would favour the *meeting*: *prefer_last_mention* and *prefer_main_role* don't have a chance to come into play at all, because they apply too late. So, in all those cases where one early rule favours one candidate, while several of the subsequent ones favour another, the input is likely to be misinterpreted.

Another drawback of the current algorithm is the application of the same heuristics to both sense disambiguation and anaphora resolution. This sometimes seems inappropriate. For instance, the rule to do with sense frequency information or the *prefer_topic* and *prefer_countable* rules, can be seen as more relevant to sense disambiguation. They shouldn't be used for anaphora, as they might be damaging.

For example, assuming hypothetically there were concepts like *computer* and *software* in the TOPIC LIST and then an antecedent for a pronoun *it* had to be chosen from among *ABC Corp, computer, software*, the first candidate would get rejected early on, because it wouldn't be present or related to items in the list of TOPICS.

A problem arises also with respect to the "prefer more frequent meaning" rule (or *prefer_common*, as it is labelled in LOLITA). In one of the training texts used for this project, this rule did, in the past, lead to an erroneous co-reference link. There

was a choice of referents for the pronoun *it*, among which one of the candidates was `New York`, which happened to be marked as frequent. Preceding heuristics didn't distinguish between candidates, so *New York* was chosen—both, as it happens, incorrectly and in a rather counter-intuitive way.

This problem also underlines the importance of the correctness of all the data to the co-reference resolution problem. If there is a mistake at the very low level of the system, it's unlikely that the higher level modules will produce a good result.

Conversely, some rules are more relevant to anaphora resolution than to sense disambiguation. For instance the rules which prefer subjects over objects, etc., or the *prefer_last_mention* rule, have no bearing on sense disambiguation, so it's both inappropriate and inefficient to apply them here.

6.3.2 No implementation of grammatical constraints

The old system provides no way of dealing with examples such as those described in sections 2.7.1 in chapter 2 and 3.2.1 in chapter 3.

So, for example a sentence like (43), processed in isolation:

(43) *Paul likes him.*

is analysed as:

(44) `Paul likes himself`

6.3.3 The drawbacks of the Context design

6.3.3.1 Referents not available in Context

It seems that in real life discourse not all types of objects are equally accessible for any type of subsequent reference. For example, the events such as those in (38) can probably be referred to via a pronoun only very soon after the utterances expressing them. At the other extreme, individuals, identified in the discourse by proper name, can be referred to via their name long after they are out of the immediate focus.

The implementation of referents history as a simple stack does not allow us to take the above into account. This might lead to a drop in recall. One example occurs in the MUC6 training article, where we have a reference to a *Mr Schweizer* in the very beginning of the text:

(45) *The Swiss chemical and pharmaceutical group also appointed Rolf W. Schweizer chief executive officer.*

After five subsequent sentences (containing over a 100 words), another reference to *Mr Schweizer* is made, however, the node representing this individual has already dropped out of the Context, so the two cannot be co-referenced.

6.3.3.2 Misleading order of referents in Context

Another problem is connected with the way the system processes its input and which portions of the Context stack it considers as relevant when examining characteristics of possible antecedents. The analysis of the following example illustrates the problem:

(46) *ABC Corp. introduced the new product. Its marketing cost the company $1,000,000.*

In the second utterance, the possessive determiner *Its* could be resolved to any of the following candidates: `ABC Corp.`, `the product` and `introducing`. (NB. the `introducing` will be rejected straight away by the heuristic preferring entities over events, so it can be ignored for the moment). The system builds a dummy node in the normal way to stand for the various possibilities for *Its*. The dummy is left unresolved until after the semantics has completed building the whole event.

On the other hand, the noun phrase *the company* is resolved immediately, still at the semantics stage: only one match, *ABC Corp*, is found and no dummy needs to be built. Thus, *the company* and *ABC Corp* are unified and the newly merged node moves towards the top of the Context, in the normal way.

The event of `costing: 96930` is finally built and the nodes are passed onto the pragmatic analysis. When the disambiguation of the *Its* eventually takes place, the top of the Context buffer contains the following:

```
20325 : cost          - type: relation
                      [Sing,Neutral,Pres,Tensed,Verb]
96930 : costing       - type: event
                      [Sing,Neutral,Term,Event]
96929 : 1,000,000     - type: entity
                      [Sing,Neutral,Ncont,Per3,Term]
96923 : ABC Corp.     - type: entity
                      [Sing,Neutral,Ncont,Acc,Per3,Propnoun,Human]
96915 : marketing     - type: event
                      [Sing,Neutral,Ncont,Nom,Per3,Event]
74196 : introduce     - type: relation
                      [Pres,Tensed,NoPer3S,Verb]
96919 : introducing   - type: event
                      [Sing,Neutral,Term,Event]
96911 : product       - type: entity
                      [Sing,Neutral,Ncont,Acc,Per3,Term,Inanimate]
```

Now, the *prefer_main_role* heuristic applies, trying to choose between `ABC Corp` and `product`. It finds that the role of the most recent occurrence of `ABC Corp.` is that of `object_`, because `ABC Corp.` was the object of the `costing 96930` event.

However, this is intuitively wrong. It could be argued that from the point of view

of the resolution of *Its* the relevant role of `ABC Corp.` is that of `subject_` of the preceding event: `introducing 96919`.

The role of the second candidate, `product` is also that of `object_`, so, the *prefer_main_role* heuristic keeps the two possibilities: `product` and `ABC Corp` still open. Next, the *prefer_last_mention* heuristic rejects `product` because it finds that, according to the order in the Context, `ABC Corp` was the more recently mentioned of the two.

Thus, the disambiguation procedure, as it stands, looks at the properties of candidates as they appeared **after** the ambiguity occurred, which is intuitively inappropriate, at least for anaphora[3].

6.3.4 Independent disambiguation within events

When dealing with events in which, say, both the subject and the objects are ambiguous, LOLITA resolves one ambiguity first and then the second one without taking into account the first resolution. In many cases, this may have no ill-effects, but there are examples where a different approach would be beneficial. E.g. (47):

(47) *Michael wrote a screenplay for a film. It was a good script.*

In the second utterance, the object of the *is_a* event, i.e. *a good script* is disambiguated first: as in the examples discussed before, it becomes resolved to the *screenplay* from the previous utterance. Next, the subject of the *is_a* (*It*) is resolved but without the knowledge of what the object was resolved to. So, at the point of disambiguation of the subject the object is still treated as ambiguous. The *It*, therefore, becomes co-referenced with *a film*, which happens to win on the basis of the *prefer_last_mention* rule in this case. The fact that the *It* is asserted to be a script is ignored.

[3]This sort of procedure might be more justified for handling cataphora, but that's not the objective here

Clearly, the independent (and erroneous) disambiguation of the subject and the object in this and similar cases can be seen as a serious shortcoming.

6.3.5 The timing of disambiguation in complex sentences

In complex and ambiguous sentences, the system sometimes makes early, erroneous decisions, which may lead to misinterpretations later. For example, the utterance in (48):

(48) *Orin Godsey said that he fainted.*

is resolved to:

(49) `Orin Godsey (human organisation) said that some male fainted`

Because the name *Orin* is unusual and unknown to the system, when LOLITA looks at *Orin Godsey*, it decides that the latter must be a proper name of some organisation. Then it comes across the pronoun *he*, but by then it's too late to realize that the pronoun most probably refers to *Orin Godsey*—at this moment, Orin Godsey can no longer be a human individual.

6.3.6 Shortcomings of the proper name matching

One of the problems in the proper name co-reference resolution is to do with the abbreviation matching. For example, the old system is unable to resolve *AT&T Corp* as a reference to *American Telephone & Telegraph Co.* This is because the system assumes that all parts of the anaphoric proper name have to be shorter or equal to the corresponding strings in the name of the antecedent, while here the *Corp* is longer than the *Co* in the antecedent. The fact that *Corp* is very likely to be synonymous with *Co* is ignored.

Another problem with this assumption is that it is likely to overgeneralise. For example, the referents such as *Kaman Unit* or *Kaman supplier* would be regarded as good antecedent for the string *Kaman*. Such matches are most likely to be erroneous. Clearly, it's not enough just to look at the strings.

6.3.7 Shortcomings of the noun phrase matching

The current rules for noun phrase matching don't always work well when it comes to matching noun phrases against `named individuals`.

In the example (32), it was shown that LOLITA represents the concept `Paul` as an instance of all `producers`, having derived this information from the expression *Paul, the TV producer....* However, if an anaphoric noun phrase *the producer* were to be used later on in the discourse, the old system ignores the extra knowledge it has about `Paul` and treats both `Paul` and `Michael` as potentially equally good matches.

Interestingly, the extra knowledge about `Paul` being `a producer` can hinder some anaphoric links. If a more general noun phrase such as *the guy* were used in the subsequent discourse, `Michael` would be considered as a possible match, while `Paul` would not. This is because of the way the hierarchy is searched to check if the concepts are compatible. Neither of the three rules for noun phrase matching would conclude that *a producer* and an anaphoric expression *the guy* could co-refer. First of all, the two are not synonyms or 'similars'. Secondly, the knowledge that *guys* can be *producers* or vice-versa is not encoded in the is_a hierarchy. Quite the contrary, the hierarchy states that if anything the two are *sisters*, and so they are potentially incompatible (or mutually exclusive), even though both are eventually subsumed under all humans.

On the other hand, since all we know about *Michael* is that he is human, the semantic analysis attaches the node representing *Michael* directly under the node `human`. This makes him compatible with every concept that comes under `human` in

the hierarchy.

To solve this sort of problem at the level of knowledge representation might require some way of marking which sets of objects are mutually exclusive and which are not. However, adopting such a solution potentially leads to the problem of having to define an arbitrarily large number of intersections.

Another problem in the noun phrase matching algorithm is related to the way proper names and individuals that have proper names are represented. Consider the node Paul again:

```
   * paul: 96910 *
universal_:
  human - 10706 - rank: universal - family: human
  paul - 94649 - rank: universal - family: propername human
  producer - 96918 - rank: universal - family: human
subject_of:
  event - 96921 - rank: individual  (ask)  - suspended_
```

Via the universal_ link to the proper name Paul: 94649, the node representing the individual Paul becomes an instance of all names: 68455 (see figure 6.5 for the relevant fragment of the hierarchy). Because of this, the phrase such as *the name* has a chance of becoming co-referenced with the individual Paul: 96910, which is clearly undesirable.

Chapter 7

General improvements to the old system

7.1 LOLITA at the time of MUC6

The system which was officially evaluated for the 6th Message Understanding Conference (the 'old system') suffered from three major problems. First, the parsing component left a lot of room for improvement. Second, the Named Entity recognition rate was fairly low, as compared with other systems. Third, the system contained a series of trivial errors in the code. Altogether, these three major shortcomings resulted in a considerable drop in performance.

In the general approach adopted by the LOLITA project, every core component plays an important role in the final result. Consequently, if any of the components is unsatisfactory, overall performance is affected.

Moreover, it must be borne in mind that the LOLITA system was not specifically designed to perform just the MUC tasks. It is a general purpose, natural language system. A lot of effort was spent on designing peripheral modules (such as those that allow the system to reproduce the original input texts and add markups to it)

had to be built from scratch. This meant that less time was left to spend on the improvements to the system's core.

7.1.1 The impact of parsing problems

The parsing stage is particularly important for the co-reference task, because correctness of many co-reference links crucially depends on the parse. For example, in the training and test sets of MUC6 articles used in the current experiment over 20% of all co-references are the 'function and values' co-reference or 'is_a based' co-references, which rely heavily on the parsing. Similarly, over 25% of all co-references were those involving proper names and acronyms. In these cases, successful parsing is needed before the names matching algorithm can be triggered.

Also, as shown in the previous chapters, many types of anaphora need syntactic information to be resolved successfully. When such information is not available, errors inevitably arise.

If a parse is not produced at all, the drop in scores can be dramatic. At the time of the original MUC6 evaluation, the parsing system was performing well below its potential: the sentence failure rate was about 12%. Once a sentence didn't parse, everything from this sentence was lost to subsequent analysis. Often, this led to a loss of co-reference links that the most naive pattern matching systems would have found without problems.

7.1.2 The 'Named Entity' and 'Co-reference' tasks dependence

A correct Named Entity recognition is often an important first step in making a correct co-reference link. If, for example, the system mistakenly identifies the name of a company as a name of a person, subsequent references to this company with

pronominal anaphora are likely to be missed or resolved incorrectly.

7.1.3 Coding errors

At the time of the MUC6 evaluation the feature system wasn't working well. In particular, the elements stored in the Context buffer often lost their original features and were allowed as referents for any anaphoric expressions. Also, the feature merging functions didn't work correctly (these are applied at final point of anaphora resolution, when two nodes are unified and a newly created node is placed in the Context). The features of newly merged nodes were often erroneous, for example the gender feature would be lost or changed from `Female` to `Sexed` or even `Neutral`; similarly, the grammatical number information would sometimes be changed from `Plur` to `NoNum`.

As as result LOLITA sometimes ended up connecting nodes representing companies or place names to individual people and connecting together personal pronouns such as *he* and *she*. Co-reference chains such as *"Sheffield - the city - he"*, *"Texas American Group - he - she - it"* were not unusual.

7.1.4 Text output errors

There were also problems with the module of LOLITA which keeps track of the input text. The module records which concepts created by the system's linguistic analysis come from which parts of the input text. This procedure is necessary for MUC6 tasks such as the co-reference task, where the system is expected to generate the input text exactly, but with added appropriate SGML mark up.

Minor errors in this module resulted in LOLITA occasionally inserting spurious space characters in some places, while deleting others. This adversely affected the final result because the scoring software is particularly sensitive to any misalignments between the answer keys and the responses.

It must be noted, however, that this module is currently regarded as peripheral with respect to LOLITA's main capabilities. Consequently, any results reported in subsequent chapters will assume that any text output errors are corrected, if necessary, before scoring.

7.1.5 General assessment of the old system

Superficially, some of the types of errors reported in the sections above could give the impression that, at the time of MUC6, LOLITA's anaphora resolution component was almost non-existent. Clearly, being able to rule out connections between pronouns of different gender or to connect together multiple occurrences of easily recognizable "Named Entities" should be quite elementary.

However, the LOLITA system aims to deal with a wide variety of natural language input — indeed, its aim is to be able to perform a deep semantic analysis on any form of NL — and it is therefore highly complex. The sub-system for resolving anaphoric expressions is also highly complex in itself. Moreover, this sub-system has to rely on the rest of the system performing well. Such complexity leads, on the one hand, to many opportunities for the whole system to go wrong. On the other hand, however, it gives the system the potential to process successfully all types of examples, from very easy to very difficult.

(Callaghan 1998) provides an interesting analysis of LOLITA's performance at MUC. He defines the notion of "easy" and "difficult" co-references (or, more generally, slots, as his framework can be used for any of the MUC tasks) based on the statistics of what the participating systems actually achieved. To put it simply, the co-references that most other systems made correctly were, by definition, regarded as "easy". On the other hand, if only a small number of systems successfully made a given co-reference link, such a link would be regarded as "difficult". Callaghan's analysis demonstrates further that in fact the LOLITA system at the time of MUC6 failed to make many "easy" co-references yet made relatively many "difficult" ones.

Given this fact, it can be assumed that if the system makes difficult co-references correctly, it must have the potential to make the easy ones. This is consistent with the hypothesis, which will be adopted here, that the system failed to make many easy co-references due to errors described above.

In summary, the worse than expected performance of LOLITA on the MUC6 tasks can be largely attributed to parsing problems on the one hand, and to almost trivial errors in the code and in the text output module, on the other.

7.2 Preliminary Development

Given the conclusion that shortcomings of the lower level modules, such as morphology, parsing and the feature system, led to a significantly reduced co-reference score, problems in these areas were addressed first.

Additionally, some changes to anaphora-critical components were made to the system and these are detailed in section 7.2.3.

Several of the changes described below were designed, tested and debugged by the author of this thesis. Some of the changes were also implemented by the author. Other members of the group were involved in the implementation of changes in the area of the grammar and parsing, often using the analysis and data provided by the author. Work on these changes took about 18 months.

7.2.1 Changes to the parsing component

Island parsing was introduced (cf. section 5.1.3.2 in chapter 5) which significantly improved the parsing success rate. Moreover, two extra parsing passes were introduced: a second pass using Brill's tagger (Brill 1994) which restricts the set of lexical entries, subsequently constraining the parsing, and a third pass using a reduced grammar, aimed at recovering constituents of complex sentences, if a

full parse wasn't possible. Finally, in cases where all three parsing passes fail, a way of recovering all the pronouns, possessive determiners and some noun phrases (those related to the TOPIC of the text) was devised. The implementation of these improvements was carried out by Prof. Garigliano and Callaghan.

7.2.2 Changes to the NE recognition component

The Named Entity recognition component was revised and many new rules were added. A major change was introduced to LOLITA's morphology module, which allowed the system to reuse names of entities previously recognized in the text, rather than treat the entities in each sentence of the text separately.

A change in the treatment of unknown proper names that appear without clear *designators* (i.e. without *Corp, Ltd, Mrs*, etc.) was introduced. In the old MUC6 system a decision as to what type of entity an unknown name stood for was made early, and usually resulted in the conclusion that it must stand for an organisation. The new improved treatment, on the other hand, involves the introduction of the concept of "human_or_organisation", the use of which allowed for a delay in the decision, until some disambiguating information became available at a later stage.

For example, given the following first sentence of an article:

(50) *Shortly after Fossett's launching Monday his competitors sent him telegrams of congratulation.*

The system cannot decide what sort of entity *Fosset* is on the basis of this name itself. However, the use of the pronoun *his* as well as the absence of any other possible referents, provide the disambiguating clues.

The implementation of the above changes was carried out by Prof. Garigliano with materials and analysis provided by the author.

7.2.3 Anaphora-critical changes

7.2.3.1 SemNet changes

The part of the SemNet relevant to the 'management succession' scenario was checked and adjusted as necessary. All the concepts in LOLITA regarding job titles, such as *director*, *CEO*, etc. were thoroughly checked and marked with a new family: `job`. This work was carried out by the author.

7.2.3.2 Grammar, semantics and MUC application changes

Grammar was improved and expanded to allow for a better parsing of the training articles. Analysis of "is_a" based co-reference and "function and value" co-references was overhauled at the parsing and semantics levels. Additionally, various "surface" rules were added to their treatment. The latter were written at the MUC co-reference application level and not as part of the core components of LOLITA. The new grammar and semantic rules were written and implemented by Prof. Garigliano on the basis of the analysis and materials provided by the author. The changes were tested and debugged by the author.

Most of the newly added constructions were needed to handle sentences like (51), below, and their variations:

(51) *Paul, director of ABC Corp, retired.*

Previously, the system had problems with these sorts of structures, particularly because it didn't have a good way of handling singular noun phrases without determiners (such as *director of ABC Corp*).

Now, the sentence in (51) is parsed in the following way:

```
sen
 full_propernoun_description
   moved_propnoun_description
     full_propernoun_simple
       propernoun_forename PAUL
   substitute_descriptions
     copula_missing_det
       relprepcl
         comnoun DIRECTOR [Sing,Sexed,Per3] * 3
         prepp
           prepNormRel OF
           propernoun_types
             full_propernoun_comp
               propernouns
                 propernoun_not_comp ABC
                 propernoun CORP [Neutral]
 verb RETIRE [Past] * 6
```

Given the above parse, the semantic analysis picks up the appropriate labels (`moved_propnoun_description`, `substitute_descriptions`, `copula_missing_det`) and builds the following event:

```
    * event: 114727 *
generalisation_:
  event - 7688 - rank: universal  (happen_)
subject_:
  paul - 114719 - rank: named individual - family: propername human
action_:
  is_a - 19894 -
object_:
  director - 114725 - rank: universal - family: job
time_:
  present_ - 20989 -
```

The MUC co-reference application then scans all the event nodes built by the system, picks out any that contain the action `is_a` (or another copula verb). If the subject of the action is of family `human` or `propername human`, and the object is of family `job`, the system marks the two as co-referent.

7.2.3.3 Article headlines

A new treatment of headlines (or titles) of the text was introduced. It is estimated that in the MUC6 texts co-referential links to headlines constitute around 10% of all co-referential links and in the MUC6 tests LOLITA only managed to make approximately one fifth of them.

The headlines in the training articles appear quite distinct from the rest of the text and were proving difficult for the system to deal with successfully. Poor analysis of headlines not only reduces the system's recall, but is also likely to lead to a serious mis-analysis of the rest of the text.

As well as having a special set of grammar rules at the parsing stage, the headlines are now analysed at the end of the text. They are processed in the context of the initial one or two sentences of the main body of the text. (This is similar to the way the Sheffield system (Gaizauskas *et al.* 1995) approached this problem.)

For example, one of the headlines from the MUC training set is as follows:

```
<HL> Who's News:
Spelling Entertainment Picks New Chairman, CEO for Blockbuster
</HL>
```

This is a particularly problematic headline to get the system to interpret as a first utterance of the input. The system doesn't know that *Spelling Entertainment* is a company (no database of company names is used). Nearly all the words in the headline could be used as common nouns. To confuse matters further, all determiners are omitted.

However, in the context of the first sentence of the article:

```
Spelling Entertainment Group Inc. named H. Wayne Huizenga, chairman
and chief executive officer of Blockbuster Entertainment Corp.,
following Blockbuster's recent purchase of a controlling interest
in Spelling.
```

at least both *Spelling Entertainment* and *Blockbuster* can easily be recognized as companies. Also, the reference to *chairman* in the first sentence aids the interpretation of the part *New Chairman, CEO....*

7.2.3.4 Pleonastic pronouns

Some improvements to the recognition of the pleonastic *there* and *it* were also carried out. Some of the recognition rules proposed by Lappin & Leass (1994, pp. 538-539), have been implemented. However, more work needs to be done in this area.

7.2.3.5 Improvements to Context

The way in which old nodes get deleted from the Context has been improved. Now, the deletion of nodes takes place according to importance or inherent prominence of the node as a referent. For example, first the oldest node denoting an action is deleted, then the oldest node denoting an implicit event (i.e. one built by LOLITA), then the oldest event node of any kind, and so on. This changes were designed and implemented by the author.

7.2.4 Other general changes

Data concerning corporate designators were added to LOLITA's knowledge base (SemNet). About 8000 new forenames were added and all the existing forenames were checked so that they were marked correctly for gender.

Other data changes which took place before the evaluation reported here were largely to do with the topic of the MUC7 competition and were concerned with airlines and air crashes.

The checking of the existing forenames was done by Yang Wang. All the remaining

data work was carried out by the author.

Additionally, as part of the on-going development of the LOLITA system and thus involving most members of the LOLITA group, considerable general debugging of the whole system was carried out.

Chapter 8

Evaluation of results

8.1 Theoretical issues

8.1.1 Evaluation of complex NL systems

One of the objectives in this project is to assess how well a sub-component of a complex natural language processing system (i.e. the anaphora resolution algorithm) contributes to the performance of the whole system on a certain task.

The system is large-scale and therefore it would be desirable to test the performance on a large corpus. The testing of large scale systems within the MUC framework involves performing a given task using as input a set of previously unseen articles.

However, the contribution of the anaphora sub-component can only come into play when other sub-components of the system can be assumed to work well. Because good performance of other sub-components can be ensured only on "open" tests (due mainly to limited resources and time), it is the "open" tests which can provide the greatest insight as to the contribution of the sub-component in question.

By contrast, the results obtained using "blind" tests cannot tell us much about

the performance of the sub-component, since we don't know how well other sub-components have performed. This is a very important point to bear in mind and its significance will be stressed again later, during the analysis of final results.

Possibly, the only way insights from a blind test can be gained would be through a detailed post-test analysis, but this is not always possible.

Using blind tests, however, does give an indication of how well the system as a whole is performing. An improvement in such overall performance is another objective of this project. (For more disscussion on evaluation see Galliers & Sparck Jones (1993))

8.1.2 The evaluation setup

The main evaluation method used in this project is similar to the one used in MUC6. A set of articles from the *Wall Street Journal* is selected and annotated with appropriate SGML tags which represent co-reference links according to the task description (the description of the task definition can be found in appendix A). The tagged articles constitute the so called *answer keys*.

LOLITA processes the original, unannotated texts and produces its own SGML markup. LOLITA's answers (the so called *responses*) are then compared with the answer keys using scoring software provided by the MUC organizers.

8.1.2.1 The training corpus

The small set of selected articles used for development consisted of 15 texts chosen from the training corpus provided by the MUC organizers (see appendix C). The articles all come from *The Wall Street Journal* and the majority of them concern the topic of 'corporate management change' (a topic chosen for the formal MUC6 evaluation).

Another training text used was a larger article from *The Guardian* newspaper (see appendix C) on a related topic. The set of 15 training articles and the single article from *The Guardian* served as tests in the process of debugging and preliminary improvements as well as in the development of the new co-reference resolution system. They constitute what will subsequently be referred to as the "open tests".

The answer keys for these articles were prepared by the author, according to the co-reference task description.

8.1.2.2 The testing corpus

A further 20 articles were randomly selected from the MUC6 training corpus and divided into two sets of 10 articles for convenience (henceforth, they will be referred to as "test A" and "test B"). The articles in test A and test B were used as semi-blind tests. They are considered *semi*-blind because the answer keys for them were prepared by the author, however, to all intents and purposes, they were used as if they were *blind.* That is, their content was never examined during development, unlike the other sets, where the analysis of particular sentences was tested and improved upon. Additionally, the set of 30 articles which constituted the MUC6 formal co-reference test was kept to be used as a completely blind test. These 30 articles have not been inspected; the answer keys for them were provided by the MUC6 organizers.

The purpose of using the tests A and B was twofold: to help assess LOLITA's performance during the improvements and development work and to be able to analyse the effects of the new algorithm on unseen texts at the end of the project. The formal MUC6 blind test set couldn't be used for the latter purpose because the LOLITA project participants wish to preserve that set as blind.

8.1.3 The MUC6 scoring method

In order to measure the accuracy of the system's output with respect to the keys, a model theoretic method for scoring is used Villain *et al.* (1995). The score is obtained by looking at co-reference chains (or 'equivalence classes') in the key and comparing them to corresponding co-reference chains in the response. The main idea behind the scorer is to find the number of links that would have to be added to or taken away from the response to make the response's equivalence classes the same as that of the key. So, for example, given sets of co-reference links like this:

Key: {A-B B-C C-D}
Response: {A-B C-D}

a key equivalence class {A-B-C-D}, with 3 links, is established[1]. The response, containing only 2 links, would need just one extra link adding (between B and C) in order to obtain the same equivalence class as that of the key.

On the other hand, in a case like this one:

Key: {A-B B-C C-D}
Response: {A-B B-C C-D D-E}

the response's equivalence class of {A-B-C-D-E} would need one link taking away in order to agree with the key's equivalence class.

The co-reference links that need to be taken away are counted as *spurious* links. The links that need to be added are counted as *missing*. The remaining links are counted as *correct*.

[1]The additional assumption behind establishing an equivalence class between these four nodes is that each link asserts identity. Therefore if A is the same as B and B is the same as C, then A is also the same as C and the same as D. It is in this sense that {A B C D} are an equivalence class

8.1.4 The *Recall* and *Precision* measures

The number of correct co-reference links made by the system divided by the number of all links in the answer key constitutes the *Recall* measure.

The number of correct links made by the system divided by the number of all links made by the system (i.e. the correct links plus the spurious links) constitutes the *Precision* measure.

8.1.5 The *F-measure*

In order to express the recall and precision as a single value, the so called F-measure (or F-value) is used. The F-measure is calculated according to the following formula:

$$F = \frac{(\beta^2 + 1.0) \times Precision \times Recall}{(\beta^2 \times Precision) + Recall}$$

The above calculation of the F-measure is due to Van Rijsbergen (1979). It is traditionally associated with the interpretation of the performance of systems on various Information Extraction tasks. The value of β depends on the relative importance given to recall over precision. If recall and precision are to be of equal weight, $\beta = 1.0$.

The F-measure was not used in MUC6 with respect to the co-reference task, however, it has since been used for the scoring of this task at the MUC7 competition. The measure is designed in a way which puts low value on results where recall is very very low and precision extremely high (or vice versa). So, for example, a system which marks a small subset, say, 10% of the co-reference links, with 100% precision will only achieve an F-measure of around 18.

The F-measure is used in this project for convenience, as it makes comparisons between systems more straight forward. The value of β adopted here is 1.0.

8.1.6 The scoring software

The main scoring software used for co-reference is that provided for the MUC7 competition. The MUC7 co-reference scorer is an improved, debugged and more efficient version of the scorer used in the MUC6 competition. The MUC6 scorer is also used but mainly for analysis purposes (the discussion of this follows in section 8.3).

8.2 Other scoring and evaluation issues

8.2.1 Key preparation

If the score of the system's results is to give a fair indication of how well the system performs, it is important to consider how the materials against which the score is calculated are prepared.

There are several problems associated with the preparation of the keys. One is to do with the interpretation of the task definition. Another, to do with errors which are inevitable in the keys. Also, it's possible that the original texts may be ambiguous.

Ideally, one would like to see the co-reference task designed in such a way that most humans would agree on it. This could be tested by taking keys prepared by one person and scoring them against those prepared by another. The aim would be to obtain a high level of agreement between the keys.

MUC6 organizers measured annotator variability for the co-reference task Sund-

heim (1995). Two independent annotators prepared a set of 17 articles and those were then scored against each other. The scores were 80% recall and 82% precision (F-measure: about 81%). It was found that the main areas of disagreement stemmed from different interpretation of vague portions of the task description, from subjective decisions when the text was ambiguous and finally from noun phrases being overlooked (i.e. annotator errors).[2]

One might argue that in view of this sort of inter-annotator score, if a system under evaluation reaches an F-measure of 80%, it means that the system's algorithm is as successful as a human annotator.

However, test results between just two annotators are not enough as a basis of such a judgement.

Also, if the system performed 80% correctly, this could be variously interpreted. For example, if in the test set a certain category were to constitute 20% of all co-references, and our system happened to ignore this category, but got all the others right, then the system's score of 80% would not be very convincing.

Sundheim reports that most human errors/differences pertained to definite descriptions, such as *the company* and bare nominals in premodifying positions, such as *aluminium* in *aluminium siding* Sundheim (1995, p. 20). At the same time very few differences pertained to proper nouns and pronouns.

In view of the above it would be advantageous to break down the scores into categories of anaphoric expressions. This would allow for an assessment of how well each category did. Moreover, the scores on those categories in which there was higher inter-annotator agreement could then be regarded with more certainty/importance.

[2]During the key preparation for the current project it was found that at least 20 co-reference links were added to the originally prepared keys by the end of the project; i.e. with respect to over 400 co-references that were marked up in the final version of the keys, about 7-8% were missing in the first version of the keys

8.3 Scoring by categories

In order to provide more information on how well LOLITA performed on certain categories of anaphoric expressions, a novel way of scoring has been designed.

First of all, the answer keys have been marked with category tags, shown in the next section. Secondly, the original scoring software has been enhanced to take these tags into account.

8.3.1 Breakdown into categories

Each noun phrase in the answer keys which enters into a co-reference relation with another noun phrase is now classified as one of the following:

Names

FLNM - full name (e.g. Eastco Industrial Safety Corp)

ABNM - abbreviated or shortened name (e.g. Eastco)

ACNM - acronym

Common nouns

APNN - inside appositive (John, *chairman* and *CEO*,...)

TRNN - noun in a ternary copula (they named *John chairman*)

CPNN - NP object of a copula verb (John is *chairman*)

PRNN - premodifier in compounds (*profit* margin)

DMNN - noun phrase with demonstrative determiner (*this cat*)

DANN - noun phrase with definite article (*the cat*)

IANN - noun phrase with indefinite article (*a cat*)

HRNN - head of relative clause (*the man* who retired)

PSNN - noun phrase with possessive determiner (*his cat*)

SGNN - noun with Saxon genitive determiner (*Roberto's cat*)

NDNN - singular noun phrase with no determiner

NDPN - plural noun phrase with no determiner

HDON - head of NP not marked otherwise

ADVR - referential adverbial (*here, there*)

Pronouns

PSPR - possessive pronoun (I like *mine*)

PSDT - possessive determiner (*his* book)

PRPR - ordinary pronoun

RFPR - reflexive pronoun

DMPR - demonstrative pronoun (*this* was the cause)

Numerical

PERC - percentage (*14%*)

ABSV - absolute values (*6 million*)

RELV - relative values (*$3 a share*)

OTHV - other values (*sales of 2 million*)

Time

DATE - dates, years in standard format

REFT - referring expression involving time (*the last three months*)

8.3.2 Difficulties in scoring by categories

There are at least two difficulties which have been encountered when trying to score the system's performance broken down by categories.

The first lies in the category tags themselves. Some of the features listed above are overlapping, for example, full names can occur inside appositive constructions. Also, some categories are relative to the discourse in which they occur: for example, if the discourse refers to an entity called *John* only using *John* and not any fuller version of the entity's name, then there would be no justification to label *John* as a shortened name. On the other hand, if the text contains the noun phrases *John Smith* and *John*, referring to the same entity, the noun phrase *John* is a shortened name with respect to *John Smith.* By contrast, some labels are absolute: a reflexive pronoun will always be a reflexive pronoun.

When it comes to tagging the keys, whenever more than one tag could apply, it was decided to use the tags from the point of view of how the system under evaluation would treat the noun phrases in question. So, if a noun phrase occurred in a construction like this:

(52) *John, the president of ABC Corp,...*

it was expected that the system would parse the construction as a appositive structure and make a co-reference link on this basis. The resolution of this noun phrase would proceed differently from resolution in this case:

(53) *John likes the president of ABC Corp.*

In the latter case, the anaphora algorithm for noun phrases with a definite determiner would be triggered.

A more difficult problem is with the question of what score to assign to the individual categories, and on what basis. In the easy cases, for example, if the whole chain of co-reference links is correct, then each category found in this chain could be assigned one correct point. Likewise, if a noun phrase is not in any chain (and it should be), then the category of this noun phrase could have one point taken away from its score (or, in other words, be treated as missing).

In the difficult cases, however, there may be elements present in the chain that shouldn't be there. How should the individual categories in this chain be regarded then? It was decided that a good enough solution for the current purposes would be to give each category that should be in the chain a score of one if the chain contains at least one other correct element. In this sense, such an item could be seen as correctly connected.

This may be a slightly generous way of scoring, however it was the best method which could be used at the time, as well as being very straightforward to implement.

Overall, the scoring by categories has provided a valuable metric of the system's changes during development. It allowed for a way of monitoring how individual categories were affected by the changes carried out at any particular time.

8.4 Interpreting the scores

It is important to consider what any co-reference score does and does not tell us. It might give us a general idea of how the system is performing, however, there are problems when it comes to evaluating different sorts of errors that a system can make.

8.4.1 Erroneous co-reference links

If a chain of referents contains one wrong item, does this mean that the whole chain should be treated as incorrect? It could be argued that the presence of one erroneous item does affect the meaning of the whole chain.

The weakness of the current evaluation method is that the score doesn't distinguish between the significance of different types of error: all erroneous links are treated as the same. In reality the errors vary, and can have different consequences. Some could totally change the interpretation of the whole discourse, while others would

have very little effect.

Considering example (54):

(54) *The judge looked at O.J. Simpson. He was accused of murdering his wife.*

if the system connects the *He* with *The judge*, rather than with *O.J. Simpson*, the consequences of such an erroneous link could be quite significant.

On the other hand, in an example like (55):

(55) *The president talked to his advisers. They decided to proceed with the defence.*

the question of whether *they* refers to *the president* and *the advisers* or just to *the advisers* is probably of less consequence.

8.4.2 Hidden errors

If, during analysis, the system connects together unmarkable phrases[3], or one markable phrase with another unmarkable one (assuming that the system recognizes the objects as unmarkable so it doesn't generate an SGML markup for them), then this sort of error is not visible in the score.

This shows the inadequacy of this particular method, were it to be used as a sole method of gauging the system's deeper understanding.

[3]these are phrases that do not appear in the key at all.

8.4.3 Problems with the task definition

Other problems with using the co-reference task for evaluation purposes arise from the task definition itself.

For example, for a text like this one:

(56) *President James C. Richardson Jr., 44, was named to the additional post of chief executive officer.*

the answer key would not link *James C. Richardson Jr* with the title *President* because the task demands that only appositives separated by a comma should be marked.

So, if a system makes such a link, this would count as an error. However, from the point of view of the meanings of the text the link can be seen as correct: both *President* and *James C. Richardson* could be used in some context as two different ways of referring to the same individual.

8.5 Final comment on evaluation

Despite its drawbacks, the MUC6 evaluation method is the best currently agreed standard. Alternatives such as detailed qualitative analyses would probably provide a much fuller picture as to how the system is performing. However, such methods are too laborious and time consuming and as such might hinder rather than encourage progress. Their cost would thus far outweigh their potential benefits. Hence, for the current purposes, the MUC6 scoring method is considered highly acceptable.

Finally, it should be added that the general ideas for system development encouraged by MUC competitions, with automatic scoring of the output on well defined

tasks, has proven to be invaluable in the development of LOLITA ever since the system first took part in a MUC competition.

Chapter 9

Performance of the improved old system

9.1 General scores

This chapter presents the results of the first evaluation carried out using the improved LOLITA system, with the old anaphora resolution algorithm ("the improved old system"). The scores are compared to those obtained using the system from the time of the MUC6 competition in October 1995 ("old LOLITA").

9.1.1 The formal blind test

Table (9.1) below shows the results obtained in the MUC6 formal test as well as those obtained for the MUC6 training set of 15 articles, for both the original MUC6 system (i.e. the old LOLITA, used in October 1995) and for the improved old system (see appendix B for details). The scores shown for the formal test are higher than those officially reported. They have been obtained after the correction of several formatting errors (cf. section 7.1.4) in the original output which had confused the scorer.

The scores are calculated according to the MUC6 scoring method discussed in chapter 8, section 8.1.3.

	Recall	Precision	F-measure
The old MUC6 system	40.1	50.2	44.6
Old LOLITA improved	61.9	62.3	62.1

Table 9.1: Evaluation results for the co-reference task using the LOLITA system on the MUC6 formal (blind) test, before and after preliminary development

In order to place LOLITA's scores in the context of other systems, Table (9.2) shows the scores of LOLITA's competitors at MUC6. The table orders these scores by F-measure (NB these were not used at the time of MUC6). Most of the scores included here are the official scores, except for the LaSIE and the Pennsylvania systems, who reported slightly better scores after fixing some very minor errors.

LOLITA's Competitors	Recall	Precision	F-measure
Pennsylvania system	63	72	67.2
FASTUS	59	72	64.8
PIE	63	63	63.0
LaSIE	54	70	61.0
New York system	53	62	57.2
RESOLVE	44	51	47.1

Table 9.2: Co-reference evaluation results of other participants of MUC6

As can be seen from the above data, the improved old LOLITA would be classed in fourth place among the other systems. This is a very encouraging result, because the improvement in the score is considerable. The system is now placed in the middle of the results table, rather than at the very bottom, where it had appeared previously.

Furthermore, the new score indicates that the shortcomings of the disambiguation algorithm were not the major problem which held LOLITA back from achieving scores comparable to those of its competitors. It seems that the changes in the morphology, parsing and NE components, as well as the addition of surface rules for co-reference (section 6.1.7, chapter 6) have made a big difference to the overall score.

9.1.2 The open test sets

The table below illustrates the results obtained for the MUC6 training set of 15 articles and for the single article from the *Guardian*.

	Recall	Precision	F-measure
LOLITA: the old MUC6 system			
MUC6 training set	35.9	56.1	43.8
Guardian art.	51.1	64.8	57.1
Old LOLITA improved			
MUC6 training set	78.8	79.3	79.0
Guardian art.	76.1	77.8	76.9

Table 9.3: Evaluation results for the co-reference task using the LOLITA system on the open test sets, before and after preliminary development

Again, the improvement in the scores is notable. Particularly so, the rise of the scores on the training set: there is a very large increase (about 43 points) in overall recall, after the preliminary improvements of the system. The increase in precision is also impressive: about 23 points. These amount to a 35 point increase in the F-measure.

For the *Guardian* article the picture is similar, even if the scores are slightly lower than those of the training set.

It is also noticeable that all the scores here are much higher than those achieved on the blind test. This is to be expected, because during preliminary improvements, the aim was to ensure that all aspects of the analysis of the texts in the open tests worked as well as possible. The same level of accuracy of analysis cannot be hoped for in the blind test sets.

9.1.3 The semi-blind test sets

Table (9.4) shows the co-reference task scores obtained for the semi-blind sets.

	Recall	Precision	F-measure
LOLITA: the old MUC6 system			
Test A	40.3	61.2	48.6
Test B	31.0	40.7	35.2
Old LOLITA improved			
Test A	56.8	71.8	63.4
Test B	69.8	69.3	69.6

Table 9.4: Evaluation results for the co-reference task using the LOLITA system on test sets A and B and, before and after preliminary development

The rise in scores on the semi-blind tests is also high, though not as high as that observed on the open tests. In fact, the score on test A (with improved old LOLITA) is now very close to that obtained in the formal blind test.

9.2 Further analysis: breakdown into categories

The following sections include the figures obtained with the new scoring method described in chapter 8. The results in the open tests are included first. This is

followed by the results obtained on the semi-blind tests. In each section, the tables of data are displayed first and then followed by a comment.

Each table contains the name of the category in rows (for explanation of the category names see chapter 8, pp. 121-122). The columns contain the following:

RESP is the number of correctly linked items in the response;

KEY is the number of items in the key;

OVG is the number of items in this category that the system connected to the wrong chain.

The remaining three columns contain recall, precision and F-value respectively. The table also shows the number of items which the system has marked and which should not have been marked, as they didn't enter into any co-reference relations in the key (hence their category is not tagged in the key). These are labelled "spuriously marked".

9.2.1 Analysis by categories of the MUC6 training set

Tables (9.5) and (9.6) (on pages 133 and 134, respectively) show detailed scores obtained for the MUC6 training set of 15 articles.

CATEGORY	RESP	KEY	OVG	REC	PREC	F-VAL
Abbreviated name	48	53	0	0.91	1.00	0.95
Full name	34	61	3	0.56	0.92	0.69
Acronym	0	1	0	undef	undef	
Inside appositive	9	22	1	0.41	0.90	0.56
Premod in compounds	8	15	0	0.53	1.00	0.69
NP inside ternary copula	5	22	5	0.23	0.50	0.31
NP with demonstrative det	1	4	0	0.25	1.00	0.40
NP with definite article	18	51	1	0.35	0.95	0.51
NP with indefinite article	0	15	0	undef	undef	
NP object of copula verb	1	8	2	0.12	0.33	0.17
Head of relative clause	2	4	0	0.50	1.00	0.66
NP with possessive det	3	20	2	0.15	0.60	0.24
NP with Saxon genitive det	0	5	0	undef	undef	
Sing noun with no det	0	24	3	undef	undef	
Plural noun with no det	2	10	2	0.20	0.50	0.28
Head of NP	1	1	0	1.00	1.00	1.00
Possessive pronoun	0	0	0	-	-	
Possessive det	15	33	5	0.45	0.75	0.56
Ordinary pronoun	6	21	1	0.29	0.86	0.43
Demonstrative pronoun	0	0	0	-	-	
Referential adverbial	0	0	0	-	-	
Percentage	0	11	0	undef	undef	
Absolute value	0	29	0	undef	undef	
Relative value	0	9	0	undef	undef	
Other value	0	12	1	undef	undef	
Date	7	14	0	0.50	1.00	0.66
Other time expression	0	1	0	undef	undef	
Total markable objects	160	446	26			
Spuriously marked: 87						

Table 9.5: Detailed results for the MUC6 training set, obtained using old LOLITA, at the time of MUC6

CATEGORY	RESP	KEY	OVG	REC	PREC	F-VAL
Abbreviated name	53	53	0	1.00	1.00	1.00
Full name	56	61	0	0.92	1.00	0.95
Acronym	1	1	0	1.00	1.00	1.00
Inside appositive	21	22	0	0.95	1.00	0.97
Premod in compounds	8	15	0	0.53	1.00	0.69
NP inside ternary copula	20	22	2	0.91	0.91	0.91
NP with demonstrative det	3	4	0	0.75	1.00	0.85
NP with definite article	36	51	3	0.71	0.92	0.80
NP with indefinite article	10	15	0	0.67	1.00	0.80
NP object of copula verb	7	8	0	0.88	1.00	0.93
Head of relative clause	2	4	0	0.50	1.00	0.66
NP with possessive det	13	20	2	0.65	0.87	0.74
NP with Saxon genitive det	4	5	0	0.80	1.00	0.88
Sing noun with no det	18	24	2	0.75	0.90	0.81
Plural noun with no det	9	10	0	0.90	1.00	0.94
Head of NP	1	1	0	1.00	1.00	1.00
Possessive pronoun	0	0	0	-	-	
Possessive det	26	33	4	0.79	0.87	0.82
Ordinary pronoun	16	21	2	0.76	0.89	0.82
Demonstrative pronoun	0	0	0	-	-	
Referential adverbial	0	0	0	-	-	
Percentage	0	11	4	undef	undef	
Absolute value	22	29	0	0.76	1.00	0.86
Relative value	7	9	0	0.78	1.00	0.87
Other value	0	12	0	undef	undef	
Date	8	14	0	0.57	1.00	0.72
Other time expression	0	1	0	undef	undef	
Total markable objects	341	446	19			
Spuriously marked: 68						

Table 9.6: Detailed results for the MUC6 training set, obtained using the improved old system

The data in the detailed breakdown shows that all the categories have improved, with the exception of premodifiers in compounds and heads of relative clauses, which have stayed the same.

There is a very big improvement in the categories which required syntax and pattern matching to work and where the implementation of surface rules has definitely made a difference. These categories are noun phrases inside appositive constructions (rise from 56 to 97 F-value), noun phrases in ternary copula constructions (rise from 31 to 91 F-value) and noun phrases as objects of copula verbs (rise from 17 to 93 F-value). 'Function and value' co-references have also improved (these are the categories such as absolute value, relative value and other value).

Adding these two groups of categories together the old MUC6 system made correct connections (in the sense described in chapter 8, section 8.3.2) for 15 out of 102 items in the 15 training texts. The improved old system, on the other hand, correctly connected 77 out of 102 items.

Resolution of pronouns and noun phrases has also improved. Considering together all the noun phrases such as NPs with definite article, NPs with indefinite article, NPs with possessive determiner, NPs with demonstrative determiner, NPs with Saxon genitive determiner and heads of relative clauses, the old system correctly connected 24 out of 99 items, while in the improved old system this figure has risen to 68.

Pronouns and possessive determiners have improved too: their recall has risen from 21 items out of 54 (39%) to 42 items out 54 (78%). The precision has risen from 50% to 88%. The new F-measure for pronouns is 82.

This is consistent with the view that the old system's algorithm for noun phrase and pronominal anaphora resolution was good, but the problems in other modules of the system didn't always allow it to be triggered.

9.2.2 Analysis by categories of the *Guardian* article

Tables (9.7) and (9.8) show a detailed breakdown of the scores obtained for the *Guardian* article. The categories not represented in the article have been omitted.

CATEGORY	RESP	KEY	OVG	REC	PREC	F-VAL
Abbreviated name	11	13	0	0.85	1.00	0.91
Full name	25	33	0	0.76	1.00	0.86
Acronym	1	1	0	1.00	1.00	1.00
Inside appositive	1	3	1	0.33	0.50	0.39
Premod in compounds	3	5	0	0.60	1.00	0.75
NP with definite article	5	14	0	0.36	1.00	0.52
NP with indefinite article	0	3	1	undef	undef	
Head of relative clause	0	1	0	undef	undef	
NP with possessive det	3	6	0	0.50	1.00	0.66
NP with Saxon genitive det	0	2	0	undef	undef	
Sing noun with no det	0	5	0	undef	undef	
Plural noun with no det	0	1	0	undef	undef	
Possessive det	6	9	0	0.67	1.00	0.80
Ordinary pronoun	10	18	3	0.56	0.77	0.64
Absolute value	0	2	0	undef	undef	
Other value	0	4	0	undef	undef	
Totals for markable objects	65	120	5			
Spuriously marked: 21						

Table 9.7: Breakdown of the co-references scores obtained for the *Guardian* article using the LOLITA system from the time of MUC6

CATEGORY	RESP	KEY	OVG	REC	PREC	F-VAL
Abbreviated name	13	13	0	1.00	1.00	1.00
Full name	29	33	0	0.88	1.00	0.93
Acronym	0	1	0	undef	undef	
Inside appositive	3	3	0	1.00	1.00	1.00
Premod in compounds	5	5	0	1.00	1.00	1.00
NP with definite article	10	14	1	0.71	0.91	0.79
NP with indefinite article	1	3	0	0.33	1.00	0.49
Head of relative clause	0	1	0	undef	undef	
NP with possessive det	4	6	0	0.67	1.00	0.80
NP with Saxon genitive det	0	2	0	undef	undef	
Sing noun with no det	2	5	1	0.40	0.67	0.50
Plural noun with no det	0	1	0	undef	undef	
Possessive det	8	9	1	0.89	0.89	0.89
Ordinary pronoun	15	18	0	0.83	1.00	0.90
Absolute value	0	2	0	undef	undef	
Other value	0	4	0	undef	undef	
Totals for markable objects	90	120	3			
Spuriously marked: 20						

Table 9.8: Breakdown of the co-reference scores obtained on the *Guardian* article using the improved old system

As in other tests, all categories show a rise in score. The most notable rises are in the category of noun phrases with definite article and in pronouns (both possessives and ordinary ones).

9.2.3 Analysis by categories of the semi-blind tests A and B

Tables (9.9) and (9.10) (pages 139 and 140, respectively) show a detailed breakdown of the scores obtained for the semi-blind test A, first using the system from the

time of MUC6 and then using the improved old system.

Tables (9.11) and (9.12) (pages 141 and 142, respectively) show a detailed breakdown of the scores obtained for semi-blind test B, first using the old MUC6 system and then using the improved old system.

CATEGORY	RESP	KEY	OVG	REC	PREC	F-VAL
Abbreviated name	16	20	0	0.80	1.0	0.88
Full name	21	42	4	0.50	0.8	0.61
Acronym	0	2	0	undef	-	
Inside appositive	8	15	0	0.53	1.0	0.69
Premod in compounds	4	13	0	0.31	1.0	0.47
NP inside ternary copula	0	2	1	undef	undef	
NP with demonstrative det	0	1	0	undef	undef	
NP with definite article	14	36	4	0.39	0.7	0.50
NP with indefinite article	0	4	1	undef	undef	
NP object of copula verb	0	0	0	-	-	
Head of relative clause	0	0	0	-	-	
NP with possessive det	5	11	1	0.45	0.8	0.57
NP with Saxon genitive det	2	6	0	0.33	1.0	0.49
Sing noun with no det	2	12	2	0.17	0.5	0.25
Plural noun with no det	1	4	0	0.25	1.0	0.40
Head of NP	0	0	0	-	-	
Possessive pronoun	0	0	0	-	-	
Possessive det	12	17	1	0.71	0.9	0.79
Ordinary pronoun	8	18	5	0.44	0.6	0.50
Demonstrative pronoun	0	0	0	-	-	
Referential adverbial	0	1	0	undef	undef	
Percentage	0	0	0	-	-	
Absolute value	0	4	0	undef	undef	
Relative value	0	2	0	undef	undef	
Other value	0	1	0	undef	undef	
Date	0	0	0	-	-	
Other time expression	0	3	0	undef	undef	
Total markable objects	93	214	19			
Spuriously marked: 29						

Table 9.9: Detailed results for **test A**, obtained using the LOLITA system from the time of MUC6

CATEGORY	RESP	KEY	OVG	REC	PREC	F-VAL
Abbreviated name	18	20	0	0.90	1.00	0.94
Full name	27	42	3	0.64	0.90	0.74
Acronym	0	2	0	undef	undef	
Inside appositive	8	15	0	0.53	1.00	0.69
Premod in compounds	5	13	0	0.38	1.00	0.55
NP inside ternary copula	1	2	0	0.50	1.00	0.66
NP with demonstrative det	0	1	0	undef	undef	
NP with definite article	24	36	3	0.67	0.89	0.76
NP with indefinite article	1	4	0	0.25	1.00	0.40
NP object of copula verb	0	0	0	-	-	
Head of relative clause	0	0	0	-	-	
NP with possessive det	5	11	0	0.45	1.00	0.62
NP with Saxon genitive det	3	6	0	0.50	1.00	0.66
Sing noun with no det	6	12	1	0.50	0.86	0.63
Plural noun with no det	2	4	0	0.50	1.00	0.66
Head of NP	0	0	0	-	-	
Possessive pronoun	0	0	0	-	-	
Possessive det	13	17	3	0.76	0.81	0.78
Ordinary pronoun	11	18	2	0.61	0.85	0.71
Demonstrative pronoun	0	0	0	-	-	
Referential adverbial	0	1	0	undef	undef	
Percentage	0	0	0	-	-	
Absolute value	1	4	0	0.25	1.00	0.40
Relative value	1	2	0	0.50	1.00	0.66
Other value	0	1	0	undef	undef	
Date	0	0	0	-	-	
Other time expression	0	3	0	undef	undef	
Total markable objects	126	214	12			
Spuriously marked: 30						

Table 9.10: Detailed results for **test A**, obtained using the improved old system

CATEGORY	RESP	KEY	OVG	REC	PREC	F-VAL
Abbreviated name	13	17	0	0.76	1.00	0.86
Full name	23	44	3	0.52	0.88	0.65
Acronym	0	1	0	undef	undef	
Inside appositive	6	16	0	0.38	1.00	0.55
Premod in compounds	2	6	0	0.33	1.00	0.49
NP inside ternary copula	3	21	1	0.14	0.75	0.23
NP with demonstrative det	0	2	0	undef	undef	
NP with definite article	2	22	1	0.09	0.67	0.15
NP with indefinite article	0	0	0	-	-	
NP object of copula verb	0	0	0	-	-	
Head of relative clause	0	0	0	-	-	
NP with possessive det	2	3	0	0.67	1.00	0.80
NP with Saxon genitive det	0	0	0	-	-	
Sing noun with no det	2	17	1	0.12	0.67	0.20
Plural noun with no det	0	7	0	undef	undef	
Head of NP	1	4	0	0.25	1.00	0.40
Possessive pronoun	0	0	0	-	-	
Possessive det	10	17	1	0.59	0.91	0.71
Ordinary pronoun	7	13	0	0.54	1.00	0.70
Demonstrative pronoun	0	0	0	-	-	
Referential adverbial	0	0	0	-	-	
Percentage	0	0	0	-	-	
Absolute value	0	9	0	undef	undef	
Relative value	0	5	0	undef	undef	
Other value	0	6	0	undef	undef	
Date	0	0	0	-	-	
Other time expression	0	0	0	-	-	
Total markable objects	71	210	7			
Spuriously marked: 67						

Table 9.11: Detailed results for **test B**, obtained using the LOLITA system from the time of MUC6

CATEGORY	RESP	KEY	OVG	REC	PREC	F-VAL
Abbreviated name	15	17	0	0.88	1.00	0.93
Full name	30	44	2	0.68	0.94	0.78
Acronym	1	1	0	1.00	1.00	1.00
Inside appositive	11	16	0	0.69	1.00	0.81
Premod in compounds	2	6	0	0.33	1.00	0.49
NP inside ternary copula	17	21	1	0.81	0.94	0.87
NP with demonstrative det	1	2	0	0.50	1.00	0.66
NP with definite article	14	22	2	0.64	0.88	0.74
NP with indefinite article	0	0	0	-	-	
NP object of copula verb	0	0	0	-	-	
Head of relative clause	0	0	0	-	-	
NP with possessive det	1	3	0	0.33	1.00	0.49
NP with Saxon genitive det	0	0	0	-	-	
Sing noun with no det	8	17	0	0.47	1.00	0.63
Plural noun with no det	5	7	0	0.71	1.00	0.83
Head of NP	1	4	0	0.25	1.00	0.40
Possessive pronoun	0	0	0	-	-	
Possessive det	14	17	1	0.82	0.93	0.87
Ordinary pronoun	13	13	0	1.00	1.00	1.00
Demonstrative pronoun	0	0	0	-	-	
Referential adverbial	0	0	0	-	-	
Percentage	0	0	0	-	-	
Absolute value	7	9	0	0.78	1.00	0.87
Relative value	5	5	0	1.00	1.00	1.00
Other value	2	6	0	0.33	1.00	0.49
Date	0	0	0	-	-	
Other time expression	0	0	0	-	-	
Total markable objects	147	210	6			
Spuriously marked: 63						

Table 9.12: Detailed results for **test B**, obtained using the improved old system

Looking at the detailed scores in all the tables for tests A and B, the picture is similar to that described for the training test set of 15 articles. In the evaluation of the improved MUC6 system, scores in all categories have risen with the exception of the heads of relative clauses (in test B), which remained the same. However, the rise in tests A and B is not as marked as that found in the training set.

The biggest gain in both tests is in the category of noun phrases with a definite determiner. In test A, the number of correctly connected noun phrases with a definite article rose from 14 to 24 (out of a total of 36 markable noun phrases). In test B, the rise in this category is from 2 to 14 (out of possible 22).

Test B shows a slightly better improvement on pronouns than test A: in test A the number of correctly connected pronouns rose from 20 to 24 (out of possible 35), while in test B the rise was from 17 to 27 (out of possible 30).

9.3 Discussion of results

Taking all the results from all test sets together, it appears that a large proportion of the rise in scores is due to the categories involving "is_a" based co-references and "function and value" co-references.

Considering the figures from all the tables, in the output of the original MUC6 system only 33 out of 109 noun phrases in the is_a based co-references were connected correctly. Additionally, 11 incorrect links were made. This constitutes recall of 30% and precision of 75% (F=43). The score for these categories from all the tables in the improved output would add up to recall of 81% and precision of 97% (F=81).

A similar scale of improvement can be found in the "function and value" co-references. The old system made no such links. The improved system made 45 out of 83 possible ones (with no incorrect links), which gives figures of 54% recall and 100% precision (F=70).

The next largest contribution to the rise is made by other definite noun phrases (i.e. **not** those involved in "is_a" based co-references). The improvement in pronouns is also noteworthy.

Again, taking the data from all the tables together, the (non-is_a) noun phrases in the old MUC6 system scored 27% recall and 84% precision (F=41). In the improved system the score rose to 63% recall, 92% precision (F=75).

The pronouns produced better results in the old MUC6 system than other categories. Considering the data from all the tables, in the old system 51% of all pronouns where correctly connected, with precision of 82% (giving F-value of 63). In the improved system, this score has risen to 79% recall, 90% precision (F=84).

This is an interesting result, because no changes to the core of the resolution algorithm were carried out before these results were obtained. It is evident that the preliminary improvements in the parsing and Named Entity recognition, etc., have had a very positive impact on the co-reference score. Now that the system is analysing more of the input (and not losing a lot of information due to parsing failures) the mechanism for resolving noun phrases and pronouns can come into play. It is likely that in the old MUC6 system, a lot of noun phrases and pronouns simply didn't make it to the resolution stage. Moreover, even if they did, their correct antecedents could have been missing from the context, due to the same basic problems.

Chapter 10

Resolving References in LOLITA: the new system

This chapter describes the revised and improved procedure for handling anaphora in the LOLITA system. Qualitative advances and innovative improvements have been made in three main areas:[1]

1. the structure and the handling of Context
2. the rules of searching for an antecedent
3. the pragmatic preference system.

The structure of Context has been changed to allow it to include several new types of information. The way the Context is used during processing has also been considerably improved.

The initial filtering rules which search the Context for possible antecedents have been substantially overhauled.

[1] All the changes described here have been designed, implemented, tested and debugged by the author.

The pragmatic preference system has undergone major design changes; its component heuristics have been expanded and the way they apply has been improved.

A detailed description of the improvements follows in the subsequent sections.

10.1 The main algorithm

The general idea behind the reference resolution algorithm remains the same as described in chapter 6. It can be summarised again thus:

As the discourse is processed, the referents found in it are stored in the Context buffer. Each time an anaphoric expression is identified in the incoming discourse, the system looks for a possible referent for this expression in the Context (obeying matching rules dictated by the type of anaphor). If the system finds no match, it introduces a new entity into the Context. If the system finds just one match, it unifies the two and adds the newly unified item in the Context. If the system finds more than one match, it builds a special structure to represent the ambiguity and passes it onto the system of preference heuristics to decide between the possibilities.

10.2 Improved Context structure

The following part of the Context buffer has been redesigned:

- REFERENTS : all the recent concepts (i.e. the ones within a defined cut-off) that have been mentioned so far by the source, the so called 'normal referents';

The cut-off point remains the same as that described in section 6.1.1, in chapter 6 (p. 75).

From now on, this part of the Context will be referred to as NORMAL REFERENTS.

The following new parts have been added:

- INDIVIDUAL REFERENTS: named individuals which have appeared in the text at any time
- FOCUS REFERENTS: a list of referents in focus in the current and immediately preceding utterance
- CURRENT REFERENTS: a list of lists of local referents, kept separate during processing for strictly grammatical reasons
- GLOBAL FOCUS REFERENTS: a list of named individuals which are likely to be the topic or global focus of the text; currently this list includes those named individuals whose names appeared in the first two sentences of the text

Also, an important new feature has been introduced: the Context now records information regarding the type of referent a given item has been mentioned with: e.g. was is a personal pronoun, was it a relative pronoun, was it a noun phrase, etc. This allows the new system to take into account the fact that different types of anaphoric expressions may trigger different rules.

10.2.1 NORMAL REFERENTS

The system of storing mentioned concepts in the list of NORMAL REFERENTS has been changed. Previously, if an anaphoric expression in the discourse was unified with an item from the Context, the newly unified node was moved to the top of the list of NORMAL REFERENTS. However, because not all ambiguities get resolved incrementally (some are resolved at the end of the sentence), this approach results in the re-ordering of the Context for some ambiguities and not others. As a result problems such as those discussed in chapter 6 (cf. example (46), page 97) arise.

In the new system, the NORMAL REFERENTS are never re-ordered. Each time a concept is mentioned in the discourse, an item which represents that mention is placed in the list of NORMAL REFERENTS and marked *New*. At the same time, all other mentions of the same node which are already in the list of NORMAL REFERENTS are marked *Old*.

The new design ensures that the original order of referents with respect to each incoming anaphoric expression is preserved.

10.2.2 INDIVIDUAL REFERENTS

Some psychological evidence (particularly the experiments of Sanford *et al.* (1988)) suggests that referents introduced into the discourse by a proper name stay "focused" in the hearer's/reader's memory the longest. It has therefore been decided to keep a separate list of all Named Individuals mentioned in the discourse, to make them available for reference (however, only via a proper name) in any part of subsequent discourse.

10.2.3 FOCUS REFERENTS

This section of the Context buffer has been introduced to record as NEW FOCUS any item which is a grammatical subject and is in the principal clause of the current sentence. The grammatical subject(s) from the previous sentence's principal clause are also stored in this list but are marked as OLD FOCUS.

The term *focus* is not intended to reflect some particular theory of discourse focus. It is used here to describe the items that are regarded as having some defined measure of saliency.

The particular measure of saliency adopted is however inspired by some psychological evidence pointing to the importance of subjects and particularly of 'first men-

tioned' subjects in the processing of pronouns (Sanford & Garrod (1981), Gernsbacher & Hargreaves (1988)).

10.2.4 CURRENT REFERENTS

The introduction of CURRENT REFERENTS into the Context buffer is a way of modelling the notion of *governing category* of GB-theory (Chomsky, 1981), useful for handling intra-sentential anaphora.

Referents found within boundaries of selected syntactic categories are kept in separate lists. For example, all the referents in a simple utterance in (57) would be kept in the same CURRENT CONTEXT:

(57) *Paul bought himself a new motorbike.*

On the other hand, in the following utterance:

(58) *Paul's granddad adores him.*

two levels of CURRENT REFERENTS will be created: one containing *Paul's granddad* and the unresolved referent of *him* and a separate level, created for the genitive noun phrase, and containing *Paul's granddad* and *Paul.*

The notion of CURRENT REFERENTS is used in the interpretation of reflexives, definite pronouns and definite noun phrases (see below for more detailed examples).

10.3 New search rules for particular types of anaphora

10.3.1 Reflexive pronouns

10.3.1.1 Reflexives used referentially

The new rule for the interpretation of reflexives used referentially states that an antecedent for a reflexive must be found within the CURRENT REFERENTS.

So, in the example (57), the referent for *himself* has to be *Paul.* In a more complex case, where the Context contains more than one male referent, hence there are two potential antecedents for *himself*, again only *Paul* is allowed to match, being the only referent that is found in the same CURRENT CONTEXT as *himself.*

(59) *Having talked to Michael, Paul bought himself a new motorbike.*

Other, successfully handled examples include those which involve an interaction of relative clause analysis and reflexive pronoun resolution, e.g.:

(60) *Feminists, who detest the actresses who reproach themselves, are happy.*

Here, the second relative pronoun *who* is unified with *the actresses* and so will be found in the CURRENT CONTEXT created by the clause *who reproach themselves*, ensuring that *the actresses* are the only possible antecedent for *themselves.*

The above treatment of reflexives works fine, as long the parsing is correct and appropriate semantic rules are in place. However, when there is a problem with either of the latter two, the reflexives may not be resolved correctly.

For example, contrast the following cases:

(61) *Paul told Michael to wash himself.*
Paul promised Michael to wash himself.

where in the first case *himself* should be *Michael* while in the second it should be *Paul*, is not analysed successfully at present. There is a problem with the parsing of the above structures: in both cases the parsing is the same and is, in fact, incorrect for both. This situation leads to an erroneous semantic analysis and to each reflexive being resolved to some unknown human male.

However, once the parsing and the semantics can be corrected, the application of the CURRENT CONTEXT solution in these two cases would be very easy.

It should be noted that the rule for reflexives implemented in the new system does not take into account the *c-command* relation that is predicted to hold between the antecedent and the reflexive (cf. the GB theory). Theoretically, there could be cases where a reflexive was found within the same CURRENT CONTEXT as some non-c-commanding entity, and so the latter could then be ruled out as a possible antecedent. In practice, however, this doesn't appear to happen very often, and when it does, co-reference is not necessarily ruled out. For instance, in (62):

(62) *Paul chatted to Jenny about herself.*

Jenny is in the same CURRENT CONTEXT as *herself*, does not c-command *herself*, yet the co-referential reading is OK.[2]

So, for current purposes, it has been decided to proceed without the use of the notion of *c-command.*

[2]Some might argue that in this example the use of *herself* is emphatic and therefore does not necessarily obey standard binding rules.

10.3.1.2 Reflexives used for emphasis

There have been no special rules developed for handling cases like the following:

(63) a. *Paul chatted to Jenny himself.*
b. *Paul chatted to Jenny herself.*
c. *Paul himself chatted to Jenny.*

or, arguably, a bit more difficult:

(64) *Paul decided to resign. No one but himself could make the decision.*

LOLITA currently resolves the reflexives correctly in examples such as (63a) and (63b), however, the overall semantic analysis of these examples is not satisfactory.

In the example (63c), erroneous parsing leads to a wrong resolution. However, once that is corrected, the analysis should improve.

With (64) the problem is more complicated, as the reflexive here appears in a sentence on its own, without the antecedent. The analysis of such usages would require a separate treatment, possibly quite different from the rules used for the other examples discussed.

10.3.2 Non-reflexive, definite personal pronouns

10.3.2.1 Third person pronouns

The rule for third person pronouns (*he/him, she/her, they/them, it*) states that an antecedent must be found outside the CURRENT REFERENTS and must exclude any antecedent found in the CURRENT REFERENTS.

Thus, in the example (58), repeated here:

(58) *Paul's granddad adores him.*

the referent for *him* can only be *Paul* and not *Paul's granddad*, as the latter was found in the same CURRENT CONTEXT as *him*.

Additionally, if the pronoun is in a main clause, possible antecedents must also exclude any referents that appear as proper names in a subordinate clause, within the same sentence. This condition aims to account for the majority of cases where a pronoun c-commands a proper name and where co-reference is therefore normally ruled out.

So, in the following example (already discussed in chapter 2 and repeated here):

(9) *He said that Michael blamed Paul.*

the *He* cannot refer to either *Michael* or *Paul* and LOLITA doesn't make such a co-reference. (NB.: Given example (9) in isolation the co-reference would not be made in any case, because there is no treatment of cataphora developed at present. However, assuming that *Michael* and *Paul* were mentioned in some preceding discourse, it is the 'main-clause condition' that will exclude them both from being potential antecedents for *He*.)

The condition might over-generalise in some cases. For example, given the following:

(65) *Paul chatted to Jenny. He discussed with her the film which Jenny had recently seen.*

the pronoun *her* is in the main clause and the name *Jenny* follows it, so the co-reference between the two is excluded — counter to intuition. The full implementation of the notion of *c-command* would solve the problem here, however, the current implementation appears to be as good as can be achieved, given LOLITA's syntactic representation and the existing interface between syntax and semantics.

10.3.2.2 First and second person pronouns

There are separate rules in LOLITA to handle first and second person pronouns. Among these, the rules for *I* and *you* have existed in the system before. They state that the pronoun *I* must refer to the source of the utterance/discourse, while *you* must invariably refer to LOLITA (unless the system is in the dialogue mode). This treatment may not seem very satisfactory and will be included in the list of future improvements.

To handle the first person plural pronoun *we* a new rule has been added to the algorithm. Apart from the usual restriction to do with CURRENT REFERENTS (which are the same here as for third person pronouns) the algorithm searches for the following possibilities, in this order:

- any straight-forward, first person plural referents, e.g. a previous mention with the use of *we*
- any sets of human entities, mentioned with the use of conjunction, e.g. *Paul and Michael* or *BMW and Rolls-Royce*
- any single mentions of companies
- any sets of people mentioned separately.

If none of the above possibilities are found, a new, first person plural referent is built.

Thus, in the example (4), discussed in chapter 2, and repeated here:

(4) *BMW is Germany's biggest auto-maker. "We employ more than a thousand workers", said the company spokesman.*

LOLITA would resolve the *we* to be *BMW*. While in this case:

(66) *Paul talked to Michael. "We discussed the new strategy", Michael said.*

the *we* in the second utterance would be resolved to *Michael and Paul*, a set of two individuals, built by the system 'on the spot' during the resolution of *we*.

Both the first and second person pronouns are more likely to be used in dialogue or in the reporting of direct speech, so their interpretation may be particularly sensitive to the handling of those kinds of discourse. However, the aspects of LOLITA relevant to direct speech are currently under development. Consequently, the interpretation of first and second person pronouns may require further improvements.

10.3.3 Possessive pronouns

10.3.3.1 Possessives used as determiners

Third person possessive pronouns: *his, her, its, their*, when used as determiners in noun phrases, look for a referent from among the NORMAL REFERENTS. In this case, antecedents found in the CURRENT REFERENTS are not usually excluded straight away, but are penalized during later processing (i.e. during the application of pragmatic heuristics, discussed below).

10.3.3.2 Possessives used with ellipsis

No new rules for possessive pronouns used on their own, i.e. with their object elided, have been developed in the new system. The current treatment of these in LOLITA requires more attention. For example, in cases like these:

(67) *Paul mislaid his pen. And Michael mislaid his.*

assuming the most usual interpretation, the existence of two distinct pens is implied

— one belonging to *Paul* and the other to *Michael.*[3] In other words, the *his* in the second utterance should be seen as a pronoun standing for *Michael's pen.* However, in the current state, the system misinterprets the second clause.

In the MUC6 corpus, possessive pronouns used on their own are extremely rare. While it would be desirable, in principle, to have a more satisfactory way of handling them, as far the MUC6 task is concerned, the lack of good rules is not seriously detrimental.

However, because LOLITA is a general purpose system, new rules for possessives must be developed, whether or not such expressions occur in the MUC corpus.

10.3.3.3 Demonstratives

Demonstrative pronouns *this, that, these* and *those* search for a referent from among NORMAL REFERENTS, with the exclusion of CURRENT REFERENTS. In the majority of the texts used in this project they are extremely rare.

10.3.4 Feature matching for pronominal anaphora

The rules discussed in sections 10.3.1 – 10.3.3 deal mainly with the problem of finding the pool of referents in the Context from which an anaphoric expression can choose an antecedent, without violating grammatical rules.

After the initial pool is selected, all candidates with non-matching features are filtered out. Personal pronouns with the feature `Human` are only allowed to match items in the Context which also have the feature `Human`. Pronouns with feature `Male` and `Female` are allowed to match any items with the corresponding gender.

The pronouns *it* and *its*, despite not carrying the feature `Human`, are allowed to

[3] A rarer interpretation would be that Michael mislaid Paul's pen, or even some other male entity's pen

refer to items marked [Human, Sing, Neutral] — a feature combination usually associated with nodes representing companies.

The above feature matching scheme has existed in the system before. The new system has been given one additional rule: now, pronouns carrying a specific gender are also allowed to match items marked Sexed. This allows for co-references to be made in cases where an antecedent has not got a defined gender, e.g. it is an ambiguous proper name like *Leslie*, or it is a noun phrase like *the doctor*.

10.3.5 Noun phrases

10.3.5.1 Definite noun phrases

Several aspects of the noun phrase matching have been improved. A rule has been implemented to bar noun phrases from referring to c-commanding antecedents, e.g. for cases such as:

(68) *Paul likes the man.*

where *Paul* and *the man* are extremely unlikely to co-refer.

Similarly, a rule has been implemented to disallow co-references between a proper name and a c-commanding noun phrase, as in the case like this one:

(69) *The man likes Paul.*

Some aspects of the matching of definite noun phrases against other definite noun phrases or against proper names have been revised.

Noun phrases referring to humans are allowed more freedom than before when it comes to choosing the initial pool of possible antecedents. Now the cases such as

those discussed in chapter 6, section 6.3.7 are resolved in a more common sense manner. So, if for example there is knowledge available that an entity in the Context is *a producer* and then the discourse contains the noun phrase *the producer*, the co-reference between the two is favoured.

Moreover, within the family of `humans` or `human organisations`, siblings in the hierarchy are now allowed to match. This relaxation of the matching rules solves the problem described earlier, whereby extra knowledge about a human entity prevented it from being matched by some fairly general noun phrases often used to refer to humans (e.g. *the guy*).

The difficulty with matching noun phrases which all fall within the hierarchy describing humans stems from the fact that terms such as `producer`, `director`, `saviour`, etc., which are siblings in the hierarchy, are not mutually exclusive — unlike siblings in other hierarchies (e.g. within `mammals`, sister terms such as `cats` and `dogs` **are** mutually exclusive). This could be seen as a problem of the *is_a* hierarchy itself, however, its proper solution lies outside the scope of this project. The solution which uses the existing hierarchy to make co-reference links, for current purposes appears to work well on the corpus tested. However, tests on other corpora with different topics would be beneficial. It seems possible that the matching rules might then need tightening.

10.3.6 Proper names

Proper names are not allowed to match against the concept of `name`. A new feature of proper name matching is also the fact that proper names can now find antecedents that appeared relatively far back in the discourse. This is because all named individuals are now stored in a separate part of the Context buffer (INDIVIDUAL REFERENTS, described in section 10.2.2) and so they are always available for subsequent reference.

10.3.7 Abbreviations

There has been a small improvement in the way proper names are matched with their abbreviations. Now the abbreviated name components are considered as matches when they are either shorter or they are synonyms of the equivalent component of the full proper name. This allows for a match to be made between *American Telegraph & Telephone Co* and *AT&T Corp.*

10.4 The heuristics system redesigned

The following sections describe the changes made to those parts of the algorithm which deal with choosing one out of several possible candidate referents, after the matching filters described above have been applied.

10.4.1 Major design changes

The first fundamental change to the design of the algorithm is a change of the heuristic preferences system: the heuristics no longer act as filters. Instead, they award penalty points to any non-preferred items.[4] Once all preference rules have been applied, the candidate with the least number of points is chosen (a blackboard architecture approach).

Another major design change which has been implemented is the separation of heuristics into those which deal with sense disambiguation, those which deal with anaphora resolution and those which can be seen as relevant to both.

The following heuristics now affect only sense disambiguation:

[4]The use of a penalty-point system, as opposed to a bonus-point system, maintains consistency with other modules of LOLITA (e.g. parsing), which also employ penalties.

- prefer_topic
- prefer_common
- prefer_countable
- prefer_more_connections.

The following rules affect both:

- *prefer_object*
- *prefer_human.*

The remaining heuristics, together with several new additions apply only to anaphora resolution.

10.4.2 New heuristics added

New heuristics, designed specifically for anaphora resolution have been added and these include:

- Prefer a referent that is present in the NEW FOCUS part of the Context to those present in the OLD FOCUS; but prefer the latter to those not present in the FOCUS part at all (*prefer_focus*).
- Prefer a referent which is seen as being in the global focus of the discourse (*prefer_global_focus*).
- Prefer a referent whose most recent mention (with respect to a given anaphoric expression) was as a grammatical subject (*prefer_gram_subject*). [5]

[5]NB this rule differs from that of *prefer_main_role*, described in section 6.1.4 (p. 81), in that it uses the term *subject* purely syntactically; by contrast, for purposes of the *prefer_main_role* heuristic the subject is the (semantic) agent in the sentence.

- Prefer a referent with the same thematic role as the anaphoric expression (*prefer_parallel_role*).
- Prefer a referent that occurred in the same sentence as the anaphoric expression (*prefer_same_sentence*).
- Prefer an explicitly mentioned referent to an implicit one (*prefer_explicit*).
- If the referent is an event implied by the current sentence or is an action of the current sentence, assume that reference to it within the current sentence is very unlikely; additionally, penalize this referent particularly heavily if the anaphoric expression is a possessive pronoun (*prefer_distant_event*).
- Prefer a candidate whose gender matches that of the anaphoric expression exactly, in other words penalize a candidate whose gender is compatible but not the same (*prefer_gender_agreement*).
- Prefer a candidate designating a non-temporal entity, i.e. penalize references to time expressions (*prefer_non_temporal*).

A new heuristic dealing with sense disambiguation has been added:

- Prefer a meaning whose family has a cumulatively closer distance to the families of the nodes in the NORMAL REFERENTS of the Context (*prefer_shorter_family_distance*).

10.4.3 Motivation for the new heuristics

The new heuristics dealing with anaphora resolution use knowledge which is more discourse based. They have been inspired by several sources, including findings of psycholinguistic experiments, other AI approaches to the anaphora problems, as well as intuition.

10.4.4 Old heuristics improved

The following three heuristics have been improved:

- *prefer_last_mention*
- *prefer_main_role*
- *prefer_human.*

Now the *prefer_last_mention* and *prefer_main_role* heuristics consider a more relevant "window" of the NORMAL REFERENTS with respect to the anaphoric expression they are applied to. That is, the subset of the NORMAL REFERENTS considered consists of concepts that appeared before the anaphoric expression.

Also, these two rules no longer affect sense disambiguation.

The *prefer_human* heuristic has been revised so that it grades the possible meanings or referents according to how far removed they are (in terms of LOLITA family hierarchy) from the family **human**. Previously, only the node closest to the family **human** would have been selected, while others were discarded.

10.4.5 Weights system for anaphora resolution

The initial preference tests look at how well a possible antecedent might fit in with LOLITA's knowledge of the world. That is, for each event being examined, the candidates for antecedents are considered with respect to the prototype associated with the current event's action. These tests check to see if there might be a major clash with the requirements of the prototype. If so, the penalties given are comparatively large and aim to exclude such candidates from further consideration.

The actual weights used for the remaining heuristics are manually engineered. They have been set mainly on the basis of examples encountered in the MUC6 training

data and in the single article from British press. Given more time and resources, it would interesting to use an automatic way of setting the weights. Possibly, some form of evolutionary algorithms could be applied to the problem, especially given the fact that this kind of approach has been used to set weights in LOLITA before (in the dialogue module), with positive results (Nettleton 1995).

10.4.6 Weights system at work: an example

Once the system has chosen a set of possible candidate antecedents for a given anaphoric expression, each candidate is examined and rated according to each of the preference heuristics. The clearest way to demonstrate how the heuristics operate is with the use of examples.

10.4.6.1 Example 1

This section shows how the system analyses the anaphoric expression *It* of the second utterance in the example from section 6.3.1, (p. 95), and repeated here:

(42) *To discuss the script, a meeting of all the staff was organized. It was very productive.*

LOLITA initially considers the following candidates for the pronoun:

```
96176 : script
96177 : discussing
13456 : discuss
96192 : organizing
74238 : organize
96190 : meeting
```

These are all the nodes which were present in the preceding Context and which had the features `Singular` and `Neutral`.

The first rules to apply assess all the candidates with respect to their compatibility with the predicate *productive*, which had already been disambiguated. This predicate (currently) imposes no restriction on its arguments, so all the candidates pass with no penalty.

The first heuristic to apply with a visible effect is the one favouring objects over events (*prefer_object*).[6] It awards penalties to the following nodes: `discussing`, `organizing` and `meeting`. The penalty awarded by the rule is shown below as "pen given". The final column shows the total penalty level collected so far after the current rule's application.

```
96176 : script          - pen given: 0    -> total: 0
96177 : discussing      - pen given: 2100 -> total: 2100
13456 : discuss         - pen given: 2100 -> total: 2100
96192 : organizing      - pen given: 2100 -> total: 2100
74238 : organize        - pen given: 2100 -> total: 2100
96190 : meeting         - pen given: 2100 -> total: 2100
```

Next, the *prefer_explicit* heuristic further penalizes all the events which were implicit in the discourse. These include `organizing` and `discussing` — nodes built by LOLITA on the basis of the verbs used in the discourse. Also, the actions `organize` and `discuss` are penalized here. By contrast, the concept of `meeting` which was actually mentioned in the text isn't penalized. After this heuristic has applied, the penalties appear as follows:

```
96176 : script          - pen given: 0     -> total: 0
96190 : meeting         - pen given: 0     -> total: 2100
74238 : organize        - pen given: 10000 -> total: 12100
96192 : organizing      - pen given: 10000 -> total: 12100
13456 : discuss         - pen given: 10000 -> total: 12100
96177 : discussing      - pen given: 10000 -> total: 12100
```

The next rule to apply is *prefer_focus*. The previous sentence mentioned *a meeting*, which was a grammatical subject of the passive clause. As a result, the node representing the `meeting` was placed in the NEW FOCUS of the CONTEXT and

[6]NB. the heuristics which did not affect any candidates have been omitted from this example.

subsequently moved to OLD FOCUS when the processing of that sentence was completed. By the time the pronoun from the current sentence is being resolved, `meeting` is in the OLD FOCUS and as such gets least penalty points. The NEW FOCUS remains empty at this point, because there are no objects appearing before the *It* in the current sentence. The resulting penalties are now:

```
96190 : meeting          - pen given: 1000  -> total: 3100
96176 : script           - pen given: 15000 -> total: 15000
74238 : organize         - pen given: 18000 -> total: 30100
96192 : organizing       - pen given: 18000 -> total: 30100
13456 : discuss          - pen given: 18000 -> total: 30100
96177 : discussing       - pen given: 18000 -> total: 30100
```

The heuristic, *prefer_gram_subject* awards no penalties to the `meeting` while penalizing all the remaining candidates:

```
96190 : meeting          - pen given: 0     -> total: 3100
96176 : script           - pen given: 12000 -> total: 27000
96177 : discussing       - pen given: 12000 -> total: 42100
13456 : discuss          - pen given: 12000 -> total: 42100
96192 : organizing       - pen given: 12000 -> total: 42100
74238 : organize         - pen given: 12000 -> total: 42100
```

The next heuristic, *prefer_main_role*, examines the thematic roles of the candidates, as they appeared in the most recent events, with respect to the position of the anaphoric expression. It gives no penalties to subjects in the events and penalizes objects slightly; other roles, such as instrument, origin, destination, location_of, etc. are awarded higher penalties. In the current example, after the application of *prefer_main_role* the penalties stand as follows:

```
96190 : meeting          - pen given: 1000 -> total: 4100
96176 : script           - pen given: 1000 -> total: 28000
74238 : organize         - pen given: 4500 -> total: 46600
96192 : organizing       - pen given: 4500 -> total: 46600
13456 : discuss          - pen given: 4500 -> total: 46600
96177 : discussing       - pen given: 4500 -> total: 46600
```

The next heuristics: *prefer_parallel_role* doesn't distinguish between candidates because none of them have a thematic role of which is the same as that of *It*, i.e. of *subject*. Each candidate gets the same penalty:

```
96190 : meeting          - pen given: 10100 -> total: 14200
96176 : script           - pen given: 10100 -> total: 38100
96177 : discussing       - pen given: 10100 -> total: 56700
13456 : discuss          - pen given: 10100 -> total: 56700
96192 : organizing       - pen given: 10100 -> total: 56700
74238 : organize         - pen given: 10100 -> total: 56700
```

The next two heuristics to apply are *prefer_same_sentence* and *prefer_last_mention.* Both of these take into account the distance (in terms of the number of sentence boundaries or concepts in the CONTEXT, respectively) which separate the antecedent from the anaphoric expressions. All the candidates considered here are from the same (immediately previous) sentence, thus the *prefer_same_sentence* heuristic doesn't distinguish between them. On the other hand, *prefer_last_mention* is able to favour the `meeting`, as that was the closest mentioned concept with respect to the pronoun being disambiguated. After the application of these two rules the penalties are now the following: (after *prefer_same_sentence*)

```
96190 : meeting          - pen given: 5000 -> total: 19200
96176 : script           - pen given: 5000 -> total: 43100
74238 : organize         - pen given: 5000 -> total: 61700
96192 : organizing       - pen given: 5000 -> total: 61700
13456 : discuss          - pen given: 5000 -> total: 61700
96177 : discussing       - pen given: 5000 -> total: 61700
```

(*after prefer_last_mention*)

```
96190 : meeting          - pen given: 7200  -> total: 26400
96176 : script           - pen given: 21600 -> total: 64700
74238 : organize         - pen given: 28800 -> total: 90500
96192 : organizing       - pen given: 28800 -> total: 90500
13456 : discuss          - pen given: 28800 -> total: 90500
96177 : discussing       - pen given: 28800 -> total: 90500
```

The chosen referent for the pronoun is thus the concept `meeting`. It has been a clear winner from the time of the application of *prefer_focus.*

The example illustrates the advantage of the new algorithm over the old one. In the old algorithm, this concept is not chosen as a referent because the heuristic *prefer_object* rejects it in the early stage of the process.

10.4.6.2 Example 2

The following example is taken from one of the training texts. The points illustrated in it are not very different from those shown by example 1. However, it is included to demonstrate that the system can deal with the real text much the same way as with a made-up example.

The following excerpt is taken out of the *Guardian* article which was used during training and testing (the full text can be found in Appendix C). The trace for the example shows the weights system applying to all the candidates for the pronoun *He*, shown in bold, from the final sentence of the excerpt.

(70) *William Grosvenor, the entrepreneur and well connected cousin of the Duke of Westminster, who is heading attempts by Texas American Group to take over the troubled Facia retail company, last night admitted that he is bankrupt. (...)*

Facia, headed by Stephen Hinchliffe, operates 850 speciality shops (...).

The DTI investigation into Mr Hinchliffe's affairs are understood to focus on the 1993 collapse of Boxgrey, a company sold by the Sheffield-based entrepreneur shortly before it collapsed.

Mr Grosvenor is known in the City as an entrepreneur who has also worked as a financial public relations adviser. His name regularly appears in newspaper social pages because of his family connections. He is related to the Aga Khan as well as the Duke of Westminster.

He *was married in 1966 to Ellen Seeliger, daughter of Germany's Ambassador to Mexico. (...)*

The system finds several individuals in the Context (all appearing in the excerpt) with the features `Male` or `Sexed` and with the feature `Human`. *Mr Hinchliffe* is one of them, *the Aga Khan* and the *Duke of Westminster*, two of the others. The concepts: *entrepreneur* and *adviser* are also considered because of the system's mis-analysis

of the sentence *Mr Grosvenor is known in the City as an entrepreneur ....* The semantic rules applied to the concepts of *being known as* or *working as* appear to be erroneous. Nevertheless these two concepts are never serious contenders to be analysed as antecedents for *He*, as can be seen below.

The following figure displays the full trace from the application of the preference heuristics.

```
"prefer_proto_child"
 ===>
no changes

"prefer_less_gen"
 ===>
no changes

"prefer_object"
 ===>
no changes

"prefer_explicit"
 ===>
no changes

"prefer_focus"
 ===>
 115902 : william             - pen given: 1000   -> total: 1000
 114717 : duke_of_westminster - pen given: 18000  -> total: 18000
 98530  : aga_khan            - pen given: 18000  -> total: 18000
 116826 : adviser             - pen given: 18000  -> total: 18000
 116815 : entrepreneur        - pen given: 18000  -> total: 18000
 115978 : hinchliffe          - pen given: 18000  -> total: 18000
 116564 : chief_executive     - pen given: 18000  -> total: 18000

"prefer_global_focus"
 ===>
 115902 : william             - pen given: 0      -> total: 1000
 114717 : duke_of_westminster - pen given: 0      -> total: 18000
 116564 : chief_executive     - pen given: 500    -> total: 18500
 115978 : hinchliffe          - pen given: 500    -> total: 18500
 116815 : entrepreneur        - pen given: 500    -> total: 18500
 116826 : adviser             - pen given: 500    -> total: 18500
 98530  : aga_khan            - pen given: 500    -> total: 18500

"prefer_gram_subj"
 ===>
```

```
 115902 : william             - pen given: 0      -> total: 1000
 114717 : duke_of_westminster - pen given: 12000  -> total: 30000
 98530  : aga_khan            - pen given: 12000  -> total: 30500
 116826 : adviser             - pen given: 12000  -> total: 30500
 116815 : entrepreneur        - pen given: 12000  -> total: 30500
 115978 : hinchliffe          - pen given: 12000  -> total: 30500
 116564 : chief_executive     - pen given: 12000  -> total: 30500

"prefer_main_role"
 ===>
 115902 : william             - pen given: 0      -> total: 1000
 115978 : hinchliffe          - pen given: 0      -> total: 30500
 114717 : duke_of_westminster - pen given: 2500   -> total: 32500
 98530  : aga_khan            - pen given: 2500   -> total: 33000
 116564 : chief_executive     - pen given: 4500   -> total: 35000
 116815 : entrepreneur        - pen given: 4500   -> total: 35000
 116826 : adviser             - pen given: 4500   -> total: 35000

"prefer_parallel_role"
 ===>
 115902 : william             - pen given: 10100  -> total: 11100
 115978 : hinchliffe          - pen given: 10100  -> total: 40600
 114717 : duke_of_westminster - pen given: 10100  -> total: 42600
 98530  : aga_khan            - pen given: 10100  -> total: 43100
 116826 : adviser             - pen given: 10100  -> total: 45100
 116815 : entrepreneur        - pen given: 10100  -> total: 45100
 116564 : chief_executive     - pen given: 10100  -> total: 45100

"prefer_same_sentence"
 ===>
 115902 : william             - pen given: 5000   -> total: 16100
 114717 : duke_of_westminster - pen given: 90000  -> total: 132600
 98530  : aga_khan            - pen given: 90000  -> total: 133100
 116815 : entrepreneur        - pen given: 150000 -> total: 195100
 116826 : adviser             - pen given: 150000 -> total: 195100
 115978 : hinchliffe          - pen given: 180000 -> total: 220600
 116564 : chief_executive     - pen given: 480000 -> total: 525100

"prefer_last_mention"
 ===>
 115902 : william             - pen given: 14400  -> total: 30500
 114717 : duke_of_westminster - pen given: 43200  -> total: 175800
 98530  : aga_khan            - pen given: 50400  -> total: 183500
 116826 : adviser             - pen given: 115200 -> total: 310300
 116815 : entrepreneur        - pen given: 129600 -> total: 324700
 115978 : hinchliffe          - pen given: 180000 -> total: 400600
 116564 : chief_executive     - pen given: 626400 -> total: 1151500

"prefer_non_temporal"
 ===>
```

```
no changes

"prefer_gender_agreement"
 ===>
 115902 : william              - pen given: 0     -> total: 30500
 114717 : duke_of_westminster - pen given: 0     -> total: 175800
 98530  : aga_khan            - pen given: 0     -> total: 183500
 116826 : adviser             - pen given: 4000  -> total: 314300
 116815 : entrepreneur        - pen given: 4000  -> total: 328700
 115978 : hinchliffe          - pen given: 0     -> total: 400600
 116564 : chief_executive     - pen given: 4000  -> total: 1155500

"prefer_namedInd"
 ===>
 115902 : william              - pen given: 0     -> total: 30500
 114717 : duke_of_westminster - pen given: 0     -> total: 175800
 98530  : aga_khan            - pen given: 0     -> total: 183500
 116826 : adviser             - pen given: 1000  -> total: 315300
 116815 : entrepreneur        - pen given: 1000  -> total: 329700
 115978 : hinchliffe          - pen given: 0     -> total: 400600
 116564 : chief_executive     - pen given: 1000  -> total: 1156500

"prefer_human"
 ===>
no changes
```

Figure 10.1: Weights system selecting a referent for a pronoun

Once the penalties have been applied, the system delays the decision about the final choice of the antecedents until it processes all the remaining referents in the sentence. If it were the case that a reference to *William Grosvenor* (via the use of proper name) occurred in the subordinate clause of the example sentence, it couldn't be chosen as a referent because of grammatical restrictions. In that case, the next best candidate would be chosen: `duke_of_westminster`.

10.5 Summary

This chapter has described the major advancements made in the area of anaphora resolution within the LOLITA system. The shortcomings of the old system, as

described in chapter 6 (see section 6.3), have been addressed and improvements to the algorithm have been introduced. The system now has a better way of handling the Context, the rules of searching for antecedents have been enhanced, grammatical constraints on anaphora resolution are taken account of, and the pragmatic preferences no longer act as filters but assign weights to possible antecedents.

The idea behind this approach to anaphora resolution is similar to that of Lappin & Leass (1994) in that it assesses various properties of the candidates for anaphora resolution and rates the candidates on that basis (cf. chapter 4, section 4.4.1). However, the algorithm described here differs from that of Lappin & Leass in that it uses a different and wider set of properties and assigns penalties rather than rewards to candidates. Also, the algorithm is designed in such a way that it can apply to a wider range of anaphoric expressions and not just pronouns.

The algorithm differs from approaches such as the one proposed by the University of Massachusetts' system RESOLVE, in that it employs no statistical techniques. It also differs from other, rule-based approaches (such as those of PIE, CogNIAC or FASTUS, cf. chapter 4) in that it attempts to gather information about candidates for anaphora resolution on the basis of a full parse and a deep semantic analysis. In the latter respect, the current system is similar to the LaSIE system (cf. chapter 4, section 4.2.2) developed by the University of Sheffield.

Chapter 11

Performance of the new system

In this chapter the results obtained in the second evaluation of the LOLITA system are presented. In this evaluation the system included all the changes to the anaphora resolution module described in chapter 10.

The results are first compared directly with those obtained using the improved old system. This is followed by an analysis of the breakdown of the scores, with particular attention paid to the categories which are most likely to be affected by the new resolution algorithm.

In the final section of the chapter, a comparison between all three systems is made.

The tables containing the general results are first presented with a brief comment and then discussed all together, in section 11.2.

The tables containing the scores by categories are presented next and compared with the corresponding scores of the improved old system.

The final sections of the chapter consider some interesting cases of co-reference resolved by the system, as well errors made.

11.1 General scores

11.1.1 The formal blind test

The table below shows the scores obtained on the MUC6 formal blind test. The results reveal a small rise, about 2 points in F-value, after the introduction of the new algorithm.

	Recall	Precision	F-measure
Old LOLITA improved	61.9	62.3	62.1
LOLITA with the new algorithm	63.1	64.7	63.9

Table 11.1: Evaluation results for the co-reference task using the LOLITA system on the MUC6 formal (blind) test, before and after the introduction of the new anaphora algorithm

11.1.2 The open tests

Table (11.2) shows the results obtained on all the open tests. On these tests, there is a big, about 9-10 points, rise in F-value.

	Recall	Precision	F-measure
Old LOLITA improved			
MUC6 training set	78.8	79.3	79.0
Guardian art.	76.1	77.8	76.9
LOLITA with the new algorithm			
MUC6 training set	87.4	89.9	88.6
Guardian art.	84.6	89.5	87.0

Table 11.2: Evaluation results for the co-reference task performed on the MUC6 training set of 15 articles and on the *Guardian* article

11.1.3 The semi-blind tests

Table (11.3) shows the results obtained on the semi-blind tests. There rise in F-values in both of the semi blind tests is somewhere in between the rises noted in the blind test and the open tests. The rise in test A is bigger (6 points) than that in test B (around 4 points).

	Recall	Precision	F-measure
Old LOLITA improved			
Test A	56.8	71.8	63.4
Test B	69.8	69.3	69.6
LOLITA with the new algorithm			
Test A	62.8	77.5	69.4
Test B	72.5	75.0	73.7

Table 11.3: Evaluation results for the co-reference task performed on the tests sets A and B, using LOLITA system before and after the introduction of the new anaphora algorithm

11.2 Analysis of the general scores

The tables presented in the sections above contain interesting results. In particular, it is interesting to note the contrast in the rise obtained on the blind test with that obtained on the open tests.

The fact that the scores on the blind system haven't risen considerably more doesn't constitute evidence against the new algorithm. As has been stressed before (section 8.1.1 in chapter 8) an evaluation carried out on a blind test like the above doesn't say very much about how a single component of a complex natural language processing system is performing. All it can provide is evidence as to how the system as a whole is performing.

The contrast between the comparatively high rise on the open tests and a comparatively low rise on the blind test can be interpreted as showing that the new algorithm only comes into play and reveals its advantage over the old algorithm when other components of the LOLITA system also work well. Therefore, on all the open tests, where it can be claimed that most (if not all) aspects of the system's analysis were improved and can be shown to work well, the new algorithm is able to achieve success rates in the high eighties.

In the blind tests, on the other hand, it isn't possible to guarantee that other components of the system work well and this is the most likely explanation why the rise in scores is smaller.

An alternative view could be proposed, suggesting, that because the new algorithm was developed using the open training sets, the rules and heuristics proposed in it might be quite specific to those particular training sets and might not generalise onto other texts even from the same domain. This could be put forward as a possible reason why the rise in scores in the blind test is not as big as the rise in the open set.

However, the main assumption behind this view can be shown to be mistaken: the rules and heuristics in the algorithm were not constructed specifically for the open texts. Other materials, such as typical anaphora examples discussed in the literature were used. Additionally, every effort was made to find some independent motivation (e.g. psychological evidence or theoretical observation) for the rules.

11.3 Analysis of selected categories

The sections below discuss briefly the detailed results obtained in the evaluation of the new system. The results for the training test sets are presented first, followed by the detailed tables for tests A and B.

11.3.1 Analysis of the MUC6 training set

The following table illustrates the scores for individual categories obtained using the new system on the MUC6 training set of 15 articles.

CATEGORY	RESP	KEY	OVG	REC	PREC	F-VAL
Abbreviated name	53	53	0	1.00	1.00	1.00
Full name	60	61	1	0.98	0.98	0.98
Acronym	1	1	0	1.00	1.00	1.00
Inside appositive	21	22	0	0.95	1.00	0.97
Premod in compounds	11	15	0	0.73	1.00	0.84
NP inside ternary copula	21	22	1	0.95	0.95	0.95
NP with demonstrative det	4	4	0	1.00	1.00	1.00
NP with definite article	37	51	2	0.73	0.95	0.82
NP with indefinite article	11	15	0	0.73	1.00	0.84
NP object of copula verb	7	8	0	0.88	1.00	0.93
Head of relative clause	3	4	1	0.75	0.75	0.75
NP with possessive det	16	20	0	0.80	1.00	0.88
NP with Saxon genitive det	4	5	0	0.80	1.00	0.88
Sing noun with no det	19	24	2	0.79	0.90	0.84
Plural noun with no det	9	10	0	0.90	1.00	0.94
Head of NP	1	1	0	1.00	1.00	1.00
Possessive pronoun	0	0	0	-	-	
Possessive det	32	33	1	0.97	0.97	0.97
Ordinary pronoun	19	21	1	0.90	0.95	0.92
Demonstrative pronoun	0	0	0	-	-	
Referential adverbial	0	0	0	-	-	
Percentage	10	11	0	0.91	1.00	0.95
Absolute value	22	29	0	0.76	1.00	0.86
Relative value	7	9	0	0.78	1.00	0.87
Other value	5	12	0	0.42	1.00	0.59
Date	9	14	0	0.64	1.00	0.78
Other time expression	0	1	0	undef	undef	
Total markable objects	382	446	9			
Spuriously marked: 33						

Table 11.4: Breakdown of the co-reference scores obtained for the MUC6 training set, using LOLITA with the new algorithm

Comparing the detailed results in Table (11.4) with those obtained with the improved old system (Table (9.6), p. 134, in chapter 9), it must be noted that the F-values of all the categories have improved, with the exception of noun phrases with Saxon genitive, plural nouns without determiner and absolute values, which have all stayed the same.

It is also noteworthy that in this test set, only 3 pronouns (out of the categories of ordinary pronoun and possessive determiner) are incorrectly linked, so the F-value for pronouns has gone up from 88% to nearly 95%. By contrast, the noun phrases haven't improved as much.

11.3.2 Analysis of the *Guardian* article

The following table contains the detailed results obtained for the *Guardian* article. The categories not represented in the text have been omitted.

CATEGORY	RESP	KEY	OVG	REC	PREC	F-VAL
Abbreviated name	13	13	0	1.00	1.00	1.00
Full name	30	33	0	0.91	1.00	0.95
Acronym	1	1	0	1.00	1.00	1.00
Inside appositive	3	3	0	1.00	1.00	1.00
Premod in compounds	5	5	0	1.00	1.00	1.00
NP with definite article	9	14	1	0.64	0.90	0.74
NP with indefinite article	0	3	0	undef	undef	
Head of relative clause	0	1	0	undef	undef	
NP with possessive det	4	6	0	0.67	1.00	0.80
NP with Saxon genitive det	0	2	0	undef	undef	
Sing noun with no det	4	5	0	0.80	1.00	0.88
Plural noun with no det	0	1	0	undef	undef	
Possessive det	9	9	0	1.00	1.00	1.00
Ordinary pronoun	17	18	0	0.94	1.00	0.96
Absolute value	0	2	0	undef	undef	
Other value	0	4	0	undef	undef	
Total markable objects	95	120	1			
Spuriously marked: 10						

Table 11.5: Breakdown of the co-references scores obtained for the *Guardian* article using LOLITA with the new algorithm

The results for the *Guardian* article reveal that all categories have either stayed the same or improved, with the exception of definite noun phrases, which have gone down slightly (however, the loss is minimal).

Further inspection of the data reveals that in neither the improved old system's output nor in the new output the analysis of the relevant fragment of the article is perfect.

In the improved old system, the following chain is marked:

{*Sears — Sears — the company — the company*}

In this chain, the link between the two occurrences of *Sears* is correct and so is the link between the two noun phrases. However, the middle link (between *Sears* and *the company*) is incorrect. For this chain, the scorer returns two points for the category of full name and two for noun phrases with definite determiner.

In the new output, the system produces the following two chains:

{*Facia — the company*}
{*Sears — Sears — the company*}

The first chain here (between *Facia* and *the company*) is correct. As a side effect of making this correct link, however, the second mention of *the company* is now counted as incorrect, because (according to the method explained in chapter 8) no item in the second chain can be seen as correct with respect to *the company.*

This points to some shortcomings of scoring co-reference in general and to this method in particular. First of all, it's difficult to assess the 'correctness' of a link, without access to the rest of the interpretation of the text. Secondly, despite the fact that the new system now makes a good co-reference link that it hadn't made before, the scoring by categories doesn't register this at all. In fact, all it shows is a loss.

11.3.3 Analysis of the semi-blind tests A and B

The following two tables contain detailed breakdown of results for the semi-blind tests A and B.

CATEGORY	RESP	KEY	OVG	REC	PREC	F-VAL
Abbreviated name	18	20	0	0.90	1.00	0.94
Full name	27	42	1	0.64	0.96	0.76
Acronym	0	2	0	undef	undef	
Inside appositive	8	15	0	0.53	1.00	0.69
Premod in compounds	4	13	0	0.31	1.00	0.47
NP inside ternary copula	1	2	0	0.50	1.00	0.66
NP with demonstrative det	0	1	0	undef	undef	
NP with definite article	21	36	1	0.58	0.95	0.72
NP with indefinite article	0	4	0	undef	undef	
NP object of copula verb	0	0	0	-	-	
Head of relative clause	0	0	0	-	-	
NP with possessive det	7	11	0	0.64	1.00	0.78
NP with Saxon genitive det	4	6	0	0.67	1.00	0.80
Sing noun with no det	6	12	0	0.50	1.00	0.66
Plural noun with no det	1	4	0	0.25	1.00	0.40
Head of NP	0	0	0	-	-	
Possessive pronoun	0	0	0	-	-	
Possessive det	17	17	0	1.00	1.00	1.00
Ordinary pronoun	15	18	0	0.83	1.00	0.90
Demonstrative pronoun	0	0	0	-	-	
Referential adverbial	0	1	0	undef	undef	
Percentage	0	0	0	-	-	
Absolute value	1	4	0	0.25	1.00	0.40
Relative value	1	2	0	0.50	1.00	0.66
Other value	0	1	0	undef	undef	
Date	0	0	0	-	-	
Other time expression	0	3	0	undef	undef	
Total markable objects	131	214	2			
Spuriously marked: 28						

Table 11.6: Detailed results for the semi-blind **test A**, obtained using LOLITA with the new algorithm

CATEGORY	RESP	KEY	OVG	REC	PREC	F-VAL
Abbreviated name	15	17	0	0.88	1.00	0.93
Full name	33	44	2	0.75	0.94	0.83
Acronym	1	1	0	1.00	1.00	1.00
Inside appositive	11	16	0	0.69	1.00	0.81
Premod in compounds	2	6	0	0.33	1.00	0.49
NP inside ternary copula	17	21	1	0.81	0.94	0.87
NP with demonstrative det	1	2	0	0.50	1.00	0.66
NP with definite article	15	22	1	0.68	0.94	0.78
NP with indefinite article	0	0	0	-	-	
NP object of copula verb	0	0	0	-	-	
Head of relative clause	0	0	0	-	-	
NP with possessive det	1	3	0	0.33	1.00	0.49
NP with Saxon genitive det	0	0	0	-	-	
Sing noun with no det	6	17	0	0.35	1.00	0.51
Plural noun with no det	4	7	0	0.57	1.00	0.72
Head of NP	2	4	0	0.50	1.00	0.66
Possessive pronoun	0	0	0	-	-	
Possessive det	16	17	1	0.94	0.94	0.94
Ordinary pronoun	12	13	0	0.92	1.00	0.95
Demonstrative pronoun	0	0	0	-	-	
Referential adverbial	0	0	0	-	-	
Percentage	0	0	0	-	-	
Absolute value	7	9	0	0.78	1.00	0.87
Relative value	5	5	0	1.00	1.00	1.00
Other value	2	6	0	0.33	1.00	0.49
Date	0	0	0	-	-	
Other time expression	0	0	0	-	-	
Total markable objects	150	210	5			
Spuriously marked: 51						

Table 11.7: Detailed results for the semi-blind **test B**, obtained using LOLITA with the new algorithm

While the general scores for the tests A and B showed a rise, the detailed results are mixed: some categories show a rise, most notably possessive determiners, which show a marked rise in both tests. However, the scores for other categories have decreased slightly, when compared with the results obtained with the improved old system (Tables (9.10) and (9.12), on pages 140 and 142, respectively).

The latter group includes some premodifiers in compounds, plural or singular nouns without determiners and noun phrases with definite determiners.

In test B, there is also a small decrease in the score for ordinary pronouns. At the same time, the score for possessive determiners has gone up. Closer analysis reveals that, as was the case with the *Guardian* article, the system gains some correct links, at the same time causing losses in other categories. That is, while the improved old system produced a chain like the following:

$\{it - its - X\}$

where X was an incorrect item, not present in the answer key, the new system, on the other hand, produces the following two chains:

$\{it - X\}$
$\{its - Y\}$

where the connections between *its* and Y is correct. However, the scoring algorithm now considers the pronoun *it* as incorrectly connected.

Overall, it seems almost that the results in tests A and B reflect the state of the global analysis of the text as transitional (between poor and good). This is confirmed by a more detailed examination of the output produced for these tests.

However, despite the fact that not all aspects of the global analysis are very good, the new algorithm is able to make some new good co-references (hence the rise in some categories and the rise in overall score).

It seems possible that mixed results, such as those obtained for tests A and B, can generally be taken as symptomatic of the problems in the underlying basic analysis. This is in contrast to the situation found in the case of the open tests. There, the majority of the categories rose in scores, some stayed the same and very few (in fact, just one) went down minimally. This sort of result could be taken to reflect a very good and improving level of the global analysis of the text.

11.4 Comparison of the three systems

In this project a subset of co-reference links is of particular interest, because, in order to make them, the new resolution algorithm has to be triggered. These co-references involve pronouns, possessive determiners, noun phrases and demonstratives — the core categories. Data on one other category — the reflexives — would be useful, however, there were no occurrences of reflexives in the tagged test sets used.

The following table takes into account all the figures from all the tagged tests for selected categories. The resulting totals for each of the three systems are presented.

For comparison purposes, the table also includes the totals for the peripheral categories such as "is_a" based co-references and "function and value" co-references, as well the category of bare nominals (the latter include singular nouns with no determiner and plural nouns with no determiner).

CATEGORY	RESP	KEY	OVG	REC	PRE	F-VAL
LOLITA: the old MUC6 system						
Noun phrases	57	210	11	27	84	41
Pronouns	74	146	16	51	82	63
Bare nominals (pl & sing)	7	80	8	9	47	15
Is_a based corefs	33	109	11	30	75	43
function and value corefs	0	83	1	und	und	-
Old LOLITA improved						
Noun phrases	132	210	11	63	92	75
Pronouns	116	146	13	79	90	84
Bare nominals (pl & sing)	50	80	4	62	93	74
Is_a based corefs	88	109	3	81	97	81
function and value corefs	45	83	0	54	100	70
LOLITA with the new algorithm						
Noun phrases	137	210	6	65	96	77
Pronouns	138	146	3	95	98	96
Bare nominals (pl & sing)	49	80	2	61	96	74
Is_a based corefs	89	109	2	82	98	89
function and value corefs	50	83	0	60	100	75

Table 11.8: Comparison of total scores for selected categories

The new system achieves a considerable rise in scores for pronouns. There is a rise in all categories except for bare nominals.

The rise at this level is particularly impressive, as it is towards the upper limit of the scale. The general improvements achieved between the old MUC6 system and the improved MUC6 system (which was shown on all the tests, including the formal blind one) was very pleasing. However, the figure of 96% (F-value) for pronouns is even more impressive.

The results for noun phrases show a very big improvement between the old system and the improved old system (from 41 to 75 in F-values, with the improved old

system achieving 92% precision). Although in the new system the score for noun phrases has risen only slightly, it is a valuable rise, particularly because both recall and precision are now higher. However, there is room for improvement, especially on recall.

11.5 Analysis of selected examples

11.5.1 Some missed co-references with noun phrases

Closer examination of the types of noun phrases that the system didn't resolve suggests that in many cases these are complex noun phrases, often requiring very sophisticated semantic analysis. For example connections like the following are missing (NB: all from different texts):

the mortgage banking company — Plaza Home Mortgage Corp.
three business executives — The other new members

Other cases involve metonymy, e.g.:

the paper — the New York Post.

There are also noun phrases which are almost resolved correctly, but some remaining errors in Name Entity recognition mean that the co-reference links cannot be counted. For example, the key contains a link like this:

the footwear manufacturer — R.G. Barry Corp.

While the system produces:

the footwear manufacturer — Barry Corp.

Even though this link doesn't contribute towards the total score, it can be seen as

an improvement in co-reference — while using the old algorithm, the resolution of this particular noun phrase was totally wrong.

Apart from the examples like the ones above, there still remain a few seemingly simpler cases, which should have been resolved by the system but are not. Further analysis indicates that these are due to coding errors in the general system. It is expected that more debugging (which, as a matter of fact, is continuously being carried out) would most certainly help in the correct resolution of these cases.

11.5.2 Interesting co-reference links made

The excerpt in (71) shows some interesting factors at play. Here, the system successfully links all the emphasized noun phrases and the possessive determiner into one chain.

(71) *The management change is the latest in a series of events that have shaken* **the company** *in recent months. As previously reported, the Securities and Exchange Commission contacted several individuals about their dealings with* **the company**. *One of those individuals said the SEC had asked about how* **the company** *valued* **its** *assets.*

The second mention of *the company* has two possible antecedents, as far as LOLITA is concerned: one is the referent designated by the first mention of *the company* and the second — *the Securities and Exchange Commission*[1].

The link between *the company* and *the Securities and Exchange Commission* has a good chance to be made on the basis of the closeness of the two expressions (cf. *prefer_last_mention*). By contrast, the other *company* (the intended referent) is quite far away. However, the grammatical rule prevents the co-reference between

[1]NB the system could possibly use the cue contained in the name: *the Securities and Exchange Commission*, to decide that it is not *the company*, however, such a deep analysis of names is extremely difficult to perform reliably

the noun phrase and a c-commanding proper name, thus enabling the correct candidate to win.

In the third sentence, the system resolves the link between *the company* and the possessive *its* mainly on the basis of *prefer_last_mention* (other heuristics, such as for example *prefer_global_focus*, strengthen this link). The link is resolved before the referent of *the company* is chosen. Here again, the grammar rule rejects *SEC* as a candidate for *the company*, thus allowing for a correct resolution.

Another type of interesting link made by the system is in the *Guardian* article and concerns the title of the article and the first two sentence of the main text:

(72) *<HL>* **Facia saviour** *is bankrupt. < /HL>*

<TXT> **William Grosvenor**, *the entrepreneur and well connected cousin of the Duke of Westminster, who is heading attempts by Texas American Group to take over the troubled Facia retail company, last night admitted that he is bankrupt.*

* Mr Grosvenor, aged 54, a pageboy at the Queen's 1953 Coronation, who is acting as chief executive officer of the US-listed company trying to buy Britain's second largest privately owned retail chain, has a spent conviction for tax fraud in Britain, for which he received a 12 month suspended sentence and a £1,000 fine in 1980. *

The system makes a correct link between *Facia saviour* and *William Grosvenor.*

The candidates that are considered as possible referents are *Grosvenor*, *the Queen* and *the Duke of Westminster.* In the improved old system, *Grosvenor* was ruled out straight away, because the links of the concept representing *Grosvenor* had several connections to other concepts (e.g. entrepreneur, cousin) which made *Grosvenor* incompatible with a sister concept of *saviour.* Now, *Grosvenor* is allowed to match (as discussed in section 10.3.5.1) and wins convincingly: because the system has

already learned from the text that *Grosvenor is bankrupt* — the *Facia saviour* who is also asserted to be bankrupt matches perfectly.

11.5.3 Types of errors made

There are cases where problems with the semantic analysis of the system together with the new rules lead to losses of co-references. For example, consider the following excerpt:

(73) *Arkla Inc. said it plans to change its name to NorAm Energy Corp.*

(...)

The company said that with the addition of Mississippi River Transmission, Entex and Minnegasco, it has outgrown the geographical boundaries suggested by the name Arkla.

This is a particularly difficult text, because the strings *Arkla Inc* or *NorAm Energy Corp* are not always used in their most usual way, i.e. as referents to a company. Sometimes, they are pointing to the actual string, which constitutes the name (the so called use vs mention distinction). Currently, LOLITA's analysis doesn't distinguish such usages, so, for example, the first utterance would lead to the creation of two separate companies in LOLITA's semantic representation.

Similarly, in the second utterance, the final occurrence of *Arkla* is taken to be a reference to the company *Arkla Inc.* As a result, the grammatical rule for the interpretation of noun phrases comes into play and rules out *Arkla Inc* as a possible referent for the initial phrase *The company.* Instead, *The company* gets connected to *NorAm Energy Corp.*

In the improved old system, due to the absence of the grammatical rule, this error does not occur.

Another example, where the grammar rule prevented co-reference (however in this case, the underlying semantic analysis was fine) can be seen in: (74):

(74) *WSMP Inc. said Cecil R. Hash, chairman and chief executive officer, resigned, and the board picked two company officers to succeed him.*

The restaurant and food-service company *said Mr. Hash "made a personal decision to return to the daily management of the 25 restaurants he presently franchises from* **WSMP***".*

Cases like these suggest that the grammatical restriction for noun phrases and proper names appears too strong. It seems that when companies are involved, and the distance between the c-commanding mention and the proper name is very long, the rule should be disabled. However, more evidence needs to be collected before this can be done.

11.6 Limitations of the new system

There is a range of co-references that cannot be solved reliably, until some form of advanced reasoning can be used. For example the contrast in:

(75) *Susan sold the car to Jenny because she decided to take up cycling.*
Susan sold the car to Jenny because she offered to pay the asking price.

poses problems for a grammar, semantics and discourse based heuristics system. It is very difficult to capture the bias towards *Susan* in the first example and towards *Jenny* in the second without making pragmatic inferences.

Currently, LOLITA resolves both sentence the same way (*she* is *Susan*).

Another interesting case which might be solved with the use of some knowledge and pragmatic inference comes from the training set:

(76) *Spelling Entertainment Group Inc. named H. Wayne Huizenga, chairman and chief executive officer of Blockbuster Entertainment Corp., following Blockbuster's recent purchase of a controlling interest in Spelling.*

(...)

Television producer **Aaron Spelling**, *the company's founder, was named vice chairman and a member of the board's executive committee.*

Steven R. Berrard, Blockbuster's vice chairman, president and chief operating officer, was named president, chief executive and a director of **Spelling**.

The system resolves the emphasized occurrence of *Spelling* to be *Aaron Spelling* rather than the company *Spelling Entertainment.* The names matching algorithm doesn't take into account the knowledge that one is more likely to be a "director of a company" rather than a "director of a person".

11.7 Performance in MUC7

Since the introduction of the new algorithm the LOLITA system took part in the MUC7 competition. The system used for MUC7 was essentially the same as the new system described in this project, with only minor modifications (details of these can be found in (Garigliano *et al.* 1998)).

The co-reference task was not substantially different from that devised for MUC6.

The following scores were achieved by all the participants in the MUC7 co-reference evaluation:

LOLITA's Competitors	Recall	Precision	F-measure
System A	56.1	68.8	61.8
System B	58.2	64.2	61.1
System C	57.5	62.7	60.0
System D	46.8	78.0	58.5
LOLITA	46.9	57.0	51.5
System E	28.4	60.6	38.6
System F	52.5	21.4	30.4

Table 11.9: Official co-reference evaluation results of LOLITA and other participants of MUC7

LOLITA's performance was considered very good, particularly because relatively little time was spent on the preparations for the co-reference task itself.

What is noticeable about all the results as compared with those from MUC6 (Table 9.2, p. 129) is that they are on average several points lower. This indicates that the task has become harder. Despite that, LOLITA came close to the higher scoring systems. Initial analysis suggests that the main reason preventing the score from being closer to that achieved on the MUC6 blind test (with the new algorithm) was a very high (around 20%) parsing failure rate.

Chapter 12

Conclusions and Future Work

In the first section of this chapter some hypotheses regarding the new algorithm and its robustness are first considered. Then, the most important lessons learned from the project as well as its implications for the NLE paradigm are discussed. This is followed by the summary of major findings.

The second section considers future work.

12.1 General conclusions and major findings

12.1.1 Possible robustness of the new algorithm under failure of basic modules

The general improvements carried out to the system have resulted in a considerable rise in scores on the MUC6 blind test set, and so one of the main objectives of the project has been achieved.

The fact that the rise in scores between the improved old system and the system with the new anaphora resolution algorithm is not as as high as the majority of

the rises reported on other tests raises interesting questions about the comparison of the old algorithm and the new one.

It is hoped that one of the great values of the new algorithm is the way in which it allows for a delayed decision with respect to the choice of antecedent. In the old system, where a filter could remove a possible referent relatively early during the interpretation process, it was essential that the basis on which the filter applied was correct. In other words, it had to be ensured that there were no coding errors or other problems which provided the algorithm with erroneous data.

In the new, weight-based, system the decision as to the choice of antecedent is made on the basis of many facts, and the rules act cumulatively. This should allow for the new algorithm to perform better under conditions such as those which occurred at the time of MUC6. (As part of other activities within the LOLITA group, a similar approach has recently been devised and applied to another domain of AI: speech recognition (Collingham *et al.* 1997). The approach is reported to be highly successful.) It would be interesting to conduct an experiment in which the new algorithm were to be implemented within the reconstructed, old MUC6 system. It is expected that the results would be better than the ones achieved with the old algorithm under the same, generally poor conditions, though probably the difference would not be very large. However, the cost of conducting this kind of experiment (in terms of time and effort) would most likely outweigh any potential benefits of such a finding.

12.1.2 The most fruitful lessons learned

It seems that improving the performance of the overall analysis of the system (parsing, morphology, semantic rules, etc.), would be most beneficial for further progress on the co-reference task. The experiments conducted in this project show that these sorts of changes led to the biggest rises in scores. So, while there may be weaknesses associated with the current rules, correcting or improving those might

lead to smaller gains than could be achieved as a result of further work and more debugging of the basic components.

The results are consistent with the general approach of Natural Language Engineering and the particular method adopted in the Laboratory of NLE at Durham. It is first and foremost the painstaking work on a whole host of smaller problems which yields biggest gains (i.e. gains on all tests, not just the open ones).

The general improvements made to the system enabled it to achieve scores well above 60% on the co-reference task, using both open and blind tests. From that point onwards, it seems that the higher the scores are to rise, the more precise the analysis of the input text is required, and a generally sophisticated resolution algorithm is called for. This has been demonstrated on the open tests, where the lower level analysis of the texts has been significantly enhanced and therefore the new algorithm could be triggered. In some categories, such as pronouns, the rise with the new algorithm was from the F-value of 84, which is already very high, to 96, which is extremely high.

It is expected, however, that the remaining 4% or 5% of co-reference links in this category (and possibly in others) could only now be made with a sophisticated use of world knowledge (cf. examples in section 11.6, in chapter 11).

Currently, work is underway (Poria, forthcoming) within the LOLITA project to semi-automatically acquire world knowledge from the Cambridge International Dictionary of English (Procter 1995). As soon as that project progresses further, an attempt to use more world knowledge and advanced forms of reasoning in anaphora resolution can be made.

12.1.3 Positive implications for the NLE paradigm and LOLITA

The LOLITA project is shown to be viable as a scientific hypothesis. The current level of performance that LOLITA has been able to attain encourages a further hypothesis: it is possible to write a large-scale, generic system which can compete on even terms with the best current systems in the world, like those taking part in the recent MUC competitions, even if those systems were designed specifically for the co-reference task.

12.1.4 Major findings in summary

Much can be accomplished in terms of improvements of the performance on the co-reference task by addressing the basic components of an NLP system.

In the project described in this thesis, the first evaluation was carried out after several general improvements to the old LOLITA system were completed.

LOLITA's score on the MUC6 blind tests improved from 44.6 to 62.1 (F-value). On the training set of 15 articles (also from the MUC6 corpus) the score improved from 43.8 to 79.0 (F-value).

Before the second evaluation was carried out, a new algorithm for anaphora resolution was designed, drawing on existing work in the fields of AI, computational linguistics, theoretical linguistics and cognitive science. In implementing the design, an NLE approach was adopted, consistent with the general design of LOLITA.

In the second evaluation LOLITA's score on the blind test rose to 63.9 F-value while on the open set (of 15 articles) the score rose to 88.6 F-value. On the open test sets the algorithm achieves outstanding results, particularly for pronouns (F-value of around 96%).

It is concluded that very high scores can be achieved by perfecting all levels of the system's core analysis and applying the new, weights-based resolution algorithm.

12.2 Future work

12.2.1 Tests in other domains

A lot of the work carried out so far has been restricted to the domain chosen by the MUC6 evaluation. It is anticipated that significant gains can be made by testing the new system on texts from other domains. Some early indication of potential benefits of this have come from the MUC7 results. Furthermore, the algorithm is currently being (so far informally) tested in the area of dialogue.

12.2.2 The weights system

The new resolution algorithm relies on a system of weights which have been manually assigned. The system would benefit from a way of setting the weights in some statistical manner.

12.2.3 More work on the core

As discussed above, significant improvements resulted from attending to relatively minor flaws in the general system. To maintain further progress it is therefore essential to continue to work on the lower levels of the system.

12.2.4 The heuristics system

Several sophisticated rules and heuristics have been implemented to deal with a range of anaphoric expressions. However, the scores for some types of expressions (particularly noun phrases and bare nominals) remain about 20 points behind those of pronouns. More analysis and further rules are needed to improve this.

Additionally, some types of anaphora, e.g. possessive pronouns, aren't well accounted for in the new system and so require more work.

12.3 Concluding remarks

The research has been conducted inside the chosen methodological approach and has addressed a well defined problem. It has successfully contributed to its solution on different fronts (the improvement to low level modules and the new anaphoric algorithm) by collecting and analysing the evidence, designing the algorithms and implementing them. It has evaluated the work in a well known framework (MUC-style) and has drawn interesting hypotheses for the future directions of NLE in general and the LOLITA system in particular.

Appendix A

MUC6 co-reference task

COREFERENCE TASK DEFINITION

(MUC-5 1995)

==========

CONTENTS

==========

1. GENERAL NOTATION

1.1. SGML TAGGING

The annotation for coreference is SGML tagging within the text stream. Referring expressions and their antecedents are tagged as follows:

<COREF ID="100">Lawson Mardon Group Ltd.</COREF> said <COREF ID="101" TYPE="IDENT" REF="100">it</COREF> ...

The basic annotation contains the information to establish some type of link between an explicitly marked pair of noun phrases. In the above example, the pronoun "it" is tagged as referring to the same entity as the phrase, "Lawson Mardon Group Ltd."

There is one markup per string. Other links can be inferred from the explicit links. We assume that the coreference relation is symmetric and transitive, so if phrase A is marked as coreferential with B (indicated by a REF pointer from A to B), we can infer that B is coreferential with A; if A is coreferential with B, and B is coreferential with C, we can infer that A is coreferential with C.

1.2. THE "TYPE" ATTRIBUTE

The purpose of the TYPE attribute is to indicate the relationship between the anaphor and the antecedent. At present only one such relationship, "IDENT" (for identity), is being annotated.

1.3. THE "ID" AND "REF" ATTRIBUTES

The ID and REF attributes are used to indicate that there is a coreference link between two strings. The ID is arbitrarily but uniquely assigned to the string during markup. The REF uses that ID to indicate the coreference link.

1.4. THE "MIN" ATTRIBUTE

The MIN attribute is used in the answer key ("key") to indicate the minimum string that the system under evaluation must include in the COREF tag in order to receive full credit for its output ("response"). So, in the next example, if the system response had omitted "of Surrey, England" from the COREF tag, the response would nonetheless receive full credit because it identified the minimum string.

<COREF ID="100" MIN="Haden MacLellan PLC">Haden MacLellan PLC of Surrey, England</COREF> ...

<COREF ID="101" TYPE="IDENT" REF="100">Haden MacLellan</COREF>

Any response which includes the MIN string and does not include any tokens beyond those enclosed in the <COREF>...</COREF> tags is valid. The MIN string will in general be the HEAD of the phrase; see section 4 for a full discussion of this issue.

1.5. THE "STATUS" ATTRIBUTE

The STATUS attribute is used in the answer key when the markup is optional. The only value for this attribute is OPT ("optional"). The evaluation software will not score a string that is marked OPT in the key unless the response has markup on that string. A potential example is given below. (It is marked OPT because a reader may not be certain that "Livingston Street" refers to the Board of Education.) Note that the optionality is marked only for the anaphor.

```
<COREF ID="102" MIN="Board of Education">Our Board of
Education</COREF> budget is just too high, the Mayor
said. <COREF ID="103" STATUS="OPT" TYPE="IDENT"
REF="102">Livingston Street</COREF> has lost control.
```

2. WHAT PART OF THE TEXT TO ANNOTATE

The <TXT> portion of the article should be annotated as well as the <HL>, the <DD>, and the <DATELINE> from the article header, but not any other lines from the header. (The DD tag sometimes doesn't appear at all, sometimes appears once, and sometimes appears twice. When it appears twice, only the SECOND instance is to be annotated.)

Lines within the <TXT> portion of the article that start with the "@" sign signify a table or other special line formatting within the text and should NOT be annotated. (However, such lines may also appear within the <HL> portion of the article, and these should be annotated.)

3. WHAT THINGS TO ANNOTATE

3.1. MARKABLES

The coreference relation will be marked between elements of the following categories: NOUNS, NOUN PHRASES, and PRONOUNS. Elements of these categories are MARKABLES. PRONOUNS include both personal and demonstrative pronouns, and with respect to personal pronouns, all cases, including the possessive. Dates ("January 23"), currency expressions ("$1.2 billion"), and percentages ("17%") are considered noun phrases.

The relation is marked only between pairs of elements both of which are markables. This means that some markables that look anaphoric will not be coded, including pronouns, demonstratives, and definite NPs whose antecedent is a clause rather than a markable. For example, in

> Program trading is "a racket," complains Edward Egnuss, a White Plains, N.Y., investor and electronics sales executive, "and *it's not to the benefit of the small investor*, *that*'s for sure."

Though "that" is related to "it's not to the benefit of the small investor", the latter is not markable, so no antecedent is annotated for "that".

3.2. NAMES AND OTHER NAMED ENTITIES

Names and other Named Entities (as defined in the MUC-6 document titled "Named Entity Task Definition" -- dates, times, currency amounts, and percentages) are all markables. A substring of a Named Entity, however, is not a markable.

Thus in

London ... *London*-based ...

the two instances of London are to be marked coreferential; in

Reuters Holding PLC ... *Reuters* announced that

"Reuters Holding PLC" and "Reuters" are to be marked coreferential. But in

Equitable of Iowa Cos. ... located in Iowa.

the two instances of "Iowa" are NOT to be marked as coreferential since the first is not a markable: it is a substring of a Named Entity.

Date expressions recognized by the Named Entity task are also treated as atomic; components of a date are not separate markables. Thus, in

In a report issued January 5, 1995, the program manager
said that there would be no new funds this year.

no relation is to be marked between "1995" and "this year".

3.3. GERUNDS

Gerunds (verbal forms using a present participle) are not markable. In

Slowing the economy is supported by some Fed officials;
it is repudiated by others.

one should not mark the relation between "slowing the economy" and "it". A phrase headed by a present participle is taken to be verbal if it can take an object (as in the above example) or can be modified by an adverb.

Present participles which are modified by other nouns or adjectives ("program trading", "excessive spending"), are preceded by "the" or are followed by an "of" phrase ("the slowing of the economy") are to be considered noun-like and ARE markable.

3.4. PRONOUNS

The possessive forms of pronouns used as determiners are markable. Thus in

its chairperson

there are are two potential markables for relations: "its" and the entire NP, "its chairperson". Similarly, in "the man's arm", there are two markables.

First, second, and third-person pronouns are all markable, so in

"There is no business reason for *my* departure",
he added.

"my" and "he" should be marked as coreferential. Reflexive pronouns are markable, so in

He shot *himself* with *his* revolver.

"He", "himself", and "his" should all be marked coreferential.

3.5. BARE NOUNS

Prenominal occurrences of nouns, e.g., in compound nouns, are markable. Thus in

The price of *aluminum* siding has steadily increased, as
the market for *aluminum* reacts to the strike in Chile.

the relation between the two occurrences of "aluminum" should be marked. Note this presupposes that the two occurrences co-refer; they do, they both refer to the type of material.

While nouns in prenominal positions are markable, the noun which appears at the head of a noun phrase is not separately markable -- it is markable only as part of the entire noun phrase. Thus in the passage

Linguists are a strange bunch. Some linguists even like
spinach.

it would not be correct to link the two instances of "linguists".

3.6. IMPLICIT PRONOUNS

Assume that English has no zero pronouns; in other words, the empty string is not markable. In

Bill called John and spoke with him for an hour.

there is no relation between the implicit subject of "spoke" and "Bill".

Do not code relations between a relative pronoun and the head it attaches to or the gap that it fills.

3.7. CONJOINED NOUN PHRASES

Noun phrases which contain two or more heads (as defined in section 4.1) are NOT markable. This restriction is imposed so that each markable can be identified by a unique contiguous head substring. Thus no coreference is to be marked for

The boys and girls enjoy their breakfast.

The individual conjuncts are markable if they are separately coreferential with other phrases:

<COREF ID="1">Edna Fribble</COREF> and <COREF ID="2">Sam
Morton</COREF> addressed the meeting yesterday. <COREF
ID="3" REF="1" TYPE="IDENT" MIN="Fribble">Ms. Fribble</COREF>

discussed coreference, and <COREF ID="4" REF="2" TYPE="IDENT" MIN="Morton">Mr. Morton</COREF> discussed unnamed entities.

If the conjuncts share modifiers, the coreference is optional:

<COREF ID="1" MIN="Fribble">Ms. Fribble</COREF> was <COREF ID="2" REF="1" TYPE="IDENT" STATUS="OPT">president</COREF> and <COREF ID="3" REF="1" TYPE="IDENT" STATUS="OPT" MIN="CEO"> CEO of Amalgamated Text Processing Inc.</COREF>

4. HOW MUCH OF THE MARKABLE TO ANNOTATE

The task is defined in order to allow maximal latitude for systems in identifying markables, and to decouple the evaluation from that of accurately parsing noun phrases. Accordingly, the string generated by a system to identify a markable must include the head of the markable (as defined below) and may include any additional text up to a maximal noun phrase (as defined below).

In preparing the key, the text element to be enclosed in SGML tags is the maximal noun phrase; the head will be designated by the MIN attribute.

[We expect that in the future it may be possible, when separate noun phrase bracketings are available, to automatically generate the maximal NP markup from a markup using only heads.]

4.1. HEAD OF A PHRASE

For most noun phrases, the head will be the main noun, without its left and right modifiers.

<COREF MIN="task" ...>the coreference task</COREF>
<COREF MIN="contract" ...>the last contract</COREF> you will ever get

<COREF MIN="quantity" ...>a large quantity of sugar</COREF>
<COREF MIN="tons" ...>about 200,000 tons of sugar</COREF>

If the head is a name, the entire name is marked. This includes suffixes such as "Sr.", "III", etc. on personal names and "Corp." on organization names; it does not include personal titles or any modifiers. We follow in this regard the rules for marking personal and organization names for the Named Entity task.

<COREF MIN="Frederick F. Fernwhistle Jr." ...>
the Honorable Frederick F. Fernwhistle Jr.</COREF>
<COREF MIN="Ford Motor Co." ...>
Ford Motor Co. of Dearborn, Michigan</COREF>
<COREF MIN="Georg Rath" ...>Herr Dr. Georg Rath</COREF>

In the case of location designators consisting of multiple names, each name is considered a separate unit (as in the Named Entity task) and the head is generally the first of these names, with the others treated as modifiers of the first name:

<COREF MIN="Newark" ...>Newark, New Jersey</COREF>

Dates, currency amounts, and percentages are also treated as atomic units, as in the Named Entity task:

```
<COREF MIN="December 7, 1941" ...>December 7, 1941, a day
        which will live in infamy, </COREF>

<COREF MIN="$1.2 million" ...>$1.2 million in crisp
        bills</COREF>

<COREF MIN="20%">20% of the shares</COREF>
```

In the case of "headless" constructions, the "head" -- for coreference purposes -- shall be the last token of the noun phrase preceding any prepositional phrases, relative clauses, and other "right modifiers":

```
<COREF MIN="seven" ...>seven of the best</COREF>
<COREF MIN="five" ...>the five who were left standing</COREF>
<COREF MIN="youngest" ...>the six youngest</COREF>
```

If the maximal noun phrase is the same as the head, the MIN need not be marked.

4.2. MAXIMAL NOUN PHRASE

The maximal noun phrase includes all text which may be considered a modifier of the noun phrase. This includes (among other modifiers) appositional phrases, non-restrictive relative clauses, and prepositional phrases which may be viewed as modifiers of the noun phrase or of a containing clause:

```
*Mr. Holland*
*the senior of the executives who will assume Holland's duties*
*the rumor that the war had ended*
*Fred Frosty, the ice cream king of Tyson's Corner,*
*the Penn Central Co., which used to run a railroad,*
XYZ Inc. formed *a joint venture with Sony*
```

Note that in the fourth and fifth cases the final comma may be viewed as part of the NP, and so is included in the maximal NP; in the last case, "with Sony" could equally well be taken to modify "venture" or "formed", and so is included as part of the maximal NP around "venture". Note also that in the "Fred Frosty" example, there is a coreference between the entire noun phrase and the appositional phrase, "the ice cream king of Tyson's Corner"; see section 5.3 for a discussion of this construct.

In the case of a pair of conjoined noun phrases with shared complements or modifiers, the maximal noun phrases will NOT include the conjunct. The maximal NP for the first conjunct will include all of the NP up to the conjunction; the maximal NP for the second conjunct will include all of the NP following the conjunction:

```
<COREF ID="1" MIN="Fribble">Ms. Fribble</COREF> was <COREF
ID="2" REF="1" TYPE="IDENT" STATUS="OPT">president</COREF> and
<COREF ID="3" REF="1" TYPE="IDENT" STATUS="OPT" MIN="CEO">CEO
of Amalgamated Text Processing Inc.</COREF>
```

4.3. EXCEPTIONS: ARTICLES

If the only difference between the head and the maximal noun phrase is the presence of an article -- the word "the", "a", or "an" at the beginning of the noun phrase -- the MIN need not be explicitly marked. (The scoring program will automatically strip leading articles before comparing strings.)

5. WHICH RELATIONSHIPS TO ANNOTATE

5.1. BASIC COREFERENCE

The basic criterion for linking two markables is whether they are coreferential: whether they refer to the same object, set, activity, etc. It is not a requirement that one of the markables is "semantically dependent" on the other, or is an anaphoric phrase.

5.2. BOUND ANAPHORS

We also make a coreference link between a "bound anaphor" and the noun phrase which binds it (even though one may argue that such elements are not coreferential in the usual sense). Thus we would link a quantified noun phrase and a pronoun dependent on that quantification:

Most computational linguists prefer *their* own parsers.

Note that a quantified noun phrase would also be linked to subsequent anaphors, outside the scope of quantification, through the usual relation of identity of coreference. Thus in the following text all three noun phrases would be linked:

Every TV network reported *its* profits yesterday.
They plan to release full quarterly statements tomorrow.

By this rule, a pronoun in a relative clause which is bound to the head of the clause would get a coreference link to the entire NP. Thus, for

every man who knows his own mind

we would establish a coreference link between "his" and the entire noun phrase "every man who knows his own mind":

<COREF ID="1" MIN="man">every man who knows
<COREF ID="2" REF="1" TYPE="IDENT">his</COREF>
own mind</COREF>

5.3. APPOSITION

A typical use of an appositional phrase is to provide an alternative description or name for an object:

Julius Caesar, the well-known emperor,

This identity of reference is to be represented by a coreference link between the appositional phrase, "the well-known emperor" and the ENTIRE noun phrase, "Julius Caesar, the well-known emperor":

<COREF ID="1" MIN="Julius Caesar">Julius Caesar, <COREF ID="2" REF="1" MIN="emperor" TYPE="IDENT"> the well-known emperor,</COREF></COREF>

The appositional phrase may be separated from the head by other modifiers. Thus

Peter Holland, 45, deputy general manager, ...

becomes

<COREF ID="1" MIN="Peter Holland">Peter Holland, 45,
<COREF ID="2" REF="1" TYPE="IDENT" MIN="manager"> deputy general manager,</COREF></COREF>

Appositional phrases that are marked indefinite are NOT considered to be coreferential. Examples of noncoreferential appositional phrases include the following:

Ms. Ima Head, a 10-year MUC veteran,
San Diego, one of America's finest cities,

Currently, only appositional phrases that are overtly marked via punctuation are considered markables. Thus, no coreference is marked in cases such as the following:

the real estate company * Century 21*
the realtor *Century 21*
presidential advisor *Joe Smarty*
Treasury Secretary *Bucks*
*the job of *manager**

5.4. PREDICATE NOMINATIVES AND TIME-DEPENDENT IDENTITY

Predicate nominatives are also typically coreferential with the subject. Thus in the example

Bill Clinton is the President of the United States.

we would record a coreference link between "Bill Clinton" and "the President of the United States". Coreference should NOT be recorded if the text only asserts the possibility of identity between two markables. In

Phinneas Flounder may be the dumbest man who ever lived.
Phinneas Flounder is a leading candidate to become president.
If elected, Phinneas Flounder would be the first Californian in the Oval Office.

no coreference is to be recorded.

Neither should coreference be recorded when the predicate nominative is marked indefinite. Examples of noncoreferential predicate nominatives include

Mediation is a viable alternative to bankruptcy.

Farm-debt mediation is one of the Farm Belt's success stories.
ARPA program managers are nice people.

Two markables should be recorded as coreferential if the text asserts them to be coreferential at ANY TIME. Thus

Henry Higgins, who was formerly sales director for Sudsy Soaps, became president of Dreamy Detergents

should be annotated as

<COREF ID="1" MIN="Henry Higgins">Henry Higgins, who was formerly <COREF ID="2" MIN="director" REF="1" TYPE="IDENT">sales director for Sudsy Soaps,</COREF></COREF> became <COREF ID="3" MIN="president" REF="1" TYPE="IDENT">president of Dreamy Detergents</COREF>

Even if the copula or inchoative verb is embedded, coreference should be marked, as in Dreamy Detergents named Henry Higgins to be president

which should be annotated as

Dreamy Detergents named <COREF ID="1">Henry Higgins</COREF> to be <COREF ID="2" REF="1" TYPE="IDENT">president</COREF>

When the copula is implied by the semantics of the verb but is not expressed overtly, the coreference relation will be marked optional in the answer key. Express ions of equivalence involving the word "as" will also be marked optional. The NP s enclosed in asterisks in the following examples will be marked optionally coref erential:

Dreamy Detergents named *Henry Higgins* *president*
Henry Higgins* is considered *Sudsy Soap's best sales director*
Higgins* will serve as *president of Dreamy Detergents*

5.5. TYPES AND TOKENS

The general principle for annotating coreference is that two markables are coreferential if they both refer to sets, and the sets are identical, or they both refer to types, and the types are identical. There are a number of problematic cases where one can argue whether something is a set or a type. There is no simple algorithm for determining the ontological category of a referent. There are, though, some useful rules. Most occurrences of bare plurals refer to types or kinds, not to sets. In

...*producers* don't like to see a hit wine increase in price... *Producers* have seen this market opening up and *they*'re now creating wines that appeal to these people.

"producers", "Producers", and "they" refer to types and they all refer to the same type. Notice that if interpreted as referring to sets, they would not all refer to the same set. More properly, there is no reason to think they would corefer; not all the producers who have seen the market opening up have created new wines.

Note that a type can be referred to by a bare plural, a definite singular np ("the tiger is fast becoming extinct") or a (bare) prenominal. In

> The action followed by one day an Intelogic announcement that it will retain an investment banker to explore alternatives "to maximize *shareholder* value," including the possible sale of the company. Mr. Edelman declined to specify what prompted the recent moves, saying they are meant only to benefit *shareholders* when "the company is on a roll."

the two starred occurrences corefer to the type: shareholder (of Intelogic).

5.6. FUNCTIONS AND VALUES

In GM announced *its third quarter profit*. *It* was *$0.02*.

all three starred phrases refer to an amount of money; they all refer to the same amount of money. Hence they are coreferential. The first phrase, in context, refers to that amount via referring to a function, say of companies and quarters of a year--or times. (In addition, the "its" in the first NP would be linked to GM.) In

> General Motors announced [their third quarter profit of *$0.02*].

the bracketed and starred phrases are coreferential. They refer to one and the same amount of money. Note that here, as in the case of apposition, the result is that a phrase is marked as being coreferential with a part of the phrase.

In

> *The temperature* is *90*....The temperature is rising.

the first occurrence of "the temperature" refers to the value of the function at arguments (places, times) supplied by context. That occurrence is coreferential with "90". In the second occurrence, "the temperature" refers to the function (indirectly, by way of referring to the derivative of the function). So it is not coreferential with the first occurrence or with "90".

There will be cases where a phrase could arguably refer to either a set or a type; in such ambiguous cases, the coreference should be recorded but marked as optional.

5.7. METONYMY

The pervasive phenomenon of metonymy raises a problem for Coreference relations. Do we annotate and recognize the relation before or after coercion? Here are some texts to consider:

(1) *The White House* sent its health care proposal to Congress yesterday. Senator Dole said *the administration*'s bill had little chance of passing.

(2) *Ford* announced a new product line yesterday. *Ford*

spokesman John Smith said *they* will start manufacturing widgets.

(3) I bought the New York Times this morning. I read that the editor of the New York Times is resigning.

(4) *The United States* is a democracy. *The United States* has an area of 3.5 million square miles.

We propose that coreference be determined with respect to coerced entities. Of course, this still leaves open the question as to the circumstances under which coercion is required. In (1) there is a coercion from the White House to the administration operating out of the White House, and that is IDENT with "the administration"; so "White House" and "administration" are IDENT. (Notice that there is also a question as to whether the administration's proposal is the same as its bill. This too requires a coercion of sorts.) In (2), while there might seem to be a coercion from Ford to a spokesman for Ford, we believe that such a coercion is not necessary, for it is plausible that corporations, as legal persons, can do many of the things that people can do--such as 'announce'. They may have to do some or all such things through other agents, but many people do many things that way. And if Ford can announce, then it, through one of its spokesmen, can "say". Believing that no coercion is required, we would mark as coreferential the first instance of "Ford", the second instance of "Ford" (in the phrase "Ford spokesman John Smith"), and "they", but would NOT mark the phrase "Ford spokesman John Smith" as coreferential with anything else in this passage. In (3) the first "New York Times" is coerced into a copy of the paper published by the New York Times and the second is coerced into the organization; so they are not IDENT. (4) is somewhat akin to (2). Countries are both geographical entities and governmental units. Thus, no coercion is necessary and the two starred occurrences are coreferential.

In the absence of general principles, a body of such decisions will need to be developed to codify the rules for coercion and coreference. In cases where there has been no clear precedent, the answer keys for formal evaluations will need to mark coreference as optional.

6. BASIS OF JUDGMENT

The coreference judgments should be based on the intelligent reader's knowledge of the world resulting from his or her best understanding of the text. It should not be based on a theory of the structure of the text, or on a linguistic theory of how NPs are resolved, or on estimates of what the typical NLP system could do. This means that some relations will be impossible for current NLP systems to recover, but this is why the task will push the technology. The annotators should assume that they are typical intelligent readers.

7. SCORING AND THE ORDERING OF LINKS

If three markables, A, B, and C, are coreferential, this relationship

could be recorded in the key in several ways: for example, by a REF pointer in both B and C pointing to A, or by a REF pointer in B pointing to A and a REF pointer in C pointing to B. A similar range of variations is possible in a system response. The current scoring rules provide that any correct key, when compared to any correct response, will yield a 100% recall/100% precision score, independent of the way the coreference relation is encoded in the key by REF pointers.

However, if the response is incomplete, its recall score CAN be affected by the way in which the coreference relation is encoded by the key. It is therefore recommended that each markable which participates in a coreference relation have a REF pointer to the most recent prior coreferential markable which does not have STATUS="OPT".

Appendix B

Evaluation scores in detail

The following sections contain co-reference scores with a separate score for each article of the given test set. First, the scores of the improved old system are included. This is followed by the scores of the new system.

B.1 Scores of the improved old system

B.1.1 Co-reference scores for the MUC6 training set

Document No.	Key Cls	Rsp Cls	Recall		Precision		f
9304120090	11	11	28/32	87.5	28/31	90.3	88.9
9403020061	8	8	20/27	74.1	20/26	76.9	75.5
9309270024	8	7	26/28	92.9	26/27	96.3	94.5
9403250108	12	13	23/35	65.7	23/34	67.6	66.7
9403100087	10	14	26/31	83.9	26/31	83.9	83.9
9303190065	11	10	16/25	64.0	16/25	64.0	64.0
9401110056	12	13	18/29	62.1	18/34	52.9	57.1
9304020074	11	11	23/31	74.2	23/29	79.3	76.7
9404250043	9	10	30/33	90.9	30/33	90.9	90.9
9401140038	9	7	24/28	85.7	24/26	92.3	88.9
9310210152	3	3	7/8	87.5	7/8	87.5	87.5
9403240097	2	3	5/6	83.3	5/6	83.3	83.3
9302090147	2	4	4/5	80.0	4/6	66.7	72.7
9404070168	3	2	4/5	80.0	4/5	80.0	80.0
9401240017	1	1	2/2	100.0	2/2	100.0	100.0
Coreference Totals:			256/325	78.8%	256/323	79.3%	79.0%

B.1.2 Scores for *The Guardian* article

	Recall		Precision		f
Coreference Totals:	70/92	76.1%	70/90	77.8%	76.9%

B.1.3 Scores for Test A

Document No.	Key Cls	Rsp Cls	Recall		Precision		f
9402020069	9	5	11/22	50.0	11/15	73.3	59.5
9402160084	7	5	5/11	45.5	5/7	71.4	55.6
9305130154	8	6	11/28	39.3	11/19	57.9	46.8
9304200015	4	5	16/19	84.2	16/17	94.1	88.9
9402100164	6	7	11/14	78.6	11/13	84.6	81.5
9305240037	6	7	11/17	64.7	11/13	84.6	73.3
9404120128	6	8	5/15	33.3	5/12	41.7	37.0
9309150121	3	4	5/8	62.5	5/9	55.6	58.8
9305120128	4	2	3/7	42.9	3/6	50.0	46.2
9312150136	3	3	6/7	85.7	6/6	100.0	92.3
Coreference Totals:			84/148	56.8%	84/117	71.8%	63.4%

B.1.4 Scores for Test B

Document No.	Key Cls	Rsp Cls	Recall		Precision		f
9403150022	8	8	16/24	66.7	16/26	61.5	64.0
9306290071	6	6	7/11	63.6	7/12	58.3	60.9
9404060109	7	7	15/19	78.9	15/19	78.9	78.9
9302190013	7	6	13/18	72.2	13/15	86.7	78.8
9404070109	6	8	9/16	56.2	9/18	50.0	52.9
9401310163	5	7	15/17	88.2	15/17	88.2	88.2
9401060179	4	4	7/9	77.8	7/8	87.5	82.4
9306100114	5	5	8/15	53.3	8/13	61.5	57.1
9307210154	3	4	10/12	83.3	10/10	100.0	90.9
9402280107	3	4	4/8	50.0	4/12	33.3	40.0
Coreference Totals:			104/149	69.8%	104/150	69.3%	69.6%

B.1.5 Scores for the blind test set

Document No.	Key Cls	Rsp Cls	Recall		Precision		f
9309140164	4	2	8/12	66.7	8/8	100.0	80.0
9401120067	4	3	5/9	55.6	5/9	55.6	55.6
9312230003	3	2	6/8	75.0	6/8	75.0	75.0
9404130062	5	4	6/9	66.7	6/11	54.5	60.0
9305050122	2	3	1/3	33.3	1/5	20.0	25.0
9304010017	3	4	5/9	55.6	5/9	55.6	55.6
9306100111	8	7	10/17	58.8	10/14	71.4	64.5
9401130019	6	5	14/20	70.0	14/17	82.4	75.7
9302030136	11	10	14/20	70.0	14/20	70.0	70.0
9301190125	4	5	14/18	77.8	14/18	77.8	77.8
9401040117	3	4	15/18	83.3	15/18	83.3	83.3
9404200037	10	6	10/22	45.5	10/15	66.7	54.1
9307190045	12	10	10/18	55.6	10/16	62.5	58.8
9311150068	8	8	22/31	71.0	22/25	88.0	78.6
9309100116	5	6	28/37	75.7	28/35	80.0	77.8
9306220057	10	7	24/35	68.6	24/35	68.6	68.6
9401190015	12	15	22/33	66.7	22/37	59.5	62.9
9310040005	11	14	30/42	71.4	30/38	78.9	75.0
9303250020	19	19	48/68	70.6	48/61	78.7	74.4
9404110093	17	21	41/57	71.9	41/72	56.9	63.6
9402230039	22	21	40/76	52.6	40/79	50.6	51.6
9311020154	18	25	41/75	54.7	41/89	46.1	50.0
9301190098	23	28	29/43	67.4	29/56	51.8	58.6
9403160006	22	25	58/77	75.3	58/84	69.0	72.0
9307290143	36	23	31/85	36.5	31/63	49.2	41.9
9402240133	15	31	96/122	78.7	96/137	70.1	74.1
9402180067	43	59	58/125	46.4	58/130	44.6	45.5
9305040023	36	46	110/157	70.1	110/164	67.1	68.5
9404040040	30	38	107/166	64.5	107/149	71.8	67.9
9401270105	41	48	53/133	39.8	53/112	47.3	43.3
Coreference Totals:			956/1545	61.9%	956/1534	62.3%	62.1%

B.2 Scores of the new system

B.2.1 Scores for the MUC6 training set of 15 articles

Document No.	Key Cls	Rsp Cls	Recall		Precision		f
9304120090	11	12	30/32	93.8	30/31	96.8	95.2
9403020061	9	8	24/27	88.9	24/27	88.9	88.9
9309270024	8	7	25/28	89.3	25/26	96.2	92.6
9403250108	12	12	28/35	80.0	28/32	87.5	83.6
9403100087	10	12	28/31	90.3	28/29	96.6	93.3
9303190065	11	9	19/25	76.0	19/26	73.1	74.5
9401110056	12	13	25/30	83.3	25/32	78.1	80.6
9304020074	11	10	24/30	80.0	24/27	88.9	84.2
9404250043	9	9	31/33	93.9	31/33	93.9	93.9
9401140038	9	7	25/28	89.3	25/26	96.2	92.6
9310210152	3	3	8/8	100.0	8/8	100.0	100.0
9403240097	2	2	6/6	100.0	6/6	100.0	100.0
9302090147	2	4	4/5	80.0	4/6	66.7	72.7
9404070168	3	3	5/5	100.0	5/5	100.0	100.0
9401240017	1	1	2/2	100.0	2/2	100.0	100.0
Coreference Totals:			284/325	87.4%	284/316	89.9%	88.6%

B.2.2 Scores for the *Guardian* article

	Recall		Precision		f
Coreference Totals:	77/91	84.6%	77/86	89.5%	87.0%

B.2.3 Scores for test A

Document No.	Key Cls	Rsp Cls	Recall		Precision		f
9402020069	9	6	15/22	68.2	15/17	88.2	76.9
9402160084	7	2	4/11	36.4	4/5	80.0	50.0
9305130154	8	7	18/28	64.3	18/23	78.3	70.6
9304200015	4	4	17/19	89.5	17/18	94.4	91.9
9402100164	6	8	10/14	71.4	10/14	71.4	71.4
9305240037	6	7	11/17	64.7	11/13	84.6	73.3
9404120128	6	7	6/15	40.0	6/12	50.0	44.4
9309150121	3	2	3/8	37.5	3/5	60.0	46.2
9305120128	4	3	4/7	57.1	4/7	57.1	57.1
9312150136	3	4	5/7	71.4	5/6	83.3	76.9
Coreference Totals:			93/148	62.8%	93/120	77.5%	69.4%

B.2.4 Score for test B

Document No.	Key Cls	Rsp Cls	Recall		Precision		f
9403150022	8	6	15/24	62.5	15/23	65.2	63.8
9306290071	6	6	7/11	63.6	7/12	58.3	60.9
9404060109	7	7	15/19	78.9	15/18	83.3	81.1
9302190013	7	7	15/18	83.3	15/16	93.8	88.2
9404070109	6	7	10/16	62.5	10/15	66.7	64.5
9401310163	5	6	16/17	94.1	16/17	94.1	94.1
9401060179	4	6	6/9	66.7	6/9	66.7	66.7
9306100114	5	5	9/15	60.0	9/13	69.2	64.3
9307210154	3	3	10/12	83.3	10/10	100.0	90.9
9402280107	3	4	5/8	62.5	5/11	45.5	52.6
Coreference Totals:			108/149	72.5%	108/144	75.0%	73.7%

B.2.5 Scores for the blind test set

Document No.	Key Cls	Rsp Cls		Recall		Precision	f
9309140164	4	3	7/12	58.3	7/8	87.5	70.0
9401120067	4	5	5/9	55.6	5/9	55.6	55.6
9312230003	3	2	7/8	87.5	7/8	87.5	87.5
9404130062	5	5	6/9	66.7	6/12	50.0	57.1
9305050122	2	5	1/3	33.3	1/5	20.0	25.0
9304010017	3	2	7/9	77.8	7/8	87.5	82.4
9306100111	8	7	9/17	52.9	9/13	69.2	60.0
9401130019	5	5	13/19	68.4	13/16	81.2	74.3
9302030136	10	8	14/19	73.7	14/17	82.4	77.8
9301190125	4	4	17/19	89.5	17/19	89.5	89.5
9401040117	3	5	16/18	88.9	16/18	88.9	88.9
9404200037	10	7	9/22	40.9	9/16	56.2	47.4
9307190045	12	12	9/18	50.0	9/17	52.9	51.4
9311150068	8	8	23/31	74.2	23/27	85.2	79.3
9309100116	5	6	29/37	78.4	29/37	78.4	78.4
9306220057	10	9	26/35	74.3	26/34	76.5	75.4
9401190015	12	15	24/33	72.7	24/37	64.9	68.6
9310040005	11	14	28/42	66.7	28/35	80.0	72.7
9303250020	18	12	48/68	70.6	48/60	80.0	75.0
9404110093	17	22	42/57	73.7	42/71	59.2	65.6
9402230039	22	26	41/76	53.9	41/73	56.2	55.0
9311020154	18	24	43/74	58.1	43/86	50.0	53.8
9301190098	23	28	28/43	65.1	28/57	49.1	56.0
9403160006	21	28	62/77	80.5	62/86	72.1	76.1
9307290143	35	26	37/85	43.5	37/67	55.2	48.7
9402240133	14	25	96/122	78.7	96/129	74.4	76.5
9402180067	43	56	58/124	46.8	58/125	46.4	46.6
9305040023	35	39	113/157	72.0	113/169	66.9	69.3
9404040040	30	35	106/166	63.9	106/141	75.2	69.1
9401270105	41	42	51/135	37.8	51/108	47.2	42.0
Coreference Totals:			975/1544	63.1%	975/1508	64.7%	63.9%

Appendix C

Training Materials

C.1 The MUC6 training set of 15

The following sections contain the 15 articles used as for training and testing. The shortest and the longest articles also have the co-reference answer keys included as examples. LOLITA's output on the co-reference task for these two articles is also shown.

C.1.1 Document 940124–0017

C.1.1.1 Input for document 940124–0017

```
<DOC>
<DOCID> wsj94_039.0198 </DOCID>
<DOCNO> 940124-0017. </DOCNO>
<HL>    Who's News:
@  Died.... </HL>
<DD> 01/24/94 </DD>
<SO> WALL STREET JOURNAL (J), PAGE B2 </SO>
<TXT>
<p>
Raymond C. Foster Jr., 74, former chairman and chief executive
officer of Stone & Webster Inc., Saturday, in Stuart, Florida, of an
aneurism.
</p>
</TXT>
</DOC>
```

C.1.1.2 Answer key for document 940124–0017

```
<DOC>
<DOCID> wsj94_039.0198 </DOCID>
<DOCNO> 940124-0017. </DOCNO>
<HL>    Who's News:
@  Died.... </HL>
<DD> 01/24/94 </DD>
<SO> WALL STREET JOURNAL (J), PAGE B2 </SO>
<TXT>
<p>
<COREF ID="1" KIND="FLNM" MIN="Raymond C. Foster Jr.|officer|chairman">
Raymond C. Foster Jr., 74, <COREF ID="0" TYPE="IDENT" REF="1" KIND="APNN"
MIN="chairman" STATUS="OPT">former chairman</COREF> and <COREF ID="2"
TYPE="IDENT" REF="1" KIND="APNN" MIN="officer" STATUS="OPT">chief
executive officer of Stone & Webster Inc.</COREF>,</COREF> Saturday,
in Stuart, Florida, of an aneurism.
</p>
</TXT>
</DOC>
```

C.1.1.3 Answer key summary for document 940124–0017

The number on left is the "ID" number of the noun phrase as marked in the key.

```
-------------------------------------------------------------------------

1   Raymond C. Foster Jr., 74, former chairman chief executive
    officer of Stone & Webster Inc.,
0   former chairman
2   chief executive officer of Stone & Webster Inc.

-------------------------------------------------------------------------
```

C.1.1.4 LOLITA's current co-reference task output for document 940124–0017

```
<DOC>
<DOCID> wsj94_039.0198 </DOCID>
<DOCNO> 940124-0017. </DOCNO>
<HL>    Who's News:
@  Died.... </HL>
<DD> 01/24/94 </DD>
<SO> WALL STREET JOURNAL (J), PAGE B2 </SO>
<TXT>
```

```
<p>
   <COREF ID="100001">Raymond C. Foster Jr</COREF>.,
74, <COREF ID="100002" TYPE="IDENT" REF="100001">former chairman</COREF>
and <COREF ID="100003" TYPE="IDENT" REF="100002">chief executive
officer of Stone & Webster Inc</COREF>., Saturday, in Stuart, Florida,
of an aneurism.
</p>
</TXT>
</DOC>
```

C.1.2 Document 940407-0168

```
<DOC>
<DOCID> wsj94_010.0193 </DOCID>
<DOCNO> 940407-0168. </DOCNO>
<HL>    Who's News:
@  Trinzic Corp.</COREF> </HL>
<DD> 04/07/94 </DD>
<SO> WALL STREET JOURNAL (J), PAGE B2 </SO>
<CO>    TRNZ </CO>
<IN> SOFTWARE (SOF) </IN>
<TXT>
<p>
   TRINZIC Corp. (Palo Alto, Calif.) -- This computer-software maker
said that its president, Frank L. Chisholm, 45 years old, resigned
to pursue other interests. Chief Executive Officer Jim Gagnard, 47,
will add the presidency to his duties.
</p>
</TXT>
</DOC>
```

C.1.3 Document 930209-0147

```
<DOC>
<DOCID> wsj93_054.0062 </DOCID>
<DOCNO> 930209-0147. </DOCNO>
<HL>    PNC Financial Changes Name </HL>
<DD> 02/09/93 </DD>
<SO> WALL STREET JOURNAL (J), PAGE A4 </SO>
<CO>    PNC </CO>
<IN> EASTERN U.S. BANKS (BAE), ALL REGIONAL BANKS (BAR),
     ALL BANKS (BNK) </IN>
<DATELINE> PITTSBURGH </DATELINE>
<TXT>
<p>
```

```
   PNC Financial Corp. changed its name to PNC Bank
Corp. and will change the names of its various subsidiaries to PNC
Bank. The name change was the second in 10 years.
</p>
</TXT>
</DOC>
```

C.1.4 Document 940324-0097

```
<DOC>
<DOCID> wsj94_016.0020 </DOCID>
<DOCNO> 940324-0097. </DOCNO>
<HL>    Who's News:
@  Mallinckrodt Group Inc. </HL>
<DD> 03/24/94 </DD>
<SO> WALL STREET JOURNAL (J), PAGE B7 </SO>
<CO>    MKG </CO>
<IN> MEDICAL SUPPLIES (MDS) </IN>
<TXT>
<p>
   MALLINCKRODT GROUP Inc. (St. Louis) -- Michael A. Rocca, 49 years
old, formerly vice president and treasurer of Honeywell Inc., was
named senior vice president and chief financial officer of this
maker of human and animal health-care products and specialty
chemicals.
</p>
</TXT>
</DOC>
```

C.1.5 Document 931021-0152

```
<DOC>
<DOCID> wsj93_095.0131 </DOCID>
<DOCNO> 931021-0152. </DOCNO>
<HL>    Kaman Unit Changes Its Name </HL>
<DD> 10/21/93 </DD>
<SO> WALL STREET JOURNAL (J), NO PAGE CITATION </SO>
<CO>    KAMNA </CO>
<IN> AEROSPACE (ARO) </IN>
<DATELINE> HUDSON, Mass. </DATELINE>
<TXT>
<p>
   EML Research Inc., a unit of Kaman Corp.,
changed its name to Kaman Electromagnetics Corp.
</p>
```

```
<p>
   The company said the name change reflects its growth and
diversification from defense research and development into products
and systems for industrial and commercial markets.
</p>
</TXT>
</DOC>
```

C.1.6 Document 940114-0038

```
<DOC>
<DOCID> wsj94_042.0228 </DOCID>
<DOCNO> 940114-0038. </DOCNO>
<HL>    Business Brief -- Dollar Time Group Inc.:
@  Company Ends Pact to Buy
@  A Real New York Bargain </HL>
<DD> 01/14/94 </DD>
<SO> WALL STREET JOURNAL (J), NO PAGE CITATION </SO>
<CO>    DLRT </CO>
<IN> LIMITED PRODUCT SPECIALTY RETAILERS (OTS),
   ALL SPECIALTY RETAILERS (RTS) </IN>
<TXT>
<p>
   Dollar Time Group Inc. said it ended its agreement to acquire A
Real New York Bargain, a privately held retailer based in New York,
in a transaction originally valued at $43.6 million.
</p>
<p>
   Dollar Time, a Hollywood, Fla., operator of 55 variety stores,
said the two companies disagreed on the direction of Dollar Time's
operations. Moreover, Dollar Time, in a statement, said it had
become "concerned" about the profitability of A Real New York
Bargain since the two sides agreed to join forces in November.
</p>
<p>
   As part of the original agreement, Joseph Sasson, chief executive
officer of A Real New York Bargain, and Jeffrey Klansky, president
of the retailer, became chief executive and president, respectively,
of Dollar Time, as well as directors. Yesterday, the two men
resigned their positions at Dollar Time.
</p>
<p>
   Dollar Time said Gary de Luca, 40 years old, the company's chief
operating officer, was named president.
</p>
</TXT>
</DOC>
```

C.1.7 Document 940425-0043

```
<DOC>
<DOCID> wsj94_004.0015 </DOCID>
<DOCNO> 940425-0043. </DOCNO>
<HL>    Business Brief -- Armco Inc.:
@  Steelmaker to Cut 200 Jobs
@  In Its Salaried Work Force </HL>
<DD> 04/25/94 </DD>
<SO> WALL STREET JOURNAL (J), PAGE A3 </SO>
<CO>    AS </CO>
<IN> STEEL MANUFACTURERS (STL) </IN>
<TXT>
<p>
   Armco Inc. said it will cut about 200 salaried positions, or 3%
of its total work force, in an effort to eliminate redundancies
between corporate and plant administrative positions.
</p>
<p>
   The specialty steelmaker projected that the cuts would save about
$15 million a year after they're completed in the fall.
</p>
<p>
   At its annual meeting, Armco also named John C. Haley, 64 years
old, chairman. Mr. Haley's appointment is for a one-year term,
during which Armco's board will study the concept of a nonexecutive
chairman. Mr. Haley, an Armco board member since 1975, is retired
chairman and chief executive officer of closely held Business
International Corp. He succeeds Robert L. Purdum, 58, who retired.
</p>
<p>
   In addition, Armco named Bruce E. Robbins, 49, and John D.
Turner, 48, to its board. Mr. Robbins is president and chief
executive officer of PNC Bank N.A., a unit of PNC Bank Corp. Mr.
Turner is president and chief executive of Copperweld Corp.
</p>
<p>
   With Mr. Purdum's retirement, Armco's board now has nine members.
</p>
</TXT>
</DOC>
```

C.1.8 Document 930402-0074

```
<DOC>
<DOCID> wsj93_034.0176 </DOCID>
<DOCNO> 930402-0074. </DOCNO>
```

```
<HL>    Time Warner Inc. Boosted
@  Levin's Pay 29% in 1992 </HL>
<DD> 04/02/93 </DD>
<SO> WALL STREET JOURNAL (J), NO PAGE CITATION </SO>
<CO>    TWX </CO>
<IN> ALL ENTERTAINMENT & LEISURE (ENT), MEDIA (MED),
   FILM, TELEVISION & MUSIC (MOV), PUBLISHING (PUB),
   RECREATIONAL PRODUCTS & SERVICES (REC) </IN>
<DATELINE> NEW YORK </DATELINE>
<TXT>
<p>
   Gerald M. Levin, Time Warner Inc.'s chairman, chief
executive officer and president, collected 1992 compensation of $3.6
million, up about 29% from 1991.
</p>
<p>
   According to the big entertainment and publishing company's 1992
proxy statement, Mr. Levin received about $1.1 million in salary and
a bonus of $2.5 million. In addition, Mr. Levin exercised options on
102,764 shares for a paper profit of $2.1 million.
</p>
<p>
   The company cited several factors for the increase, including the
rise in the company's stock price last year, debt restructuring and
completion of a strategic alliance with Toshiba Corp. and Itochu
Corp., formerly C. Itoh & Co. that resulted in a $1 billion
investment. Moreover, Mr. Levin, though filling the post of co-chief
executive officer in 1992, served as the company's sole top officer
for most of the year because of the cancer-related illness of the
late Steven J. Ross, co-chief executive and chairman.
</p>
<p>
   Mr. Levin has since been named to the company's top three posts.
</p>
</TXT>
</DOC>
```

C.1.9 Document 940111-0056

```
<DOC>
<DOCID> wsj94_044.0048 </DOCID>
<DOCNO> 940111-0056. </DOCNO>
<HL>    Business Brief -- McCormick & Co.:
@  Net Income Rose by 16%
@  During the Fourth Quarter </HL>
<DD> 01/11/94 </DD>
<SO> WALL STREET JOURNAL (J), PAGE B4 </SO>
<CO>    MCCRK </CO>
```

```
<IN> FOOD PRODUCTS (FOD), FOOD PRODUCERS, EXCLUDING FISHING (OFP) </IN>
<TXT>
<p>
   McCormick & Co., Sparks, Md., said fourth-quarter net income rose
16%, due to strength in its international, industrial and packaging
operations.
</p>
<p>
   Net for the quarter ended Nov. 30 rose to $40.5 million, or 50
cents a share, from the year-earlier $35 million, or 43 cents a
share. Sales rose 1.9% to $460.8 million from $452.2 million.
</p>
<p>
   Fourth-quarter results include a charge of two cents a share, or
$1.2 million, due to an increase in the federal corporate tax rate
and post-retirement benefits.
</p>
<p>
   Bailey Thomas, the spice company's chairman and chief executive
officer, said the company had a successful year despite onion costs,
unfavorable foreign-exchange rates, and weak economic conditions in
many markets. "We expect 1994 to be another good year," he said.
</p>
<p>
   For the full year, net income fell 23% to $73 million or 89 cents
a share, from the year-earlier $95.2 million, or $1.16 a share. The
latest year includes a $26.6 million, or 33 cent a share, charge for
cumulative effects of accounting changes. Sales rose 5.7% to $1.56
billion from $1.47 billion.
</p>
</TXT>
</DOC>
```

C.1.10 Document 930319-0065

```
<DOC>
<DOCID> wsj93_040.0010 </DOCID>
<DOCNO> 930319-0065. </DOCNO>
<HL>    Business Brief -- Earth Technology Corp.:
@  Firm Expects to Post a Profit
@  For Quarter Ended Feb. 26 </HL>
<DD> 03/19/93 </DD>
<SO> WALL STREET JOURNAL (J), PAGE A5D </SO>
<CO>    ETCO </CO>
<IN> POLLUTION CONTROL, WASTE MANAGEMENT (POL),
   INDUSTRIAL & COMMERCIAL SERVICES (SVC) </IN>
<TXT>
<p>
```

Earth Technology Corp. expects to post net income of between $380,000 and $430,000, or 12 cents to 14 cents a share, for its second quarter, ended Feb. 26, Creighton Early, chief financial officer, said.

That compares with a net loss of $638,000, or 22 cents a share, in the year-earlier period. Revenue for the latest quarter rose to around $15.5 million from $11.7 million, Mr. Early said in an interview in Long Beach, Calif.

For the six months, the company will post net income of between $750,000 and $800,000, or 24 cents to 26 cents a share, on revenue of about $30 million, he said. For the first six months of fiscal 1992, it had a net loss of $576,000, or 19 cents a share, on revenue of $25.7 million.

Diane Creel, recently promoted to chief executive officer, said the return to profitability reflects the success of the company's "back-to-basics" strategy focusing on its core government and commercial hazardous-waste consulting businesses. The company floundered in recent years when its diversification efforts soured as the recession hit.

</TXT>
</DOC>

C.1.11 Document 940310-0087

<DOC>
<DOCID> wsj94_021.0167 </DOCID>
<DOCNO> 940310-0087. </DOCNO>
<HL> Who's News:
@ Madison Group Says
@ Board Has Dismissed
@ Lucas as Its President </HL>
<DD> 03/10/94 </DD>
<SO> WALL STREET JOURNAL (J), PAGE B10 </SO>
<CO> MADI </CO>
<IN> REAL ESTATE INVESTMENTS (REA) </IN>
<DATELINE> FORT LAUDERDALE, Fla. </DATELINE>
<TXT>

Madison Group Associates Inc. said its board dismissed Kenneth Lucas, president, naming Dean J. Trantalis as interim president.

The company also said two new directors -- Roland Breton and Steve Gibboney -- had been appointed to its board.

Mr. Lucas became chief executive of the media concern less than two months ago, when William T. Craig resigned from his job as a director and chief executive officer.

The company gave no reason for Mr. Lucas's dismissal. Neither he nor Madison executives could be reached for comment.

The management change is the latest in a series of events that have shaken the company in recent months. As previously reported, the Securities and Exchange Commission contacted several individuals about their dealings with the company. One of those individuals said the SEC had asked about how the company valued its assets.

Those assets consist largely of video libraries. According to a recent securities filing, an accountant formerly hired by Madison recommended that an independent specialist be hired to evaluate the video libraries.

</TXT>
</DOC>

C.1.12 Document 940325-0108

<DOC>
<DOCID> wsj94_015.0175 </DOCID>
<DOCNO> 940325-0108. </DOCNO>
<HL> Sandoz's 1993 Profit
@ Rose 14%; Company
@ Appoints New CEO </HL>
<DD> 03/25/94 </DD>
<SO> WALL STREET JOURNAL (J), PAGE A5 </SO>
<CO> Z.SAN </CO>
<IN> DRUG MANUFACTURERS (DRG) </IN>
<DATELINE> ZURICH </DATELINE>
<TXT>

Sandoz AG announced a 14% jump in 1993 net income and increased its dividend. The Swiss chemical and pharmaceutical group also appointed Rolf W. Schweizer chief executive officer.

```
<p>
   Although analysts widely expected a double-digit rise in
earnings, they said a dividend improvement of 23% to 58 Swiss francs
($40.62) a share from 47 francs was greater than expected.
</p>
<p>
   Sandoz also said its share capital would be simplified by
converting nonvoting participation certificates into registered
shares. The remaining two categories of registered and bearer shares
would be split in a ratio of 1 to 5.
</p>
<p>
   Sandoz's profit rose to 1.71 billion francs in 1993 from 1.5
billion francs a year earlier. The company said the growth was aided
by increased drug sales and improved margins in the
chemicals-and-environment sector.
</p>
<p>
   The appointment of the 63-year-old Mr. Schweizer as chief
executive also was unexpected. He will succeed Marc Moret, 70, who
kept the chairman and chief executive positions after Daniel C.
Wagniere, 57, became chief operating officer last year. Up until
then, Mr. Moret had held all three positions. Mr. Moret will
continue as chairman.
</p>
</TXT>
</DOC>
```

C.1.13 Document 930927-0024

```
<DOC>
<DOCID> wsj93_105.0058 </DOCID>
<DOCNO> 930927-0024. </DOCNO>
<HL>    Business Brief -- Eastco Industrial Safety Corp.:
@  Company Expects to Post
@  A Loss for Its Fiscal Year </HL>
<DD> 09/27/93 </DD>
<SO> WALL STREET JOURNAL (J), NO PAGE CITATION </SO>
<CO>    ESTO </CO>
<IN> GENERAL INDUSTRIAL & COMMERCIAL SERVICES (ICS),
   ALL INDUSTRIAL & COMMERCIAL SERVICES (SVC) </IN>
<TXT>
<p>
   Eastco Industrial Safety Corp., Huntington Station, N.Y., said it
expects to post a loss for fiscal 1993 and "disappointing" results
for fiscal first quarter of 1994.
</p>
<p>
```

The maker and distributor of industrial safety products said its president and chief executive officer stepped down.

Joel Poznansky, 35 years old, who joined the company in July as vice president of manufacturing, was named president and chief executive officer, succeeding Alan E. Densen, 60, who became senior vice president. Mr. Densen will continue as a director and chairman.

Eastco said it's likely the company will report a loss for fiscal 1994, which ends in June. It attributed the poor performance to "weak industry conditions and a shortage of working capital."

In order to improve its working capital, Eastco extended its credit arrangements with its principal lender by another year, and the senior management collectively loaned $250,000 to the company, an amount which was matched by the bank. The company, which employs about 200 people, said it's implementing a series of cost-cutting measures, including layoffs.

</TXT>
</DOC>

C.1.14 Document 940302-0061

<DOC>
<DOCID> wsj94_024.0274 </DOCID>
<DOCNO> 940302-0061. </DOCNO>
<HL> AT&T Plans to Hang Up
@ Its Real Corporate Name </HL>
<DD> 03/02/94 </DD>
<SO> WALL STREET JOURNAL (J), PAGE A4 </SO>
<CO> T </CO>
<IN> COMMUNICATIONS TECHNOLOGY (CMT), COMPUTERS (CPR),
LONG DISTANCE TELEPHONE PROVIDERS (LDS), TELEPHONE SYSTEMS (TLS) </IN>
<DATELINE> NEW YORK </DATELINE>
<TXT>

Another venerable corporate name belongs to the ages:
American Telephone & Telegraph Co. wants to be just plain AT&T Corp.

The company, of course, has an unsentimental explanation for wanting to change its name. It's more practical. And most people know the telecommunications giant by its initials.

```
<p>
   "It's the perfect time, being that it's 10 years since our
breakup and telephone and telegraph doesn't describe the company we
are today," says Marilyn Laurie, AT&T's senior vice president of
public relations. AT&T's own survey showed that 58% of people over
18 years old don't even know what the letters stand for.
</p>
<p>
   What they used to stand for was a name that symbolized a bygone
American era, when unpretentious people with names like Bell and
Edison could invent things that became the foundations for corporate
empires.
</p>
<p>
   American Telephone & Telegraph Co. became the biggest of them
all: When it was broken up in 1984, AT&T was the world's largest
company with more than a million employees, revenues that topped
$100 billion and what is still widely regarded as the world's
premiere research and development organization.
</p>
</TXT>
</DOC>
```

C.1.15 Document 930412–0090

C.1.15.1 Input for document 930412–0090

```
<DOC>
<DOCID> wsj93_031.0105 </DOCID>
<DOCNO> 930412-0090. </DOCNO>
<HL>    Who's News:
@  Spelling Entertainment
@  Picks New Chairman,
@  CEO for Blockbuster </HL>
<DD> 04/12/93 </DD>
<SO> WALL STREET JOURNAL (J), PAGE B5 </SO>
<CO>    BV CHR SP </CO>
<IN> ALL ENTERTAINMENT & LEISURE (ENT), FILM,
   TELEVISION & MUSIC (MOV),
   LIMITED PRODUCT SPECIALTY RETAILERS (OTS),
   RECREATIONAL PRODUCTS & SERVICES (REC),
   ALL SPECIALTY RETAILERS (RTS) </IN>
<DATELINE> LOS ANGELES </DATELINE>
<TXT>
<p>
   Spelling Entertainment Group Inc. named H. Wayne
Huizenga, chairman and chief executive officer of Blockbuster
```

```
Entertainment Corp., following Blockbuster's recent purchase of a
controlling interest in Spelling.
</p>
<p>
   He succeeds investor Carl H. Lindner, whose closely-held American
Financial Corp. had held a 48.2% stake in Spelling that Blockbuster
acquired last month in a $140 million stock swap. Combined with
Spelling shares previously acquired by Blockbuster, the Fort
Lauderdale, Fla., video store chain now holds a 53.8% stake in
Spelling, a producer and distributor of films and television
programs, including the current Fox shows "Beverly Hills 90210" and
"Melrose Place." Television producer Aaron Spelling, the company's
founder, was named vice chairman and a member of the board's
executive committee.
</p>
<p>
   Steven R. Berrard, Blockbuster's vice chairman, president and
chief operating officer, was named president, chief executive and a
director of Spelling, succeeding Mr. Lindner's son S. Craig Lindner,
who remains a board member.
</p>
</TXT>
</DOC>
```

C.1.15.2 Answer key for document 930412–0090

```
<DOC>
<DOCID> wsj93_031.0105 </DOCID>
<DOCNO> 930412-0090. </DOCNO>
<HL>    Who's News:
@  <COREF ID="1" KIND="ABNM">Spelling Entertainment</COREF>
@  Picks <COREF ID="40" TYPE="IDENT" REF="3" KIND="NDNN"
   MIN="Chairman">New Chairman</COREF>,
@  <COREF ID="5" TYPE="IDENT" REF="3" KIND="NDNN" MIN="CEO">CEO for
   <COREF ID="7" KIND="ABNM">Blockbuster</COREF></COREF> </HL>
<DD> 04/12/93 </DD>
<SO> WALL STREET JOURNAL (J), PAGE B5 </SO>
<CO>    BV CHR SP </CO>
<IN> ALL ENTERTAINMENT & LEISURE (ENT), FILM,
   TELEVISION & MUSIC (MOV),
   LIMITED PRODUCT SPECIALTY RETAILERS (OTS),
   RECREATIONAL PRODUCTS & SERVICES (REC),
   ALL SPECIALTY RETAILERS (RTS) </IN>
<DATELINE> LOS ANGELES </DATELINE>
<TXT>
<p>
   <COREF ID="0" TYPE="IDENT" REF="1" KIND="FLNM">Spelling Entertainment
Group Inc.</COREF> named <COREF ID="3" KIND="FLNM">H. Wayne
```

```
Huizenga</COREF>, <COREF ID="2" TYPE="IDENT" REF="3" KIND="TRNN"
STATUS="OPT">chairman</COREF> and <COREF ID="4" TYPE="IDENT" REF="3"
MIN="officer" KIND="TRNN" STATUS="OPT">chief executive officer of
<COREF ID="6" TYPE="IDENT" REF="7" KIND="FLNM">Blockbuster Entertainment
Corp.</COREF></COREF>, following <COREF ID="8" TYPE="IDENT" REF="6"
KIND="ABNM">Blockbuster</COREF>'s recent purchase of <COREF ID="12"
KIND="IANN" MIN="interest">a controlling interest in <COREF ID="9"
TYPE="IDENT" REF="0" KIND="ABNM">Spelling</COREF></COREF>.
</p>
<p>
   <COREF ID="10" TYPE="IDENT" REF="3" KIND="PRPR">He</COREF> succeeds
<COREF ID="39" KIND="FLNM" MIN="Carl H. Lindner">investor
Carl H. Lindner</COREF>, whose closely-held American Financial Corp. had
held <COREF ID="11" TYPE="IDENT" REF="12" KIND="HRNN" MIN="stake">a
<COREF ID="13" TYPE="IDENT" REF="11" KIND="PERC">48.2%</COREF> stake in
<COREF ID="14" TYPE="IDENT" REF="9" KIND="ABNM">Spelling</COREF> that
<COREF ID="15" TYPE="IDENT" REF="8" KIND="ABNM">Blockbuster</COREF>
acquired last month in <COREF ID="17" KIND="IANN" MIN="swap">a
<COREF ID="16" TYPE="IDENT" REF="17" KIND="ABSV">$140 million</COREF>
stock swap</COREF></COREF>. Combined with <COREF ID="18" TYPE="IDENT"
REF="14" KIND="ABNM">Spelling</COREF> shares previously acquired by
<COREF ID="19" TYPE="IDENT" REF="15" KIND="ABNM">Blockbuster</COREF>,
<COREF ID="20" TYPE="IDENT" REF="19" KIND="DANN" MIN="chain">the Fort
Lauderdale, Fla., video store chain</COREF> now holds <COREF ID="22"
KIND="IANN" MIN="stake">a <COREF ID="21" TYPE="IDENT" REF="22"
KIND="PERC">53.8%</COREF> stake in <COREF ID="23" TYPE="IDENT" REF="18"
KIND="ABNM" MIN="Spelling">Spelling, a producer and distributor of films
and <COREF ID="25" KIND="PRNN">television</COREF> programs, including
the current Fox shows "Beverly Hills 90210" and "Melrose Place."</COREF>
</COREF> <COREF ID="28" KIND="FLNM" MIN="Aaron Spelling|founder"><COREF
ID="24" TYPE="IDENT" REF="25" KIND="PRNN">Television</COREF> producer
Aaron Spelling, <COREF ID="27" TYPE="IDENT" REF="28" KIND="APNN"
MIN="founder"><COREF ID="26" TYPE="IDENT" REF="23" KIND="DANN">the
company</COREF>'s founder</COREF></COREF>, was named <COREF ID="29"
TYPE="IDENT" REF="28" MIN="chairman"  KIND="TRNN" STATUS="OPT">vice
chairman</COREF> and a member of <COREF ID="41" KIND="DANN">the
board</COREF>'s executive committee.
</p>
<p>
   <COREF ID="31" KIND="FLNM" MIN="chairman|Steven R. Berrard|officer">
Steven R. Berrard, <COREF ID="30" TYPE="IDENT" REF="31" MIN="chairman"
KIND="APNN" STATUS="OPT"><COREF ID="32" TYPE="IDENT" REF="19"
KIND="ABNM">Blockbuster</COREF>'s vice chairman</COREF>, <COREF ID="33"
TYPE="IDENT" REF="31" KIND="APNN" STATUS="OPT">president</COREF> and
<COREF ID="34" TYPE="IDENT" REF="31" MIN="officer" KIND="APNN"
STATUS="OPT">chief operating officer</COREF></COREF>, was named
<COREF ID="35" TYPE="IDENT" REF="31" KIND="TRNN">president</COREF>,
<COREF ID="36" TYPE="IDENT" REF="31" KIND="TRNN" MIN="executive">chief
executive</COREF> and a director of <COREF ID="37" TYPE="IDENT" REF="26"
KIND="ABNM">Spelling</COREF>, succeeding <COREF ID="38" TYPE="IDENT"
```

```
REF="39" KIND="FLNM" MIN="Lindner">Mr. Lindner</COREF>'s son S. Craig
Lindner, who remains a <COREF ID="42" TYPE="IDENT" REF="41"
KIND="PRNN">board</COREF> member.
</p>
</TXT>
</DOC>
```

C.1.15.3 Anwer key summary for document 930412-0090

```
---------------------------------------------------------------------

 1   Spelling Entertainment
 0   Spelling Entertainment Group Inc.
 9   Spelling
 14  Spelling
 18  Spelling
 23  Spelling, a producer and distributor of films and television
     programs, including the current Fox shows "Beverly Hills 90210"#
     and "Melrose Place."
 26  the company
 37  Spelling

---------------------------------------------------------------------

 7   Blockbuster
 6   Blockbuster Entertainment Corp.
 8   Blockbuster
 15  Blockbuster
 19  Blockbuster
 20  the Fort Lauderdale, Fla., video store chain
 32  Blockbuster

---------------------------------------------------------------------

 3   H. Wayne Huizenga
 5   CEO Blockbuster
 2   chairman
 40  New Chairman
 4   chief executive officer of Blockbuster Entertainment Corp.
 10  He

---------------------------------------------------------------------

 12  a controlling interest in Spelling
 11  a 48.2% Spelling that Blockbuster acquired last month in a
     $140 million stock swap
 13  48.2%
```

---

39 investor Carl H. Lindner
38 Mr. Lindner

---

17 a $140 million stock swap
16 $140 million

---

25 television
24 Television

---

22 a 53.8% Spelling, a producer and distributor of films
and television programs, including the current Fox shows
"Beverly Hills 90210" and "Melrose Place."
21 53.8%

---

28 Television producer Aaron Spelling, the company's founder
29 vice chairman
27 the company's founder

---

41 the board
42 board

---

31 Steven R. Berrard, Blockbuster's vice chairman, president
and chief operating officer
35 president
36 chief executive
33 president
34 chief operating officer
30 Blockbuster's vice chairman

C.1.15.4 LOLITA's current co-reference task output for document 930412–0090

```
<DOC>
<DOCID> wsj93_031.0105 </DOCID>
<DOCNO> 930412-0090. </DOCNO>
<HL>    Who's News:
@  <COREF ID="100007" TYPE="IDENT" REF="100006">Spelling
Entertainment</COREF> Picks <COREF ID="200003" TYPE="IDENT"
REF="200002">New Chairman</COREF>, <COREF ID="200004" TYPE="IDENT"
REF="200003">CEO for <COREF ID="300007" TYPE="IDENT" REF="300006">
Blockbuster</COREF></COREF> </HL>
<DD> 04/12/93 </DD>
<SO> WALL STREET JOURNAL (J), PAGE B5 </SO>
<CO>    BV CHR SP </CO>
<IN> ALL ENTERTAINMENT & LEISURE (ENT), FILM,
   TELEVISION & MUSIC (MOV),
   LIMITED PRODUCT SPECIALTY RETAILERS (OTS),
   RECREATIONAL PRODUCTS & SERVICES (REC),
   ALL SPECIALTY RETAILERS (RTS) </IN>
<DATELINE> LOS ANGELES </DATELINE>
<TXT>
<p>
   <COREF ID="100001">Spelling Entertainment Group Inc</COREF>.
named <COREF ID="200001">H. Wayne
Huizenga</COREF>, <COREF ID="200005" TYPE="IDENT" REF="200004">
chairman</COREF> and <COREF ID="200006" TYPE="IDENT" REF="200005">
chief executive officer of <COREF ID="300001">Blockbuster
Entertainment Corp</COREF></COREF>., following <COREF ID="300002"
TYPE="IDENT" REF="300001">Blockbuster</COREF>'s recent purchase of a
controlling interest in <COREF ID="100002" TYPE="IDENT" REF="100001">
Spelling</COREF>.
</p>
<p>
   <COREF ID="200002" TYPE="IDENT" REF="200001">He</COREF> succeeds
investor <COREF ID="400001">Carl H. Lindner</COREF>, whose closely-held
American Financial Corp. had held <COREF ID="500002" TYPE="IDENT"
REF="500001">a <COREF ID="500001">48.2%</COREF> stake</COREF> in
<COREF ID="100003" TYPE="IDENT" REF="100002">Spelling</COREF> that
<COREF ID="300003" TYPE="IDENT" REF="300002">Blockbuster</COREF>
acquired last month in <COREF ID="600001">a <COREF ID="600002"
TYPE="IDENT" REF="600001">$140 million</COREF> stock swap</COREF>.
Combined with <COREF ID="100004" TYPE="IDENT" REF="100003">Spelling
</COREF> shares previously acquired by <COREF ID="300004" TYPE="IDENT"
REF="300003">Blockbuster</COREF>, <COREF ID="300005" TYPE="IDENT"
REF="300004">the Fort Lauderdale, Fla., video store chain</COREF>
now holds <COREF ID="700002" TYPE="IDENT" REF="700001">a <COREF
ID="700001">53.8%</COREF> stake in <COREF ID="100005" TYPE="IDENT"
REF="100004">Spelling</COREF>, a producer and distributor of films
```

```
and <COREF ID="800001">television</COREF> programs, including the
current <COREF ID="1100001">Fox</COREF> shows "Beverly Hills 90210"
and "Melrose Place</COREF>." <COREF ID="800002" TYPE="IDENT"
REF="800001">Television</COREF> producer <COREF ID="900001">Aaron
Spelling</COREF>, <COREF ID="900003" TYPE="IDENT" REF="900002">the
<COREF ID="1100002" TYPE="IDENT" REF="1100001">company</COREF>'s
founder</COREF>, was named <COREF ID="900002" TYPE="IDENT"
REF="900001">vice chairman</COREF> and a member of the <COREF
ID="1200001">board</COREF>'s executive committee.
</p>
<p>
   <COREF ID="1000001">Steven R. Berrard</COREF>, <COREF ID="1000002"
TYPE="IDENT" REF="1000001"><COREF ID="300006" TYPE="IDENT" REF="300005">
Blockbuster</COREF>'s vice chairman</COREF>, <COREF ID="1000003"
TYPE="IDENT" REF="1000002">president</COREF> and <COREF ID="1000004"
TYPE="IDENT" REF="1000003">chief operating officer</COREF>, was named
<COREF ID="1000006" TYPE="IDENT" REF="1000005">president</COREF>,
<COREF ID="1000005" TYPE="IDENT" REF="1000004">chief executive</COREF>
and a director of <COREF ID="100006" TYPE="IDENT" REF="100005">Spelling
</COREF>, succeeding Mr. <COREF ID="400002" TYPE="IDENT" REF="400001">
Lindner</COREF>'s son S. Craig Lindner, who remains a <COREF ID="1200002"
TYPE="IDENT" REF="1200001">board</COREF> member.
</p>
</TXT>
</DOC>
```

C.2 Other training materials

C.2.1 The long article from *The Guardian*

C.2.1.1 Input for the *Guardian* article

```
<DOC>
<DOCID> guardian.june.1.1996 </DOCID>
<DOCNO> 01061996. </DOCNO>
<HL> Facia saviour is bankrupt. Grosvenor says firm was not told. </HL>
<DD> 01/06/96 </DD>
<AUTHOR>Patrick Donovan, City editor</AUTHOR>

<TXT>
<p> William Grosvenor, the entrepreneur and well connected cousin of
the Duke of Westminster, who is heading attempts by Texas American
Group to take over the troubled Facia retail company, last night admitted
```

that he is bankrupt.

 Mr Grosvenor, aged 54, a pageboy at the Queen's 1953 Coronation, who is acting as chief executive officer of the US-listed company trying to buy Britain's second largest privately owned retail chain, has a spent conviction for tax fraud in Britain, for which he received a 12 month suspended sentence and a #1,000 fine in 1980.

 He pleaded guilty to plotting to defraud the Inland Revenue by attempting to pass off the #8,500 costs of a grouse shoot as a tax deductible business expense. The conviction is considered spent under the Rehabilitation of Offenders Act.

 Texas American announced on Thursday that it had agreed to buy Facia, which had been seeking capital. Facia, headed by Stephen Hinchliffe, operates 850 speciality shops, with the high-street names including Sock Shop, Salisburys, Red or Dead, and Contessa.

 Mr Hinchliffe, whose affairs have been investigated by the Department of Trade and Industry, also controls shoe shops including Freeman, Hardy & Willis and Curtess, which he took over from Sears.

 As the crisis surrounding Facia escalated, Sears yesterday served a petition in the High Court to put the shoe operations into administration. Sears still has an interest in the companies as part of a staggered settlement deal. It is putting #25 million aside to cover the disposal costs and is suspending plans to sell a further interest in the Saxone chain to Facia.

 It is understood Sears has been increasingly concerned at the running of Facia and took the action after the company defaulted on #4 million worth of rental payments which became due yesterday. Sears is believed to have little confidence in Texas American's plans for the company.

 Mr Grosvenor yesterday accepted that his bankruptcy could affect his credibility as a businessman. He added that he had told his US partners but had not informed Facia about his bankruptcy.

 Helen Clark, a lawyer at the City firm of Eversheds, said that she had made a bankruptcy order in the High Court on 19 October 1994. "As yet there has been no distribution to creditors and Mr Grosvenor remains bankrupt", she said.

 Later that year, on November 24, a Dublin solicitor, Denis Murnaghan, obtained judgment against Mr Grosvenor and two other defendants for #350,000 plus costs which have been estimated at more than #100,000. The judgment remains unpaid.

 Mr Grosvenor yesterday said he was not a director of the Texas

American Group but was acting as its chief executive. He added: "We have come to an agreement to make a contract (to take over Facia)". The deal would be funded out of the company's own resources.

 Mr Grosvenor said that the company had not submitted up to date filings with the US Securities and Exchange Commission because of the need to take into account recent acquisitions.

 According to other documents about Texas American filed in Washington, the company is said to have interests in Internet lottery and casino games. It has stakes in a Nevada hotel development and Portugese holiday businesses.

 Facia last night declined to make any comment. The company has been looking for a capital injection of around #40 million. It is more than six months late filing its accounts for the year ending January 1995. Accounts for its Sock Shop subsidiary are heavily qualified by the auditors.

 The DTI investigation into Mr Hinchliffe's affairs are understood to focus on the 1993 collapse of Boxgrey, a company sold by the Sheffield-based entrepreneur shortly before it collapsed.

 Mr Grosvenor is known in the City as an entrepreneur who has also worked as a financial public relations adviser. His name regularly appears in newspaper social pages because of his family connections. He is related to the Aga Khan as well as the Duke of Westminster. He was married in 1966 to Ellen Seeliger, daughter of Germany's Ambassador to Mexico. His mother was one of the four daughters of the third Lord Churston.
</TXT>
</DOC>

C.2.1.2 Answer key for the *Guardian* article

<DOC>
<DOCID> guardian.june.1.1996 </DOCID>
<DOCNO> 01061996. </DOCNO>
<HL> <COREF ID="1" KIND="NDNN" MIN="saviour"><COREF ID="3" KIND="FLNM">
Facia</COREF> saviour</COREF> is bankrupt. </HL> <HL> <COREF ID="0"
TYPE="IDENT" REF="1" KIND="ABNM">Grosvenor</COREF> says <COREF ID="2"
TYPE="IDENT" REF="3" KIND="NDNN">firm</COREF> was not told. </HL>
<DD> 01/06/96 </DD>

<AUTHOR><COREF ID="106" KIND="FLNM" MIN="Patrick Donovan|editor">Patrick
Donovan, <COREF ID="107" TYPE="IDENT" REF="106" KIND="NDNN" MIN="editor">
<COREF ID="112" KIND="PRNN">City</COREF> editor</COREF></COREF></AUTHOR>

```
<TXT>
<p> <COREF ID="4" TYPE="IDENT" REF="0" KIND="FLNM" MIN="William
Grosvenor|cousin">William Grosvenor, <COREF ID="5" TYPE="IDENT" REF="4"
KIND="APNN">the entrepreneur</COREF> and <COREF ID="6" TYPE="IDENT"
REF="4"KIND="APNN" MIN="cousin">well connected cousin of <COREF ID="97"
KIND="FLNM" MIN="Duke of Westminster">the Duke of Westminster</COREF>
</COREF>, who is heading attempts by <COREF ID="12" KIND="FLNM">Texas
American Group</COREF> to take over <COREF ID="7" TYPE="IDENT" REF="3"
KIND="DANN" MIN="company|Facia">the troubled Facia retail company</COREF>
</COREF>, last night admitted that <COREF ID="8" TYPE="IDENT" REF="4"
KIND="PRPR">he</COREF> is bankrupt. </p>


<p> <COREF ID="9" TYPE="IDENT" REF="8" KIND="ABNM" MIN="Grosvenor|
officer">Mr Grosvenor, aged 54, a pageboy at the Queen's 1953 Coronation,
who is acting as <COREF ID="10" TYPE="IDENT" REF="9" KIND="NDNN"
MIN="officer">chief executive officer of <COREF ID="11" TYPE="IDENT"
REF="12" KIND="DANN" MIN="company">the <COREF ID="110" KIND="FLNM">
US</COREF>-listed company trying to buy <COREF ID="13" TYPE="IDENT"
REF="7" KIND="SGNN" MIN="chain"><COREF ID="108" KIND="FLNM">Britain
</COREF>'s second largest privately owned retail chain</COREF></COREF>
</COREF></COREF>, has <COREF ID="17" KIND="IANN" MIN="conviction">a spent
conviction for tax fraud in <COREF ID="109" TYPE="IDENT" REF="108"
KIND="FLNM">Britain</COREF>, for which <COREF ID="14" TYPE="IDENT"
REF="9" KIND="PRPR">he</COREF> received a 12 month suspended sentence
and a #1,000 fine in 1980</COREF>.</p>

<p> <COREF ID="15" TYPE="IDENT" REF="14" KIND="PRPR">He</COREF>
pleaded guilty to plotting to defraud the Inland Revenue by
attempting to pass off the <COREF ID="103" KIND="OTHV" MIN="costs">
<COREF ID="102" TYPE="IDENT" REF="103" KIND="ABSV">#8,500</COREF>
costs of a grouse shoot</COREF> as a tax deductible business expense.
<COREF ID="16" TYPE="IDENT" REF="17" KIND="DANN">The conviction</COREF>
is considered spent under the Rehabilitation of Offenders Act.</p>

<p> <COREF ID="18" TYPE="IDENT" REF="11" KIND="ABNM">Texas American
</COREF> announced on Thursday that <COREF ID="19" TYPE="IDENT" REF="18"
KIND="PRPR">it</COREF> had agreed to buy <COREF ID="20" TYPE="IDENT"
REF="13" KIND="FLNM" MIN="Facia">Facia, which had been seeking
capital</COREF>.  <COREF ID="21" TYPE="IDENT" REF="20" KIND="FLNM"
MIN="Facia">Facia, headed by <COREF ID="23" KIND="FLNM">Stephen
Hinchliffe</COREF></COREF>, operates 850 speciality shops, with the
high-street names including <COREF ID="104" KIND="FLNM">Sock
Shop</COREF>, Salisburys, Red or Dead, and Contessa.</p>

<p> <COREF ID="22" TYPE="IDENT" REF="23" KIND="ABNM" MIN="Hinchliffe">Mr
Hinchliffe, whose <COREF ID="116" KIND="PSNN">affairs</COREF> have been
investigated by <COREF ID="87" KIND="FLNM">the Department of Trade and
Industry</COREF></COREF>, also controls <COREF ID="29" KIND="NDPN"
MIN="shops"><COREF ID="31" KIND="PRNN">shoe</COREF> shops including
```

```
Freeman, Hardy & Willis and Curtess, which <COREF ID="24" TYPE="IDENT"
REF="22" KIND="PRPR">he</COREF> took over from <COREF ID="27"
KIND="FLNM">Sears</COREF></COREF>.</p>

<p> As the crisis surrounding <COREF ID="25" TYPE="IDENT" REF="21"
KIND="FLNM">Facia</COREF> escalated, <COREF ID="26" TYPE="IDENT" REF="27"
KIND="FLNM">Sears</COREF> yesterday served a petition in <COREF ID="57"
KIND="FLNM">the High Court</COREF> to put <COREF ID="28" TYPE="IDENT"
REF="29" KIND="DANN" MIN="operations">the <COREF ID="30" TYPE="IDENT"
REF="31" KIND="PRNN">shoe</COREF> operations</COREF> into administration.
<COREF ID="32" TYPE="IDENT" REF="26" KIND="FLNM">Sears</COREF> still has
an interest in the companies as part of a staggered settlement deal.
<COREF ID="33" TYPE="IDENT" REF="32" KIND="PRPR">It</COREF> is putting
#25 million aside to cover the disposal costs and is suspending plans to
sell a further interest in the Saxone chain to <COREF ID="34"
TYPE="IDENT" REF="25" KIND="FLNM">Facia</COREF>.</p>

<p> It is understood <COREF ID="35" TYPE="IDENT" REF="33"
KIND="FLNM">Sears</COREF> has been increasingly concerned at the running
of <COREF ID="36" TYPE="IDENT" REF="34" KIND="FLNM">Facia</COREF> and
took the action after <COREF ID="37" TYPE="IDENT" REF="36"
KIND="DANN">the company</COREF> defaulted on <COREF ID="39" KIND="OTHV"
MIN="#4 million|payments"><COREF ID="38" TYPE="IDENT" REF="39"
KIND="ABSV" MIN="#4 million">#4 million</COREF> worth of rental payments
which became due yesterday</COREF>.  <COREF ID="40" TYPE="IDENT" REF="35"
KIND="FLNM">Sears</COREF> is believed to have little confidence in <COREF
ID="41" TYPE="IDENT" REF="19" KIND="ABNM">Texas American</COREF>'s plans
for <COREF ID="42" TYPE="IDENT" REF="37" KIND="DANN">the
company</COREF>.</p>

<p> <COREF ID="43" TYPE="IDENT" REF="15" KIND="ABNM" MIN="Grosvenor">Mr
Grosvenor</COREF> yesterday accepted that <COREF ID="53"
KIND="PSNN"><COREF ID="44" TYPE="IDENT" REF="43" KIND="PSDT">his</COREF>
bankruptcy</COREF> could affect <COREF ID="45" TYPE="IDENT" REF="44"
KIND="PSDT">his</COREF> credibility as a businessman.  <COREF ID="46"
TYPE="IDENT" REF="45" KIND="PRPR">He</COREF> added that <COREF ID="47"
TYPE="IDENT" REF="46" KIND="PRPR">he</COREF> had told <COREF ID="49"
TYPE="IDENT" REF="41" KIND="PSNN" MIN="partners" STATUS="OPT"><COREF
ID="48" TYPE="IDENT" REF="47" KIND="PSDT">his</COREF> <COREF ID="111"
TYPE="IDENT" REF="110" KIND="FLNM">US</COREF> partners</COREF> but had
not informed <COREF ID="50" TYPE="IDENT" REF="42"
KIND="FLNM">Facia</CO="IDENT" REF="48" KIND="PSNN">his</COREF>
bankruptcy</COREF>.</p>

<p> <COREF ID="55" KIND="FLNM" MIN="Helen Clark">Helen Clark, a lawyer at
the <COREF ID="113" TYPE="IDENT" REF="112" KIND="PRNN">City</COREF> firm
of Eversheds</COREF>, said that <COREF ID="54" TYPE="IDENT" REF="55"
KIND="PRPR">she</COREF> had made a <COREF ID="118" TYPE="IDENT" REF="53"
KIND="PRNN">bankruptcy</COREF> order in <COREF ID="56" TYPE="IDENT"
REF="57" KIND="FLNM">the High Court</COREF> on 19 October 1994.  "As yet
```

```
there has been no distribution to creditors and <COREF ID="58"
TYPE="IDENT" REF="51" KIND="ABNM" MIN="Grosvenor">Mr Grosvenor</COREF>
remains bankrupt", <COREF ID="59" TYPE="IDENT" REF="54"
KIND="PRPR">she</COREF> said.</p>

<p> Later that year, on November 24, a Dublin solicitor, Denis Murnaghan,
obtained <COREF ID="64" KIND="NDNN" MIN="judgment">judgment against <COREF
ID="60" TYPE="IDENT" REF="58" KIND="ABNM" MIN="Grosvenor">Mr
Grosvenor</COREF> and two other defendants for #350,000 plus <COREF ID="62"
KIND="HRNN" MIN="costs">costs which have been estimated at <COREF ID="61"
TYPE="IDENT" REF="62" KIND="OTHV">more than #100,000</COREF></COREF>
</COREF>.  <COREF ID="63" TYPE="IDENT" REF="64" KIND="DANN">The
judgment</COREF> remains unpaid.</p>

<p> <COREF ID="65" TYPE="IDENT" REF="60" KIND="ABNM" MIN="Grosvenor">Mr
Grosvenor</COREF> yesterday said <COREF ID="66" TYPE="IDENT" REF="65"
KIND="PRPR">he</COREF> was not a director of <COREF ID="67" TYPE="IDENT"
REF="41" KIND="FLNM" MIN="Texas American Group">the Texas American
Group</COREF> but was acting as <COREF ID="68" TYPE="IDENT" REF="66"
KIND="PSNN" STATUS="OPT"><COREF ID="115" TYPE="IDENT" REF="67"
KIND="PSDT">its</COREF> chief executive</COREF>.  <COREF ID="69"
TYPE="IDENT" REF="66" KIND="PRPR">He</COREF> added: "<COREF ID="119"
TYPE="IDENT" REF="67" KIND="PRPR">We</COREF> have come to <COREF ID="72"
KIND="IANN" MIN="agreement">an agreement to make a contract (to take over
<COREF ID="70" TYPE="IDENT" REF="50" KIND="FLNM">Facia</COREF>)</COREF>".
<COREF ID="71" TYPE="IDENT" REF="72" KIND="DANN" STATUS="OPT">The deal
</COREF> would be funded out of <COREF ID="73" TYPE="IDENT" REF="67"
KIND="DANN">the company</COREF>'s own resources.</p>

<p> <COREF ID="74" TYPE="IDENT" REF="69" KIND="ABNM" MIN="Grosvenor">Mr
Grosvenor</COREF> said that <COREF ID="75" TYPE="IDENT" REF="73"
KIND="DANN">the company</COREF> had not submitted up to date filings with
the US Securities and Exchange Commission because of the need to take into
account recent acquisitions.</p>

<p> According to other documents about <COREF ID="76" TYPE="IDENT"
REF="75" KIND="ABNM">Texas American</COREF> filed in Washington, <COREF
ID="77" TYPE="IDENT" REF="76" KIND="DANN">the company</COREF> is said to
have interests in Internet lottery and casino games.  <COREF ID="78"
TYPE="IDENT" REF="77" KIND="PRPR">It</COREF> has stakes in a Nevada hotel
development and Portugese holiday businesses.</p>

<p> <COREF ID="79" TYPE="IDENT" REF="70" KIND="FLNM">Facia</COREF> last
night declined to make any comment.  <COREF ID="80" TYPE="IDENT" REF="79"
KIND="DANN">The company</COREF> has been looking for <COREF ID="82"
KIND="IANN" MIN="injection">a capital injection of <COREF ID="81"
TYPE="IDENT" REF="82" KIND="OTHV">around #40 million</COREF></COREF>.
<COREF ID="83" TYPE="IDENT" REF="80" KIND="PRPR">It</COREF> is more than
six months late filing <COREF ID="84" TYPE="IDENT" REF="83"
KIND="PSDT">its</COREF> accounts for the year ending January 1995.
```

```
Accounts for <COREF ID="85" TYPE="IDENT" REF="84" KIND="PSDT">its</COREF>
<COREF ID="105" TYPE="IDENT" REF="104" KIND="FLNM">Sock Shop</COREF>
subsidiary are heavily qualified by the auditors.</p>

<p> The <COREF ID="86" TYPE="IDENT" REF="87" KIND="ACNM">DTI</COREF>
investigation into <COREF ID="117" TYPE="IDENT" REF="116" KIND="SGNN">
<COREF ID="88" TYPE="IDENT" REF="24" KIND="ABNM" MIN="Hinchliffe">Mr
Hinchliffe</COREF>'s affairs</COREF> are understood to focus on the 1993
collapse of <COREF ID="91" KIND="FLNM" MIN="Boxgrey">Boxgrey, a company
sold by <COREF ID="89" TYPE="IDENT" REF="88" KIND="DANN"
MIN="entrepreneur">the Sheffield-based entrepreneur</COREF> shortly
before <COREF ID="90" TYPE="IDENT" REF="91" KIND="PRPR">it</COREF>
collapsed</COREF>.</p>

<p> <COREF ID="92" TYPE="IDENT" REF="74" KIND="ABNM" MIN="Grosvenor">Mr
Grosvenor</COREF> is known in <COREF ID="114" TYPE="IDENT" REF="113"
KIND="DANN">the City</COREF> as an entrepreneur who has also worked as a
financial public relations adviser.  <COREF ID="93" TYPE="IDENT" REF="92"
KIND="PSDT">His</COREF> name regularly appears in newspaper social pages
because of <COREF ID="94" TYPE="IDENT" REF="93" KIND="PSDT">his</COREF>
family connections.  <COREF ID="95" TYPE="IDENT" REF="94" KIND="PRPR">
He</COREF> is related to the Aga Khan as well as <COREF ID="96"
TYPE="IDENT" REF="97" KIND="FLNM" MIN="Duke of Westminster">the Duke of
Westminster</COREF>.  <COREF ID="98" TYPE="IDENT" REF="95" KIND="PRPR">
He</COREF> was married in 1966 to <COREF ID="100" KIND="FLNM"
MIN="daughter|Ellen Seeliger">Ellen Seeliger, <COREF ID="99" TYPE="IDENT"
REF="100" KIND="APNN" MIN="daughter">daughter of Germany's Ambassador to
Mexico</COREF></COREF>.  <COREF ID="101" TYPE="IDENT" REF="98"
KIND="PSDT">His</COREF> mother was one of the four daughters of the third
Lord Churston.</p>
</TXT>
</DOC>
```

C.2.1.3 Answer key summary for the *Guardian* article

```
-------------------------------------------------------------------------

3       Facia
7       the troubled Facia retail company
13      Britain's second largest privately owned retail chain
20      Facia, which had been seeking capital
21      Facia, headed by  Stephen Hinchliffe
25      Facia
34      Facia
6       Facia
37      the company
42      the company
50      Facia
```

70 Facia
79 Facia
80 The company
83 It
84 its
85 its
2 firm

---

1 Facia saviour
0 Grosvenor
4 William Grosvenor, the entrepreneur and well connected cousin of the Duke of Westminster, who is heading attempts by Texas American Group to take over the troubled Facia retail company
6 well connected cousin of the Duke of Westminster
5 the entrepreneur
8 he
9 Mr Grosvenor, aged 54, a pageboy at the Queen's 1953 Coronation, who is acting as chief executive officer of the US -listed company trying to buy Britain 's second largest privately owned retail chain
14 he
15 He
43 Mr Grosvenor
44 his
45 his
46 He
47 he
48 his
51 his
58 Mr Grosvenor
60 Mr Grosvenor
65 Mr Grosvenor
66 he
69 He
74 Mr Grosvenor
92 Mr Grosvenor
93 His
94 his
95 He
98 He
101 His
68 its chief executive
10 chief executive officer of the US -listed company trying to buy Britain's second largest privately owned retail chain

---

112 City

```
113     City
114     the City

-------------------------------------------------------------------------

106     Patrick Donovan, City editor
107     City editor

-------------------------------------------------------------------------

97      the Duke of Westminster
96      the Duke of Westminster

-------------------------------------------------------------------------

12      Texas American Group
11      the US-listed company trying to buy Britain's second largest
        privately owned retail chain
18      Texas American
19      it
41      Texas American
67      the Texas American Group
119     We
73      the company
75      the company
76      Texas American
77      the company
78      It
115     its
49      his US partners

-------------------------------------------------------------------------

110     US
111     US

-------------------------------------------------------------------------

108     Britain
109     Britain

-------------------------------------------------------------------------

17      a spent conviction for tax fraud in Britain, for which he
        received a 12 month suspended sentence and a #1,000 fine in 1980
16      The conviction

-------------------------------------------------------------------------

103     #8,500 costs of a grouse shoot
```

```
102     #8,500

-------------------------------------------------------------------------

23      Stephen Hinchliffe
22      Mr Hinchliffe, whose affairs have been investigated by the
        Department of Trade and Industry
24      he
88      Mr Hinchliffe
89      the Sheffield-based entrepreneur

-------------------------------------------------------------------------

104     Sock Shop
105     Sock Shop

-------------------------------------------------------------------------

116     affairs
117     Mr Hinchliffe's affairs

-------------------------------------------------------------------------

87      the Department of Trade and Industry
86      DTI

-------------------------------------------------------------------------

31      shoe
30      shoe

-------------------------------------------------------------------------

27      Sears
26      Sears
32      Sears
33      It
35      Sears
40      Sears

-------------------------------------------------------------------------

29      shoe shops including Freeman, Hardy & Willis and Curtess, which
        he took over from Sears
28      the shoe operations

-------------------------------------------------------------------------

57      the High Court
56      the High Court
```

```
---------------------------------------------------------------------------

39      #4 million worth of rental payments which became due yesterday
38      #4 million

---------------------------------------------------------------------------

53      his  bankruptcy
118     bankruptcy
52      his  bankruptcy

---------------------------------------------------------------------------

55      Helen Clark, a lawyer at the  City  firm of Eversheds
54      she
59      she

---------------------------------------------------------------------------

62      costs which have been estimated at  more than #100,000
61      more than #100,000

---------------------------------------------------------------------------

64      judgment against  Mr Grosvenor and two other defendants for
        #350,000 plus  costs which have been estimated at more than
        #100,000
63      The judgment

---------------------------------------------------------------------------

72      an agreement to make a contract (to take over  Facia  )
71      The deal

---------------------------------------------------------------------------

82      a capital injection of  around #40 million
81      around #40 million

---------------------------------------------------------------------------

91      Boxgrey, a company sold by the Sheffield-based entrepreneur
        shortly before  it  collapsed
90      it

---------------------------------------------------------------------------

100     Ellen Seeliger,  daughter of Germany's Ambassador to Mexico
99      daughter of Germany's Ambassador to Mexico
```

C.2.1.4 LOLITA's current co-reference task output for the *Guardian* article

```
<DOC>
<DOCID> guardian.june.1.1996 </DOCID>
<DOCNO> 01061996. </DOCNO>
<HL> <COREF ID="21000026" TYPE="IDENT" REF="21000025">
<COREF ID="8000018" TYPE="IDENT" REF="8000017">Facia</COREF>
saviour</COREF> is bankrupt. <COREF ID="21000027" TYPE="IDENT"
REF="21000026">Grosvenor</COREF> says <COREF ID="8000019" TYPE="IDENT"
REF="8000018">firm</COREF> was not told. </HL>
<DD> 01/06/96 </DD>
<AUTHOR><COREF ID="400001">Patrick Donovan</COREF>,
<COREF ID="400002" TYPE="IDENT" REF="400001">
<COREF ID="100001">City</COREF> editor</COREF></AUTHOR>

<TXT>
<p> <COREF ID="2100002" TYPE="IDENT" REF="2100001">William Grosvenor
</COREF>, <COREF ID="21000028" TYPE="IDENT" REF="21000027">
the entrepreneur</COREF> and <COREF ID="21000029" TYPE="IDENT"
REF="21000028">well connected cousin of the <COREF ID="500001">
Duke of Westminster</COREF></COREF>, who is heading attempts by
<COREF ID="600001">Texas American Group</COREF> to take over the
troubled <COREF ID="800001">Facia</COREF> <COREF ID="700001">retail
</COREF> company, last night admitted that <COREF ID="2100003"
TYPE="IDENT" REF="2100002">he</COREF> is bankrupt. </p>

<p> Mr <COREF ID="2100004" TYPE="IDENT" REF="2100003">Grosvenor
</COREF>, aged 54, a pageboy at the Queen's 1953 Coronation, who
is acting as chief executive officer of the US-listed company trying
to buy <COREF ID="900001">Britain</COREF>'s second largest privately
owned retail chain, has a spent conviction for tax fraud in
<COREF ID="900002" TYPE="IDENT" REF="900001">Britain</COREF>,
for which <COREF ID="2100005" TYPE="IDENT" REF="2100004">he</COREF>
received a 12 month suspended sentence and a #1,000 fine in 1980.
</p>

<p> <COREF ID="2100006" TYPE="IDENT" REF="2100005">He</COREF>
pleaded guilty to plotting to defraud the Inland Revenue by
attempting to pass off the #8,500 costs of a grouse shoot as a tax
deductible <COREF ID="700002" TYPE="IDENT" REF="700001">business</COREF>
expense. The conviction is considered spent under the Rehabilitation of
Offenders Act.</p>

<p> <COREF ID="600002" TYPE="IDENT" REF="600001">Texas American
</COREF> announced on Thursday that <COREF ID="600003" TYPE="IDENT"
REF="600002">it</COREF> had agreed to buy <COREF ID="800002"
TYPE="IDENT" REF="800001">Facia</COREF>, which had been seeking
capital. <COREF ID="800003" TYPE="IDENT" REF="800002">Facia</COREF>,
```

```
headed by <COREF ID="1100001">Stephen Hinchliffe</COREF>, operates
850 speciality shops, with the high-street names including
<COREF ID="1300001">Sock Shop</COREF>, Salisburys, Red or Dead,
and Contessa.</p>

<p> Mr <COREF ID="1100002" TYPE="IDENT" REF="1100001">Hinchliffe
</COREF>, whose affairs have been investigated by the <COREF ID=
"1200001">Department of Trade and Industry</COREF>, also controls
<COREF ID="1400001">shoe</COREF>shops including Freeman, Hardy & Willis
and Curtess, which <COREF ID="1100003" TYPE="IDENT" REF="1100002">
he</COREF> took over from <COREF ID="1500001">Sears</COREF>. </p>

<p> As the crisis surrounding <COREF ID="800004" TYPE="IDENT"
REF="800003">Facia</COREF> escalated, <COREF ID="1500002"
TYPE="IDENT" REF="1500001">Sears</COREF> yesterday served a petition in
the <COREF ID="1000001">High Court</COREF> to put the <COREF ID="1400002"
TYPE="IDENT" REF="1400001">shoe</COREF> operations into administration.
<COREF ID="1500003" TYPE="IDENT" REF="1500002">Sears</COREF> still
has an interest in the companies as part of a staggered settlement deal.
<COREF ID="1500004" TYPE="IDENT" REF="1500003">It</COREF> is putting
#25 million aside to cover the disposal costs and is suspending
<COREF ID="1600001">plans to sell a further interest in the Saxone chain
to <COREF ID="800005" TYPE="IDENT" REF="800004">Facia</COREF></COREF>.
</p>

<p> It is understood <COREF ID="1500005" TYPE="IDENT" REF="1500004">
Sears</COREF> has been increasingly concerned at the running of
<COREF ID="800006" TYPE="IDENT" REF="800005">Facia</COREF> and took
the action after <COREF ID="800007" TYPE="IDENT" REF="800006">the company
</COREF> defaulted on #4 million worth of rental payments which became
due yesterday.  <COREF ID="1500006" TYPE="IDENT" REF="1500005">Sears
</COREF> is believed to have little confidence in <COREF ID="1600002"
TYPE="IDENT" REF="1600001"><COREF ID="600004" TYPE="IDENT" REF="600003">
Texas American</COREF>'s plans</COREF> for <COREF ID="600005"
TYPE="IDENT" REF="600004">the company</COREF>.</p>

<p> Mr <COREF ID="2100007" TYPE="IDENT" REF="2100006">Grosvenor</COREF>
yesterday accepted that <COREF ID="1700001"><COREF ID="2100008"
TYPE="IDENT" REF="2100007">his</COREF> bankruptcy</COREF> could affect
<COREF ID="2100009" TYPE="IDENT" REF="2100008">his</COREF> credibility
as a businessman.  <COREF ID="21000013" TYPE="IDENT" REF="21000012">He
</COREF> added that <COREF ID="21000010" TYPE="IDENT" REF="2100009">he
</COREF> had told <COREF ID="21000011" TYPE="IDENT" REF="21000010">his
</COREF> <COREF ID="200001">US</COREF> partners but had not informed
<COREF ID="800008" TYPE="IDENT" REF="800007">Facia</COREF> about
<COREF ID="1700002" TYPE="IDENT" REF="1700001"><COREF ID="21000012"
TYPE="IDENT" REF="21000011">his</COREF> bankruptcy</COREF>.</p>

<p> <COREF ID="1800001">Helen Clark</COREF>, a lawyer at <COREF
ID="800009" TYPE="IDENT" REF="800008">the <COREF ID="100002"
```

```
TYPE="IDENT" REF="100001">City</COREF> firm of Eversheds</COREF>,
said that <COREF ID="1800002" TYPE="IDENT" REF="1800001">she</COREF>
had made a <COREF ID="1700003" TYPE="IDENT" REF="1700002">bankruptcy
</COREF> order in the <COREF ID="1000002" TYPE="IDENT" REF="1000001">
High Court</COREF> on 19 October 1994.  "As yet there has been no
distribution to creditors and Mr <COREF ID="21000014" TYPE="IDENT"
REF="21000013">Grosvenor</COREF> remains bankrupt", <COREF ID="1800003"
TYPE="IDENT" REF="1800002">she</COREF> said.</p>

<p> <COREF ID="2000001">Later that year</COREF>, on November 24,
a Dublin solicitor, Denis Murnaghan, obtained <COREF ID="1900001">
judgment</COREF> against Mr <COREF ID="21000015" TYPE="IDENT"
REF="21000014">Grosvenor</COREF> and two other defendants for
#350,000 plus costs which have been estimated at more than #100,000.
<COREF ID="1900002" TYPE="IDENT" REF="1900001">The judgment</COREF>
remains unpaid.</p>

<p> Mr <COREF ID="21000016" TYPE="IDENT" REF="21000015">Grosvenor
</COREF> yesterday said <COREF ID="21000017" TYPE="IDENT" REF="21000016">
he</COREF> was not a director of the <COREF ID="600006" TYPE="IDENT"
REF="600005"> Texas American Group</COREF> but was acting as <COREF
ID="2100001"><COREF ID="600007" TYPE="IDENT" REF="600006">its</COREF>
chief executive</COREF>.  <COREF ID="21000018" TYPE="IDENT" REF=
"21000017">He</COREF> added: "<COREF ID="600008" TYPE="IDENT" REF=
"600007">We</COREF> have come to an agreement to make a contract
(to take over <COREF ID="8000010" TYPE="IDENT" REF="800009">Facia
</COREF>)". The deal would be funded out of the <COREF ID="8000011"
TYPE="IDENT" REF="8000010">company</COREF>'s own resources.</p>

<p> Mr <COREF ID="21000019" TYPE="IDENT" REF="21000018">Grosvenor
</COREF> said that <COREF ID="8000012" TYPE="IDENT" REF="8000011">
the company</COREF> had not submitted up to date filings
with the <COREF ID="200002" TYPE="IDENT" REF="200001">US</COREF>
<COREF ID="200003" TYPE="IDENT" REF="200002">Securities and Exchange
Commission</COREF> because of the need to take
into account recent acquisitions.</p>

<p> According to other documents about <COREF ID="600009"
TYPE="IDENT" REF="600008">Texas American</COREF> filed in Washington,
<COREF ID="6000010" TYPE="IDENT" REF="600009">the company</COREF>
is said to have interests in Internet lottery and casino games.
<COREF ID="6000011" TYPE="IDENT" REF="6000010">It</COREF> has stakes
in a Nevada hotel development and Portugese holiday businesses.</p>

<p> <COREF ID="8000013" TYPE="IDENT" REF="8000012">Facia</COREF>
last night declined to make any comment.  <COREF ID="8000014"
TYPE="IDENT" REF="8000013">The company</COREF> has been looking
for a capital injection of around #40 million.  <COREF ID="8000015"
TYPE="IDENT" REF="8000014">It</COREF> is more than six months late
filing <COREF ID="8000016" TYPE="IDENT" REF="8000015">its</COREF>
```

```
accounts for <COREF ID="2000002" TYPE="IDENT" REF="2000001">
the year</COREF> ending January 1995. Accounts for <COREF ID=
"8000017" TYPE="IDENT" REF="8000016">its</COREF> <COREF ID="1300002"
TYPE="IDENT" REF="1300001">Sock Shop</COREF> subsidiary are heavily
qualified by the auditors.</p>

<p> The <COREF ID="1200002" TYPE="IDENT" REF="1200001">DTI</COREF>
investigation into Mr <COREF ID="1100004" TYPE="IDENT" REF="1100003">
Hinchliffe</COREF>'s affairs are understood to focus on the 1993 collapse
of Boxgrey, a company sold by <COREF ID="1100005" TYPE="IDENT"
REF="1100004">the <COREF ID="300001">Sheffield</COREF>-based entrepreneur
</COREF> shortly before <COREF ID="300002" TYPE="IDENT" REF="300001">
it</COREF> collapsed.</p>

<p> Mr <COREF ID="21000020" TYPE="IDENT" REF="21000019">Grosvenor
</COREF> is known in the <COREF ID="100003" TYPE="IDENT" REF="100002">
City</COREF> as an entrepreneur who has also worked as a financial
public relations adviser.  <COREF ID="21000021" TYPE="IDENT" REF=
"21000020">His</COREF> name regularly appears in newspaper social
pages because of <COREF ID="21000022" TYPE="IDENT" REF="21000021">
his</COREF> family connections.  <COREF ID="21000023" TYPE="IDENT"
REF="21000022">He</COREF> is related to the Aga Khan as well
as the <COREF ID="500002" TYPE="IDENT" REF="500001">Duke of
Westminster</COREF>.  <COREF ID="21000024" TYPE="IDENT" REF=
"21000023">He</COREF> was married in 1966 to <COREF ID="2200001">
Ellen Seeliger</COREF>, <COREF ID="2200002" TYPE="IDENT"
REF="2200001">daughter of Germany's Ambassador to Mexico</COREF>.
<COREF ID="21000025" TYPE="IDENT" REF="21000024">His</COREF> mother was
one of the four daughters of the third Lord Churston.</p>
</TXT>
</DOC>
```

References

Allen, J. (1994). *Natural Language Understanding (2nd edition).* The Benjamin Cummings Publishing Company Inc, Redwood City, CA.

Alshawi, H. (1987). *Memory and Context for Language Interpretation.* CUP, Cambridge.

Anderson, A., Garrod, S., & Sanford, A. J. (1983). *The accessibility of pronominal antecedents as a function of episode shifts in narrative text. Quarterly Journal of Experimental Psychology* **35a**, 427–440.

Appelt, D. E., Hobbs, J. R., Bear, J., Israel, D., Kameyama, M., A., K., Martin, D., K., M., & Tyson, M. (1995). SRI International FASTUS system. MUC-6 Test Results and Analysis. In *Proceedings of the Sixth Message Understanding Conference*, pages 237–248. Morgan Kaufmann Publishers.

ARPA (1993). *Proceedings of the Fifth Message Understanding Conference (MUC-5).* Morgan Kaufmann Publishers.

Baldwin, B. (1995). *CogNIAC: A discourse Processing Engine.* PhD thesis, Department of Computer and Information Sciences University of Pennsylvania.

Baldwin, B. (1997). CogNIAC: high precision coreference with limited knowledge and linguistic resources. In *Proceedings of the ACL'97/EACL'97 workshop on operational factors in practical, robust, anaphora resolution for unrestricted texts*, Madrid. Association for Computational Linguistics, ACL.

Baldwin, B., Reynar, J., Collins, M., Eisner, J., Ratnaprkhi, A., Rosenzweig, J., & Srinivas, A. S. (1995). Description of the University of Pennsylvania System Used for MUC-6. In *Proceedings of the Sixth Message Understanding Conference*, pages 177–191. Morgan Kaufmann Publishers.

Baring-Gould, S. (forthcoming). *The knowledge representation of LOLITA.* PhD thesis, Laboratory for Natural Language Engineering, Department of Computer Science, University of Durham.

Boguraev, B. (1979). *Automatic resolution of linguistic ambiguities.* PhD thesis, University of Cambridge, Computer Laboratory.

Brennan, S. E., Friedman, M. W., & Pollard, C. J. (1987). A centering approach to pronouns. In *Proceedings of the 25st Annual Meeting of the Association of Computational Linguistics,*, pages 155–162, Cambridge, MA.

Brill, E. (1994). Some advances in rule-based part of speech tagging. In *AAAI-94.*

Brown, G. & Yule, G. (1983). *Discourse Analysis.* Cambridge Textbooks in Linguistics. Cambridge University Press.

Callaghan, P. (1998). *An Evaluation of LOLITA and Related Natural Language Processing Systems.* PhD thesis, Laboratory for Natural Language Engineering, Department of Computer Science, University of Durham.

Caramazza, A. & Gupta, S. (1979). *The roles of topicalisation, parallel function and verb semantics in the interpretation of pronouns. Linguistics* **17**, 497–518.

Carter, D. (1987). *Interpreting Anaphors in natural language texts.* Ellis Horwood Series in Artificial Intelligence. Ellis Horwood Limited, Chichester.

Charniak, E. (1972). Toward a model of children's story comprehension. Technical Report AI TR-266, MIT, Artificial Intelligence Laboratory.

Chomsky, N. (1965). *Aspects of the Theory of Syntax.* MIT Press, Cambridge, MA.

Chomsky, N. (1981). *Lectures on Government and Binding.* Foris, Dordrecht.

Chomsky, N. (1986). *Knowledge of Language: Its Nature, Origin and Use.* Praeger, New York.

Christopherson, P. (1939). *The Articles: A Study of Their Theory and Use in English.* Oxford University Press.

Clarke, H. H. & Sengul, C. J. (1979). *In search of referents for nouns and pronouns. Memory and Cognition* **7**, 35–41.

Collingham, R. J., Johnson, K., Nettleton, D. J., Dempster, G., & Garigliano, R. (1997). *The Durham Telephone Enquiry System. International Journal of Speech Technology* **2**, 113–119.

Costantino, M. (1997). *User Defined Templates in Information Extraction.* PhD thesis, (submitted) Laboratory for Natural Language Engineering, Department of Computer Science, University of Durham.

DARPA (1995). *Proceedings of the Sixth Message Understanding Conference (MUC-6).* Morgan Kaufmann Publishers.

Dyer, M. (1983). *In-Depth Understanding: A Computer Model of Integrated Processing for Narrative Comprehension.* MIT Press, Cambridge, MA.

Fiengo, R. & May, R. (1994). *Indices and Identity.* Linguistic Inquiry Monographs. MIT Press, Cambridge, MA.

Fisher, D., Soderland, S., McCarthy, J., Feng, F., & Lehner, W. (1995). Description of the UMass system as used for MUC-6. In *Proceedings of the Sixth Message Understanding Conference.* Morgan Kaufmann Publishers.

Fraurud, K. (1990). *Definiteness and the processing of NPs in natural discourse. Journal of Semantics* **7**, 395–434.

Gaizauskas, R., Wakao, T., Humphreys, K., Cunnigham, H., & Wilks, Y. (1995). University of Sheffield: Description of the LaSIE System as Used for MUC-6. In *Proceedings of the Sixth Message Understanding Conference*, pages 207–220. Morgan Kaufmann Publishers.

Galliers, J. & Sparck Jones, K. (1993). Evaluating natural language processing systems. Technical Report 291, University of Cambridge Computer Laboratory.

Garigliano, R. (1995). *Editorial. Natural Language Engineering* **1**.

Garigliano, R., Morgan, R. G., & Smith, M. H. (1992). Lolita: progress report i. Technical report, Laboratory for Natural Language Engineering, Department of Computer Science, University of Durham.

Garigliano, R., Morgan, R. G., & Smith, M. H. (1993). The LOLITA System as a Contents Scanning Tool. In *Avignon '93.*

Garigliano, R., Urbanowicz, A. J., & Nettleton, D. J. (1998). University of Durham: Description of the LOLITA System as used in MUC-7. In *Seventh Messages Understanding Conference (MUC-7)*. Morgan Kaufmann.

Garnham, A., Oakhill, J., & Cruttenden, H. (1992). *The Role of Implicit Causality and Gender Cue in the Interpretation of Pronouns. Language and Cognitive Processes* **7**, 231–255.

Gernsbacher, M. A. & Hargreaves, D. (1988). *Accessing sentence participants: The advantage of first mention. Journal of Memory and Language* **27**, 699–717.

Gernsbacher, M. A., Hargreaves, D., & Beeman, M. (1989). *Building and accessing clausal representations: The advantage of first mention versus the advantage of clause recency. Journal of Memory and Language* **28**, 735–755.

Gordon, P. C., Grosz, B. J., & Gilliom, L. A. (1993). *Pronouns, Names, and the Centering of Attention in Discourse. Cognitive Science* **17**, 311–347.

Grodzinsky, Y. & Reinhart, T. (1993). *The Innateness of Binding and Coreference. Linguistic Inquiry* **24**, 69–101.

Grosz, B. J. (1977). The representation and the use of focus in a system for understanding dialogs. In *Proceedings of IJCAI*, pages 67–76, Cambridge, MA.

Grosz, B. J., Joshi, A. K., & Weinstein, S. (1986/1995). *Towards a computational theory of discourse interpretation. Computational Lingusitics* **21**.

Grosz, B. J. & Sidner, C. L. (1986). *Attention, Intentions and the Structure of Discourse. Computational Linguistics* **12**, 175–204.

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

Hirst, G. (1981). *Anaphora in Natural Language Understanding.* Springer-Verlag, Berlin.

Hirst, G. (1987). *Semantic Interpretation and the Resolution of Ambiguity.* CUP, Cambridge.

Hobbs, J. R. (1979). *Resolving pronoun references. Lingua* **44**, 311–338.

Hudak, P., Peyton Jones, S., & Wadler, P. (1992). Report on the functional programming language haskell. Technical Report Version 1.2,

Department of Computer Science, University of Glasgow, also available at ftp://ftp.dcs.gla.ac.uk/pub/haskell/report.

Huls, C., Bos, E., & Claassen, W. (1995). *Automatic Referent Resolution of Deictic and Anaphoric Expressions. Computational Linguistics* **21**, 59–79.

Jackendoff, R. (1972). *Semantic Interpretation in Generative Grammar.* MIT Press, Cambridge, MA.

Jacobs, P. S. & Rau, L. F. (1993). *Innovations in text interpretation. Artificial Intelligence* **63 (Special Issue on NLP)**, 143–191.

Johnson, D. (1977). *On relational constraints on grammars.* In **Cole, P. & J., S.**, editors, *Syntax and Semantics*, volume 8, pages 151–178. Academic Press.

Jones, C. (1994). *Dialogue Structure Models: An Engineering Approach to the Analysis and Generation of Natural English Dialogues.* PhD thesis, Laboratory for Natural Language Engineering, Department of Computer Science, University of Durham.

Kamp, H. (1981). *A Theory of Truth and Semantic Representation.* In **Groenendijk, J. A., Janssen, T. M. V., & Stokhof, M. B. J.**, editors, *Formal Methods in the Study of Language*, volume 136, pages 227–322. Mathematical Centre Tracts, Amsterdam.

Katz, J. (1971). *Generative semantics is interpretive semantics. Linguistic Inquiry* **2**, 313–31.

Katz, J. & Fodor, J. A. (1963). *The structure of a semantic theory. Language* **39**, 170–210.

Keenan, E. & Comrie, B. (1977). *Noun phrase accessibility and universal grammar. Linguistic Inquiry* **8**, 62–100.

Kehler, A. (1998). *Current Theories of Centering for Pronoun Interpretation: A Critical Evaluation. Computational Lingusitics* **23**, 467–475.

Kempson, R. M. (1990). Lecture course on "advanced syntax", university college london.

Kennedy, C. & Boguraev, B. (1996). Anaphora in a Wider Context: Tracking

Discourse Referents. In **Wahlster, W.**, editor, *Proceedings of the 12th European Conference on Artificial Intelligence*, pages 582–586. John Wiley & Sons, Ltd.

Lappin, S. & Leass, H. J. (1994). *An algorithm for Prononminal Anaphora Resolution. Computational Linguistics* **20**, 535–561.

Lasnik, H. (1989). *Essays on Anaphora.* Studies in natural language and linguistic theory. Kluwer, Dordrecht.

Lin, D. (1995). University of Manitoba: Description of the PIE System Used for MUC-6. In *Proceedings of the Sixth Message Understanding Conference*, pages 113–126. Morgan Kaufmann Publishers.

Long, D. & Garigliano, R. (1994). *Reasoning by Analogy and Causality: A Model And Application.* Ellis Horwood Series in Artificial Intelligence. Ellis Horwood.

Lyons, J. (1977). *Semantics.* Cambridge University Press.

Marslen-Wilson, W. D., Tyler, L. K., & Koster, C. (1993). *Integrative processes in utterance resolution. Journal of Memory and Language* **32**, 56–75.

McCarthy, J. F. (1996). *A Trainable Approach to Coreference Resolution for Information Extraction.* PhD thesis, Department of Computer Science, University of Massachisetts Amherst.

McCawley, J. (1968). *The Role of Semantics in a Grammar.* In **E., B. & Harms, R. T.**, editors, *Universals in Linguistic Theory*, pages 124–69. Holt, Rinehart and Winston, New York.

McKoon, G., Ward, G., Ratcliff, R., & Sproat, R. (1993). *Morphosyntactic and pragmatic factors affecting the accessibility of discourse entities. Journal of Memory and Language* **32**, 56–75.

Miller, G. (1990). *WordNet: An online lexical database. International Journal of Lexicography* **3**.

Morgan, R. G., Garigliano, R., Callaghan, P., Poria, S., Smith, M. H., Urbanowicz, A. J., Collingham, R. J., Costantino, M., Cooper, C., & the LOLITA Group (1995). University of Durham: Description of the

LOLITA System as used in MUC-6. In *Sixth Messages Understanding Conference (MUC-6)*. Morgan Kaufmann.

Morgan, R. G., Smith, M. H., & Short, S. (1994). Translation by Meaning and Style in LOLITA. In *Proceedings of Machine Translation: Ten years on.* Cranfield University and British Computer Society.

Neale, S. (1990). *Descriptions.* MIT Press, Cambridge, MA.

Nettleton, D. J. (1995). *Evolutionary Algorithms in Artificial Intelligence: A Comparative Study Through Applications.* PhD thesis, Laboratory for Natural Language Engineering, Department of Computer Science, University of Durham.

Poesio, M. & Vieira, R. (1997). A corpus-based investigation of definite description use. Available from http://xxx.lanl.gov/cmp-lg/9710007. To appear in Computational Linguistics.

Poria, S. (forthcoming). *An Engineering Approach to Knowledge Acquisition by the Interactive Analysis of Dictionary Definitions.* PhD thesis, Laboratory for Natural Language Engineering, Department of Computer Science, University of Durham.

Procter, P., editor (1995). *Cambridge International Dictionary of English.* Cambridge University Press.

Reinhart, T. (1976). *The syntactic domain of anaphora.* PhD thesis, MIT.

Reinhart, T. (1986). *Centre and periphery in the grammar of anaphora.* In **Lust, B.**, editor, *Studies in the acquisition of anaphora*, volume 1, pages 123–150. Reidel, Dordrecht.

Sanford, A. J. & Garrod, S. C. (1981). *Understanding Written Language.* Wiley, Chichester.

Sanford, A. J., Moar, K., & Garrod, S. C. (1988). *Proper names as controllers of discourse focus. Language and Speech* .

Sidner, C. L. (1979). Towards a computational theory of definite anaphora comprehension in english discourse. Technical Report TR 537, MIT Artificial Intel-

ligence Laboratory.

Sidner, C. L. (1981). *Focusing for Interpretation of Pronouns. American Journal of Computational Linguistics* **7**, 217–231.

Sidner, C. L. (1983). *Focusing in the comprehension of definite anaphora.* In **Brady & Berwick**, editors, *Computational Models of Discourse*, pages 267–330. MIT Press, Cambridge, MA.

Smith, M. H. (1995). *Natural Language Generation in the LOLITA System: An Engineering Approach.* PhD thesis, University of Durham.

Smith, N. V. & Wilson, D. (1979). *Modern Linguistics. The Results of Chomsky's Revolution.* Penguin, Harmondsworth.

Sowa, J. F. (1984). *Conceptual Structures, information processing in mind and machine.* Addison-Wesley.

Stevenson, R., Nelson, A. W. R., & Stenning, K. (1995). *The role of parallelism in strategies of pronoun comprehension. Language and Speech* **38**, 393–418.

Stevenson, R. & Urbanowicz, A. J. (1995). The effects of sentence subject, initial mention and pragmatic plausibility on the accessibility of a pronoun's antecedents. Technical Report HCRC/RP-58, Human Communication Research Centre, Edinburgh.

Stevenson, R. J., Crawley, R. A., & Kleinman, D. (1994). *Thematic roles, focus and the representation of events. Language and Cognitive Processes* **9**, 519–548.

Sundheim, B. M. (1995). Overview of Results of the MUC-6 Evaluation. In *Proceedings of the Sixth Message Understanding Conference (MUC-6).* DARPA, Morgan Kaufmann Publishers.

Tomita, M. (1986). *Efficient Parsing of NL: A Fast Algorithm for Practical Systems.* Kluwer Academic Publishers, Boston, Ma.

Van Rijsbergen, C. J. (1979). *Information Retrieval.* Butterworths, London.

Villain, M., Burger, J., Aberdeen, J., Connolly, D., & Hirschman, L.

(1995). A Model-Theoretic Coreference Scoring Scheme. In *Proceedings of the Sixth Message Understanding Conference*, pages 45–52. Morgan Kaufmann Publishers.

Walker, M. A., Iida, M., & Cote, S. (1994). *Japanese discourse and the process of centering. Computational Lingusitics* **20**.

Wang, Y. (1994). *An Intelligent Computer-based Tutoring Approach for the Management of Negative Transfer.* PhD thesis, University of Durham.

Webber, B. L. (1979). *A Formal Approach to Discourse Anaphora.* Outstanding Dissertations in Linguistics. Garland Publishing, Inc, New York and London.

Webber, B. L. (1983). *So what can we talk about now.* In **Brady, M. & Berwick, B.**, editors, *Computational Models of Discourse*, pages 331–370. MIT Press, Cambridge, MA.

Wilks, Y. (1975). *Preference semantics.* In **Keenan, E. L. K.**, editor, *Formal Semantics of Natural Language*, pages 329–348. CUP, Cambridge.