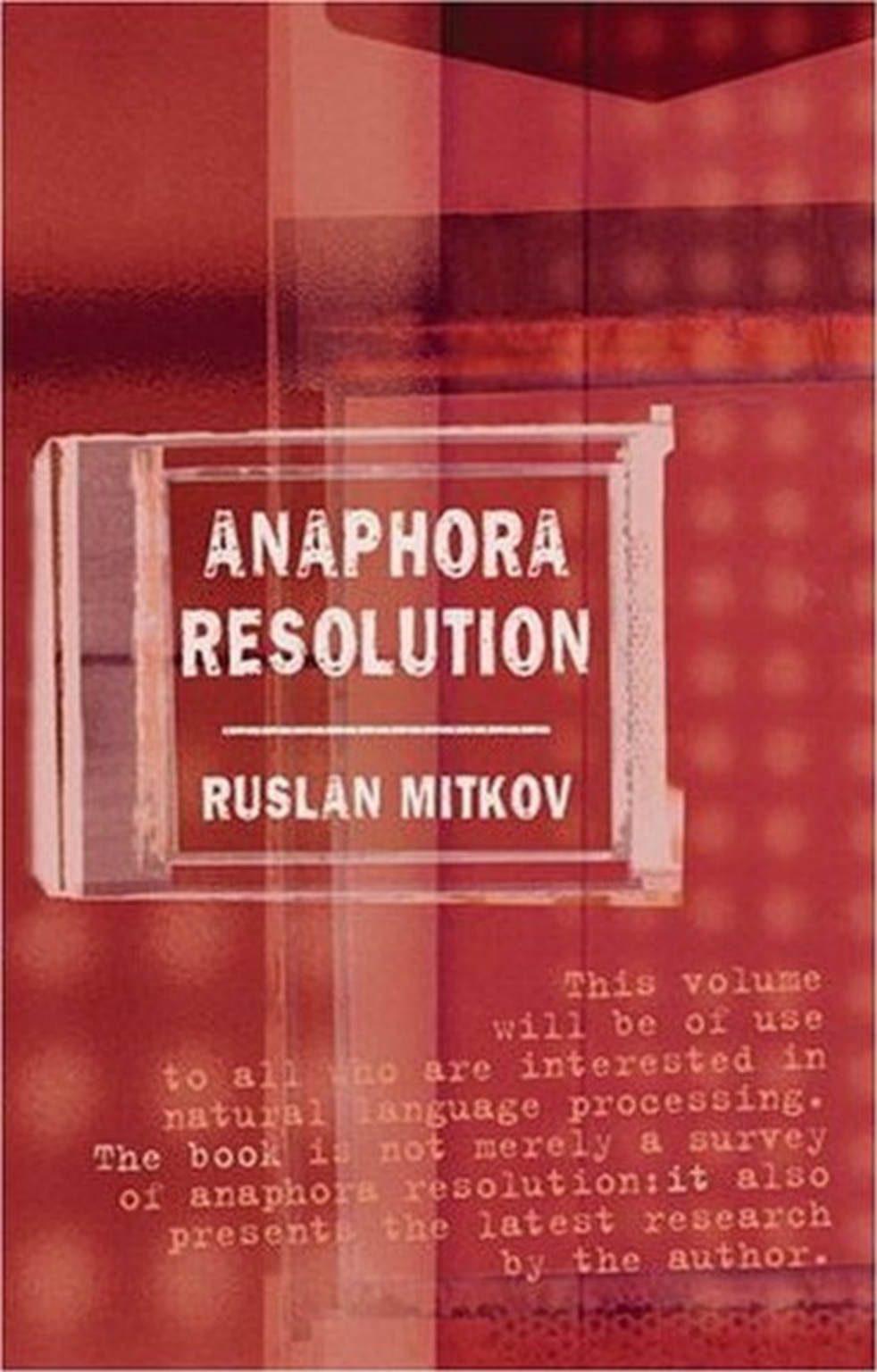
ANAPHORA RESOLUTION
RUSLAN MITKOV
This volume will be of use to all who are interested in natural language processing. The book is not merely a survey of anaphora resolution: it also presents the latest research by the author.

Anaphora Resolution

Anaphora Resolution

RUSLAN MITKOV

An imprint of Pearson Education

London • New York • Toronto • Sydney • Tokyo • Singapore • Hong Kong • Cape Town
New Delhi • Madrid • Paris • Amsterdam • Munich • Milan • Stockholm

PEARSON EDUCATION LIMITED

Head Office:
Edinburgh Gate
Harlow CM20 2JE
Tel: +44 (0)1279 623623
Fax: +44 (0)1279 431059

London Office:
128 Long Acre
London WC2E 9AN
Tel: +44 (0)20 7447 2000
Fax: +44 (0)20 7240 5771
Website: www.history-minds.com

First published in Great Britain in 2002

ISBN 0 582 32505 6

British Library Cataloguing in Publication Data
A CIP catalogue record for this book can be obtained from the British Library

Library of Congress Cataloguing in Publication Data
A CIP Catalogue record for this book can be obtained from the Library of Congress

10 9 8 7 6 5 4 3 2 1

Set in Palatino 9.5/12pt by Graphicraft Limited, Hong Kong
Produced by Pearson Education (China) Ltd.

The Publishers' policy is to use paper manufactured from sustainable forests.

Contents

To the loving memory of my mother

This book is dedicated to my mother Penka Georgieva Moldovanska who encouraged me to begin a career in research, specifically in the area of Computational Linguistics, and who sadly did not live to see in its published form the book she knew was to be dedicated to her. This book is but a pale reflection of her constant and powerful inspiration and this dedication is a modest token of appreciation for all that she did for me.

Acknowledgements

I would like to thank a number of people for their advice and help in the preparation of this book. I am particularly grateful to Prof. Geoffrey Leech, the editor of the *Studies in Language and Linguistics* series, whose comments on various drafts of the manuscript proved to be crucial and whose encouragement was greatly appreciated. I am most indebted to Linda C. Van Guilder who read the entire manuscript with great care and made very helpful suggestions for its improvement. Andrew Caink is another colleague who deserves my unreserved gratitude, as is Vince Robbins for his meticulous proof-reading of the first draft. I would like to express sincere thanks to Catalina Barbu, Richard Evans and Constantin Orasan from the Research Group in Computational Linguistics at the University of Wolverhampton not only for their helpful comments but also for implementing some of the approaches presented in this book. In addition, I would like to thank a number of colleagues who provided further comments regarding different parts or chapters of the book: Amit Bagga, Antonio Ferrández, Shalom Lappin, Shikego Nariyama, Yasuko Obana, Monique Rolbert, Maximiliano Saiz-Noeda and Hristo Tanev. Last but not least, I would like to acknowledge the support of the University of Wolverhampton which made completion of this book possible.

Preface

This book aims to present the state of the art in the expanding and increasingly important task of anaphora resolution, which plays a vital role in a number of Natural Language Processing applications including machine translation, automatic abstracting, information extraction and question answering. In surveying this material, the book aims to fill an existing gap in the literature with an up-to-date survey of the field, given that the previous books of similar nature were published some time ago. To help researchers and students involved in anaphora resolution projects, this book addresses various issues related to the practical implementation of anaphora systems, such as rules employed, algorithms implemented or evaluation techniques used.

I have not covered the work prior to 1986 in detail because this has been extensively presented in Hirst's book *Anaphora in Natural Language Understanding* (1981) as well as in Carter's book *Interpreting Anaphora in Natural Language Texts* (1987a). I have chosen instead to focus on the important work carried out after the publication of these two excellent volumes. In particular, I have discussed in detail some of the work in the 1990s, this decade being characterised by the advent of numerous new approaches and projects.

While the book intends to present an objective, comprehensive and up-to-date survey of the field, it also includes considerable discussion of my own research (more specifically Chapters 7 and 8, parts of Chapter 9, and to a lesser extent Chapters 2 and 6). At the risk of seeming somewhat less than objective, I have included this in-depth discussion of my work as something of a case study, as I know no work in greater detail than my own. I hope the readers will accept this in the manner in which it was intended: as a detailed exemplar to be used as a foil in their survey, with the understanding that this book necessarily reflects my own views on the subject. It is intended for an audience of readers interested in anaphora resolution and in Natural Language Processing or Computational Linguistics in general, including but not limited to researchers, lecturers, students and NLP software developers.

Ruslan Mitkov
October 2001

Acronyms and abbreviations

Adj	adjective
Cb	backward-looking center
CSI	common-sense inference
GB	government and binding (theory)
DRS	discourse representation structure
DRT	discourse representation theory
ENGCG	English constraint grammar
ESG	English slot grammar
FDG	functional dependency grammar
iff	if and only if
MARS	Mitkov's anaphora resolution system
MUC	Message Understanding Conference
N	noun
NLP	natural language processing
NLU	natural language understanding
NP	noun phrase
PI (rule)	pronoun interpretation (rule)
PP	prepositional phrase
PREP	preposition
S	sentence
V	verb
VP	verb phrase
VT	veins theory

Introduction

This book concerns the automatic resolution of anaphors, which is a crucial task in the understanding of natural language by computers. Before introducing concepts central to the book in Chapter 1, I shall discuss why 'understanding' natural language is so difficult and hint at how computer systems attempt this task by analysing the input at different levels. The sketch provided here should help the reader to see where anaphora resolution fits in the bigger picture of Natural Language Understanding. I shall briefly review the levels of linguistic and extra-linguistic analysis, and the various relevant forms of knowledge.

Why is it so difficult for computers to understand natural language?

Understanding natural language is a daunting task for computers. The main difficulty arises from the fact that natural languages are inherently ambiguous. Whereas humans generally manage to pick out the intended meaning from a set of possible interpretations, computers are less likely to do so due, among other reasons, to their limited 'knowledge' and inability to get their bearings in complex contextual situations.

Ambiguity can occur at the lexical level where words may have more than one meaning (e.g. *bank*, *file*, *chair*), but also at the syntactic level when more than one structural analysis is possible (e.g. *Flying planes can be dangerous*, *I saw the man with the telescope*). Furthermore, ambiguity is exhibited at the semantic level (*The rabbit is ready for lunch* – where *the rabbit* can be interpreted as both agent and patient) or pragmatic level (*Can you open the window?* – where this phrase can act both as a request and as a question, depending on the contextual situation). The automatic resolution of ambiguity requires a huge amount of linguistic and extralinguistic knowledge as well as inferring and learning capabilities, and is therefore realistic only in restricted domains.

The levels of linguistic analysis

A natural language system requires considerable knowledge about language, including how to identify words, how words are arranged into sentences, what the words mean and how their individual meanings combine to produce

sentence meanings. At a higher level, it must be able to identify sentences in a text, establish the relationships among them, glean the intentions behind each sentence, etc. In addition, if an automatic natural language system is to be able to understand language like humans, it should be supplied with world and domain knowledge as well as reasoning abilities.

A Natural Language Understanding (NLU) program should be able to determine the acceptability of a sentence from the point of view of various levels of analysis and should establish connections between the different components of a sentence or text.

In order to illustrate how natural language input is analysed and what knowledge is needed, consider a hypothetical analysis of the following sentence:

(1) This book outlines the state of the art of anaphora resolution. It discusses the complexity of this NLU task.

I assume that the computer is dealing with written text input and not voice input, so at this stage no **phonetic analysis** would be needed. To start, the **morphological** and **lexical analysis** must identify the words, their lexical classes (parts of speech) and possible derivations. Therefore *this* would be identified as a determiner, *book* as a noun and so on. In addition, *outlines* and *discusses* would be recognised as third person present tense forms of the verbs *to outline* and *to discuss*, respectively and *state of the art* would be analysed as a compound word. **Morphological** and **lexical knowledge** in the form of rules[1] and a dictionary would be needed to perform this level of analysis successfully. A domain-specific dictionary could help the program to find that the acronym *NLU* stands for *Natural Language Understanding*. Next, after identifying sentence boundaries, **syntactic analysis** would determine whether the sentences in the text are syntactically acceptable by breaking up each sentence into smaller syntactic components and applying relevant grammar rules. As a result, in the first sentence *this book*, *the state of the art*, *anaphora resolution* and *the state of the art of anaphora resolution* would be recognised as noun phrases, and *outlines the state of the art of anaphora resolution* as a verb phrase. Similar analysis would then be applied to the second sentence. **Syntactic knowledge** in the form of grammar rules would be necessary for the completion of this level of analysis. **Semantic analysis** would then look at the semantics of each word and how the words relate to one another. This analysis would tell us that the verb *to outline* requires an agent which is either human or a written work (e.g. paper, book, article) and that the patient of the verb should be a problem, area, event, etc. The semantic analysis would identify that *book* is a written work from the category of inanimate and non-human concepts and is the agent of the sentence, and that *the state of the art of anaphora resolution* is non-human and is the patient. **Semantic knowledge** is typically encoded in a dictionary or ontology and is expressed via formalisms such as first-order logic, semantic networks, attribute value pairs, knowledge representation languages, etc. In this particular example the compound word *anaphora resolution* would have to be identified as an NLU task which would normally require **domain knowledge**. It can already be seen here that distinctions between different kinds of knowledge are not always clear.

Moreover, in order to understand example (1) properly, one of the tasks of **discourse analysis** is to establish anaphoric relations. For example, the program has to find that *it* in the second sentence refers to *this book* and that *this NLU task* stands for *anaphora resolution*. Thus it becomes evident that knowledge about the semantics of the verb *to discuss* and of the (compound) words *book* and *anaphora resolution* will be very helpful at this stage. If the sentence *Will you be able to read it?* followed example (1), **pragmatic analysis** would be necessary to identify the speaker's intention by finding out if the new sentence represented a request or a question regarding the ability of the addressee to read the book (e.g. if he/she has free time or if he/she has the necessary background which will enable him/her to read the book). A further analysis might require inferential processing in order to interpret the text within the application domain or genre correctly. In these cases domain or **real-world knowledge** might have to be resorted to.

Useful NLP programs, tools and resources

Various Natural Language Processing (NLP) programs, tools and resources such as the following are needed to carry out the different levels of analysis.[2] **Morphological analysers** are programs that analyse each word and establish derivations. **Dictionaries** in machine-readable form (also termed **lexicons**) often contain information useful for semantic analysis such as animacy, gender, required agent (for verbs), etc. An **ontology** is a dictionary in which the words are represented as hierarchical concepts with relations such as *part-of* and *is-a* given. **Part-of-speech (POS) taggers** are important corpus-based tools for identifying the grammatical category of each word and some return additional information such as syntactic function (e.g. subject, object, etc.). **Parsers** are programs which provide syntactic analysis of sentences. They use knowledge about words (and word meanings) from the lexicon and a set of rules formalised as **grammar**. A 'lighter' version of a parser that does not deliver full syntactic analysis but is limited to parsing smaller constituents such as noun phrases or prepositional phrases, is called a **shallow parser** (parsers restricted specifically to NP analysis are often termed **NP extractors**). In terms of practical semantic analysis tools, **word-sense disambiguation** programs represent the state of the art.

Chapter 2 gives more details on tools and resources needed for anaphora resolution. Readers not familiar with NLP are advised to consult *Computational Linguistics: An Introduction* (Grishman 1986), *Natural Language Understanding* (Allen 1995) or the *Oxford Handbook of Computational Linguistics* (Mitkov 2002).

Notes

1 In this introduction I refer to 'rules' in their broadest sense: nowadays machine learning algorithms often replace traditional 'if–then' rules.

2 I restrict this brief outline to a selection of widely used NLP tools and do not discuss programs for performing higher level discourse and pragmatic analysis. See Allen (1995) for a detailed account on the latter.

CHAPTER ONE

Linguistic fundamentals

This chapter offers an introduction to anaphora and the concepts associated with it. It outlines the related phenomenon of coreference and classifies the various types of anaphora. The chapter does not aim to provide an all-encompassing theoretical linguistics account of the pervasive phenomenon of anaphora, but seeks to provide the basics for those who wish to familiarise themselves with the field of automatic resolution of anaphora or who plan to undertake practical work in this field, with particular reference to the types of anaphora most widely used.

1.1 Basic notions and terminology

Cohesion is a phenomenon accounting for the observation (and assumption) that what people try to communicate in spoken or written form[1] in 'normal circumstances' is a coherent whole, rather than a collection of isolated or unrelated sentences, phrases or words. Cohesion occurs where the interpretation of some element in the discourse is dependent on that of another and involves the use of abbreviated or alternative linguistic forms which can be recognised and understood by the hearer or the reader, and which refer to or replace previously mentioned items in the spoken or written text.

Consider the following extract from Jane Austen's *Pride and Prejudice*:

(1.1) *Elizabeth* looked archly, and turned away. *Her* resistance had not injured her with the gentleman.[2]

Although it is not stated explicitly, it is normal to assume that the second sentence is related to the first one and that *her* refers to *Elizabeth*. It is this reference which ensures the cohesion between the two sentences. If now the text is changed by replacing *her* with *his* in the second sentence or the whole second sentence is replaced with *This book is about anaphora*, cohesion does not occur any more: the interpretation of the second sentence in both cases no longer depends on the first sentence.

Discourse (1.1) features an example of anaphora with the possessive pronoun *her* referring to the previously mentioned noun phrase *Elizabeth*. Halliday and Hasan (1976) describe **anaphora**[3] as 'cohesion which points back to some previous

item'.[4] The 'pointing back' word or phrase[5] is called an **anaphor**[6] and the entity to which it refers or for which it stands is its **antecedent**. The process of determining the antecedent of an anaphor is called **anaphora resolution**.[7] When the anaphor refers to an antecedent and when both have the same referent in the real world, they are termed **coreferential**. Consider the following example from Huddleston (1984):

(1.2) *The Queen* is not here yet but *she* is expected to arrive in the next half an hour.

In this example, the pronoun *she* is an anaphor, *the Queen* is its antecedent and *she* and *the Queen* are coreferential. Note that the antecedent is not the noun *Queen* but the noun phrase (NP) *the Queen*.

The relation between the anaphor and the antecedent is not to be confused with that between the anaphor and its referent; in the above example the referent *the Queen* is a person in the real world (e.g. Queen Elizabeth) whereas the antecedent *the Queen* is a linguistic form. Next, consider (1.3):

(1.3) *This book* is about anaphora resolution. *The book* is designed to help beginners in the field and *its* author hopes that *it* will be useful.

In this example there are three anaphors referring to the antecedent *this book* – the noun phrase *the book*, the possessive pronoun *its* and the personal pronoun *it* (section 1.4 below will discuss different varieties of anaphora). For all three anaphors, the referent in the real world is the book being read and therefore the anaphors and their antecedent(s)[8] are coreferential.

On the other hand, look at this example:

(1.4) Stephanie *balked*, as *did* Mike.[9]

This sentence features the verb anaphor *did* (see also section 1.4.4) which is a substitution for the antecedent *balked*; however, since the two terms in this anaphoric relation do not have a common referent, one cannot speak of coreference between the two.

1.2 Coreference

The previous section introduced examples of **coreference**, which is the act of picking out the same referent in the real world. As seen in (1.3), a specific anaphor and more than one of the preceding (or following) noun phrases may be coreferential thus forming a **coreferential chain** of entities which have the same referent. As a further illustration, in (1.5) *Sophia Loren*, *she* (from the first sentence), *the actress*, *her* and *she* (second sentence) are coreferential. Coreferential chains partition discourse entities into equivalence classes. In (1.5) the following coreferential chains can be singled out: {Sophia Loren, she, the actress, her, she}, {Bono, the U2 singer}, {a thunderstorm}, {a plane}.[10]

(1.5) *Sophia Loren* says *she* will always be grateful to Bono. *The actress* revealed that the U2 singer helped *her* calm down when *she* became scared by a thunderstorm while travelling on a plane.[11]

Definite noun phrases in copular relation are considered as coreferential, hence in the example

(1.6) *David Beckham* is *the Manchester United midfielder*.[12]

the proper name *David Beckham* and the definite description *the Manchester United midfielder* are coreferential.[13] Coreferential are also *David Beckham* and *the second best player in the world* in (1.7)

(1.7) *David Beckham* was voted *the second best player in the world* behind Rivaldo.[14]

Other examples of copular relations include the relation of apposition illustrated by (1.8):

(1.8) *Dominique Voynet*, *the French Environment Minister*, launched a bitter attack on Mr. Prescott's 'chauvinism'.[15]

In this example the definite noun phrase *the French Environment Minister* is coreferential with the NP to which it applies, in this case *Dominique Voynet*. Since proper names are regarded as definite, in the example

(1.9) Bulger is a fugitive and *his sister*, *Jean Holland*, had tried to stop the Justice Department from seizing Bulger's winnings, one-sixth of a 1991 $14.3 million jackpot.[16]

the NP *Jean Holland* is coreferential with the NP to which it applies (*his sister*). On the other hand, the indefinite predicate nominal *a fugitive* is not normally regarded as coreferential[17] with *Bulger*: the fact that it is not specific enough means that it cannot be viewed as an NP having the same referent in the real world as *Bulger*.[18]

It is important to point out that in some cases an NP without a 'definiteness' modifier (such as *the*, *this*, *that*) can still be regarded as specific and definite, and therefore coreferential with the NP with which it is in a copular relation:

(1.10) *Nicolas Clee*, *editor of the Bookseller*, describes him as a journalist's dream contact.[19]

In this example *editor of the Bookseller* is specific enough to be regarded as definite, and therefore coreferential with *Nicolas Clee*.

Coreference is typical of anaphora realised by pronouns and non-pronominal definite noun phrases (see varieties of anaphora in 1.4), but does not apply to varieties of anaphora that are not based on referring expressions, such as verb anaphora. However, as was already seen with indefinite noun phrases, not every NP triggers coreference. Bound anaphors which have as their antecedent quantifying NPs such as *every man*, *most computational linguistics*, *nobody*, etc., are another example where the anaphor and the antecedent do not corefer. As an illustration, the relation in (1.11) is only anaphoric, whereas in (1.12) it is both anaphoric and coreferential.

(1.11) *Every man* has *his* own destiny.[20]
(1.12) *John* has *his* own destiny.

A **substitution test** can be used to establish coreference in (1.12) resulting in the semantically equivalent sentence

(1.13) *John* has *John's* own destiny.

No such equivalence can be yielded with (1.11) however, where a substitution test produces

(1.14) Every man has every man's destiny.

which is not the same statement as (1.11).

Finally, in the example

(1.15) The man who gave his paycheck to his wife was wiser than the man who gave it to his mistress.[21]

the anaphor *it* and the antecedent *paycheck* do not correspond to the same referent in the real world but to one of a similar description (such type of anaphora is called **identity-of-sense anaphora** as opposed to **identity-of-reference anaphora** in examples (1.3) and (1.5); see also section 1.6 for more details). Therefore, *it* and *his paycheck* are not coreferential.[22]

On the other hand, there may be cases where two items are coreferential without being anaphoric. **Cross-document coreference** is an obvious example: two mentions of the same person in two different documents will be coreferential, but will not stand in anaphoric relation.

Having seen some of the differences between anaphora and coreference, it is worth emphasising that identity-of-reference nominal anaphora[23] involves coreference by virtue of the anaphor and its antecedent having the same real-world referent. Consequently, for anaphora of that type, it would be logical to regard each of the preceding lexical noun phrases[24] that are coreferential with the anaphor(s) as a legitimate antecedent. In the light of this observation, the task of automatic anaphora resolution will be considered successful, if any of the preceding non-pronominal entities in the coreferential chain[25] is identified as an antecedent. Consider again (1.5). Here the antecedent of the anaphors *she* (first sentence) and *the actress* is the noun phrase *Sophia Loren*; both *Sophia Loren* and *the actress* can be considered antecedents for the anaphors *her* and *she* from the second sentence.

This book will focus more on the task of anaphora resolution and less on coreference resolution.[26] Whereas the task of anaphora resolution has to do with tracking down an antecedent of an anaphor, coreference resolution seeks to identify all coreference classes (chains). For more on coreference resolution, it is suggested that the reader consult the Message Understanding Conference (MUC) Proceedings in which coreference resolution is covered extensively (Hirschman and Chinchor 1997).

1.3 Discourse entities

When the antecedent is an NP, it becomes convenient to abstract away from its syntactic realisation in order to capture certain subtleties of its semantics. The

abstraction, termed a **discourse entity**, allows the NP to be modelled as a set of one or more elements and provides a natural metaphor for describing what may on the surface seem to be grammatical number conflicts.

For example, consider (1.16):

(1.16) Lisa could almost see the stars in the black sky, how they had looked that night.[27]

The discourse entity described by the noun phrase *Lisa* consists of one element – the specific person in question, whereas the discourse entity represented by the noun phrase *the stars* incorporates all the stars in the sky that Lisa could 'almost see'.

Consider now (1.17):

(1.17) The teacher gave *each child* a crayon. *They* started drawing colourful pictures.

The discourse entity represented by the noun phrase *each child* comprises all children in the teacher's class and is therefore referred to by a plural anaphor.

Finally, in (1.18) the antecedent of the plural anaphor *they* is *the police*, which as a noun phrase is singular:

(1.18) Had *the police* taken all the statements *they* needed from her?[28]

If the discourse entity associated with the NP *the police* is now considered, it is easy to explain the number 'mismatch': this discourse entity as a set contains more than one element.

Therefore, the anaphor agrees with the number of the discourse entity (for more on agreement, see Chapter 2, section 2.1.1) associated with its antecedent rather than the number of the NP representing it.[29]

For the sake of simplicity, I shall often limit the treatment of the antecedent to its classical definition as a linguistic form (e.g. surface constituent such as noun phrase) and, therefore, refrain from searching for an associated discourse entity (e.g. semantic set). This is an approach widely adopted by a number of anaphora resolution systems that do not have recourse to sophisticated semantic analysis. It should be borne in mind, however, that there are cases where more detailed semantic description or processing is required for the successful resolution (see Chapter 2, section 2.1.3).

1.4 Varieties of anaphora according to the form of the anaphor

Nominal anaphora arises when a **referring expression** (pronoun, definite noun phrase or proper name) has a non-pronominal noun phrase as its antecedent. This most important and frequently occurring class of anaphora has been researched and covered most extensively, and is the best understood in the Natural Language Processing (NLP) literature. As a consequence, this book will be looking mainly at the computational treatment of nominal anaphora.

1.4.1 Pronominal anaphora

The most widespread type of anaphora is that of **pronominal anaphora**. Pronominal anaphora occurs at the level of *personal pronouns* (The most difficult for Dalí was to tell her, between two [sic] of nervous laughter, that *he* loved her.[30]), *possessive pronouns* (But the best things about Dalí are *his* roots and *his* antennae.[31]), *reflexive pronouns* (Dalí once again locked *himself* in his studio . . .[32]) and *demonstrative pronouns* (Dalí, however, used photographic precision to transcribe the images of his dreams. *This* would become one of the constraints of his work . . .[33]). *Relative pronouns* are regarded as anaphoric too (Dalí, a Catalan *who* was addicted to fame and gold, painted a lot and talked a lot.[34]).

The set of anaphoric pronouns consists of all third person personal (*he*, *him*, *she*, *her*, *it*, *they*, *them*), possessive (*his*, *her*, *hers*, *its*, *their*, *theirs*) and reflexive (*himself*, *herself*, *itself*, *themselves*) pronouns plus the demonstrative (*this*, *that*, *these*, *those*) and relative (*who*, *whom*, *which*, *whose*) pronouns both singular and plural (*where* and *when* are anaphoric too, see section 1.4.4 for locative and temporal anaphora). Pronouns first and second person singular and plural are usually used in a deictic manner[35] (*I* would like *you* to show me the way to San Marino) although their anaphoric function is not uncommon in reported speech or dialogues as the use of *I* in (1.19) and (1.25), and the use of *you* in (1.20):

(1.19) 'He is beautiful,' *Isabel* told the woman, of her own son. '*I* feel incomplete when I am not with him.'[36]
(1.20) *James*, don't cross-examine me. *You* sound like a prosecuting counsel.[37]

1.4.1.1 PLEONASTIC *IT*

In addition to the first and second person pronouns, the pronoun *it* can often be non-anaphoric. For example, in (1.21) *it* is not specific enough to be considered anaphoric:

(1.21) It is dangerous to be beautiful – that is how women have learned shame.[38]

Non-anaphoric uses of *it* are also referred to as **pleonastic**[39] (Lappin and Leass 1994) or **prop it** (Quirk et al. 1985). Examples of pleonastic *it* include non-referential instances of

(a) *It* appearing in constructions with modal adjectives such as *It is dangerous*, *It is important*, *It is necessary*, *It is sufficient*, *It is obvious*, *It is useful*, etc.
(b) *It* in various constructions with cognitive verbs such as *It is believed that . . .*, *It appears that . . .*, *It should be pointed out that . . .*, etc.
(c) *It* appearing in constructions describing weather conditions such as *It is raining*, *It is sunny*, *It is drizzling*, etc.
(d) *It* in temporal constructions such as *It is five o'clock*, *It is high time* (we set off), *It is late*, *It is tea time*, *It is winter*, *What day is it today?*, etc.
(e) *It* in constructions related to distance such as *How far is it to Wolverhampton?*, *It is a long way from here to Tokyo*.

(f) *It* in idiomatic constructions such as *At least we've made it*, *Stick it out*, *Call it quits*, *How's it going?*[40]
(g) *It* in cleft constructions such as *It was Mr. Edgar who recruited Prudence Adair*.[41]

Non-anaphoric uses of *it* are not always a clear cut case and some occurrences of *it* appear to be less unspecified than others and are therefore a matter of debate in linguistics. For further discussion of this issue see Morgan (1968).

The automatic identification of pleonastic *it* in English is not a trivial task. For further discussion see section 2.2.1.

1.4.1.2 OTHER NON-ANAPHORIC USES OF PRONOUNS

In addition to pleonastic *it*, there are other non-anaphoric uses of third person pronouns in English. The **generic** use of pronouns is frequently observed in proverbs or sayings:

(1.22) He that plants thorns must never expect to gather roses.
(1.23) He who dares wins.

The **deictic** use (see note 34; see also section 1.11) of third person pronouns is not uncommon in conversation. For example, some time ago I went shopping with my son, then 3 years old. Upon reaching the till he explained to me that we had spent a lot of money so that we now had less money than we had started the shopping trip with. The cash assistant must have overheard his comments and I was chuffed when she said:

(1.24) He seems remarkably bright for a child of his age.

In this case *he* was not used anaphorically but deictically; in fact there had been no mention of the little boy prior to the utterance.

1.4.2 Lexical noun phrase anaphora

Lexical noun phrase anaphora is realised syntactically as definite noun phrases, also called **definite descriptions** (Russell 1905), and **proper names**. Although personal, reflexive, possessive and demonstrative pronouns[42] as well as definite descriptions and proper names are all considered definite expressions, only lexical noun phrases and not pronouns have a meaning independent of their antecedent. Furthermore, definite descriptions do more than just refer. They convey some additional information, as in (1.25), where the reader can learn more about *Roy Keane* through the definite description *Alex Ferguson's No. 1 player*.

(1.25) *Roy Keane* has warned Manchester United *he* may snub their pay deal. *United's skipper* is even hinting that unless the future Old Trafford Package meets *his* demands, *he* could quit the club in June 2000. *Irishman Keane*, 27, still has 17 months to run on *his* current £23,000-a-week contract and wants to commit *himself* to United for life. *Alex Ferguson's No. 1 player* confirmed: 'If it's not the contract *I* want, *I* won't sign.'[43]

In this text, *Roy Keane* has been referred to by anaphoric pronouns (*he*, *his*, *himself*, *I*), but also by definite descriptions (*United's skipper*, *Alex Ferguson's No. 1 player*) and a proper name (*Irishman Keane*).[44] Furthermore, *Manchester United* is referred to by the definite description *the club* and by the proper name *United*.

The additional information conveyed by definite referring expressions frequently stands in predictable semantic relation to the antecedent, and thus increases the cohesiveness of the text. Lexical noun phrase anaphors may have the same head as their antecedents (*these footprints* and *the footprints*, see example (1.27)) or the relationship between the referring expression and its antecedent may be that of synonymy (*shop . . . the store*), generalisation/hypernymy (*boutique . . . the shop*, also *Manchester United . . . the club* as in (1.25)) or specialisation/hyponymy (*shop . . . the boutique*, also *their pay deal . . . his current £23,000-a-week contract* as in (1.25)).[45] Proper names[46] often refer to antecedents whose names they match in whole or in part (*Manchester United . . . United*) with exact repetition not being uncommon:

(1.26) *Alice* was as nervous as a kitten on the eve of Miles' party. That's *Alice* for you.[47]

Certain determiners such as *the*, *this*, *these*, *that* and *those* signal that the noun phrase they modify is coreferential to a previous noun phrase.

(1.27) Both noses went down to *the footprints* in the snow. *These footprints* were very fresh.[48]

We have already seen that coreferential noun phrases may have identical heads, but also that noun phrases may be coreferential even if their heads are not identical. On the other hand, identity of heads does not necessarily imply coreference of two noun phrases. For example:

(1.28) The rooms on the first floor and ground floor did not reveal anything odd.[49]

In this example, *the first floor* is not coreferential with *ground floor*. Similarly, in (1.25) *his current £23,000-a-week contract* and *the contract I want* are not coreferential.

Finally, definite descriptions are not always anaphoric and their **generic** use is not uncommon:

(1.29) No one knows precisely when *the wheel* was invented.
(1.30) George enjoys playing *the piano*.

1.4.3 Noun anaphora

Noun phrase anaphora should not be confused with **noun anaphora** – the anaphoric relation between a non-lexical proform and the head noun or nominal group[50] of a noun phrase. Noun anaphora represents a particular case of identity-of-sense anaphora (see example (1.15) above).

(1.31) I don't think I'll have a sweet *pretzel*, just a plain *one*.[51]

The non-lexical proform *one* constitutes an example of a noun anaphor. Note that *one* points to the noun *pretzel* and not to the noun phrase *a sweet pretzel*.

1.4.4 Verb anaphora, adverb anaphora

Among the other varieties of anaphora according to the form of the anaphor, **verb anaphora** should be mentioned. In the sentence:

(1.32) When Manchester United swooped to lure Ron Atkinson away from the Albion, it was inevitable that his midfield prodigy would *follow*, and in 1981 he *did*.[52]

the interpretation of *did* is determined by its anaphoric relation[53] to its antecedent in the preceding clause. Whereas in (1.32) the anaphor *did* stands for the verb *followed*, the verb anaphor *did* in (1.33) replaces the verb phrase *begged for reinforcements*:

(1.33) Romeo Dallaire, the Canadian general in charge, *begged for reinforcements*; so *did* Boutros-Ghali.[54]

We also distinguish **adverb anaphora** which can be *locative* such as *there* (1.34) or *temporal* anaphora such as *then* (1.35).

(1.34) Will you walk with me to *the garden*? I've got to go down *there* and Bugs has to go to the longhouse.[55]

(1.35) For centuries archaeologists have argued over descriptions of how Archimedes used concentrated solar energy to destroy the Roman fleet *in 212BC*. Historians have said nobody *then* knew enough about optics and mirrors.[56]

As previously illustrated with first and second person pronouns, adverbs of this type are frequently used not anaphorically but deictically, taking their meaning from contextual elements such as the time or location of utterance.

It has already been shown that the anaphors can be verbs and adverbs, as well as nouns and noun phrases,[57] and thus span the major part-of-speech categories.

1.4.5 Zero anaphora

Another important class of anaphora according to the form of the anaphor is the so-called **zero anaphora** or **ellipsis**. Zero anaphors (signalled below by ∅) are 'invisible' anaphors – at first glance they do not appear to be there because they are not overtly represented by a word or phrase. Since one of the properties and advantages of anaphora is its ability to reduce the amount of information to be presented via abbreviated linguistic forms, ellipsis may be the most sophisticated variety of anaphora.[58]

Ellipsis is the phenomenon associated with the deletion of linguistic forms, thus enhancing rather than damaging the coherence of a sentence or a discourse segment. The resultant 'gap' (zero anaphor) signals the necessity of recovering the meaning via its antecedent.

The most common forms of ellipsis are zero pronominal anaphora, zero noun anaphora and verb (phrase) ellipsis.

1.4.5.1 ZERO PRONOMINAL ANAPHORA

Zero pronominal anaphora occurs when the anaphoric pronoun is omitted but is nevertheless understood. This phenomenon occurs in English in a somewhat restricted environment, but is so pervasive in other languages such as Spanish, Italian, Portuguese, Polish, Chinese, Japanese, Korean and Thai, that NLP applications covering these languages cannot circumvent the problem of zero anaphora resolution. Consider the first sentence in this paragraph.

(1.36) *Zero pronominal anaphora* occurs when the anaphoric pronoun is omitted but ∅ is nevertheless understood.

The third clause in this sentence features zero pronominal anaphora (the expected full form would have been *but it is nevertheless understood*).[59]

Similarly the second clause of the sentence

(1.37) *Willie* paled and ∅ pulled the sock up quickly.[60]

contains a zero pronominal anaphor.

In some languages verb agreement points to a zero pronoun. As an illustration, consider the following example in Spanish:

(1.38) *Marta* está muy cansada. ∅ Ha estado trabajando todo el día.
Marta is very tired. (*She*) Has been working all day long.

Japanese, Chinese and Korean are languages with extensive use of zero pronouns.[61] The following is an example of zero pronominal anaphora in Japanese.

(1.39) Nihongo o hanasu no wa kantan desu ga kaku no wa muzukashii desu.
Speaking *Japanese* is easy but writing ∅ (= it) is difficult.

A study of anaphoric pronouns in parallel English and Japanese texts conducted by Uehara (1996) exemplifies the pervasive distribution of zero pronouns in Japanese. This study found[62] that 14.5% of the English anaphoric pronouns were retained in Japanese as overt pronouns, 29% were replaced by overt noun phrases and 56.5% were 'deleted' as zero pronouns.[63]

1.4.5.2 ZERO NOUN ANAPHORA

Zero noun anaphora arises when the head noun only – and not the whole NP – is elliptically omitted (the reference is realised by the 'non-omitted', overt modifiers). Typical overt modifiers of zero anaphoric nouns in English are the indefinites *several*, *few*, *some*, *many*, *more*.

(1.40) George was bought a huge box of *chocolates* but few ∅ were left by the end of the day.
(1.41) Jenny ordered three *copies* of the document and Conny ordered several ∅ too.

In (1.40) and (1.41) the empty set sign ∅ stands for the elliptically omitted *chocolates* and *copies* respectively.

1.4.5.3 ZERO VERB ANAPHORA

Zero verb anaphora occurs when the verb is omitted elliptically and the zero anaphor points to a verb in a previous clause or sentence:

(1.42) *Win* a Golf GTi or ∅ a week in Florida or ∅ weekend in Paris.[64]

The zero verb anaphors, ∅, stand for the verb *win* in the clause *Win a Golf GTi*.

1.4.5.4 VERB PHRASE ZERO ANAPHORA (ELLIPSIS)

Verb phrase zero anaphora, also termed **ellipsis**, is the omission of a verb phrase which leaves a gap pointing to a verb phrase antecedent, usually in a previous clause, and which enhances the readability and coherence of the text by avoiding repetition.

(1.43) I have never *been to Miami* but my father has ∅, and he says it was wonderful.

In this example ∅ stands for the verb phrase *been to Miami*.

Finally, it is interesting to note that the antecedent can be elliptically omitted too as in (1.44):

(1.44) I have not got a car myself but Tom has ∅, and I think I'll be able to persuade him to let us borrow *it*.[65]

1.5 Types of anaphora according to the locations of the anaphor and the antecedent

The varieties of anaphora discussed so far are based on the different types of words which refer back to (or replace) a previously mentioned item. Depending on the location of the antecedent, **intrasentential** (sentence) anaphora and **intersentential** (discourse) anaphora can be observed.

Intrasentential anaphora arises if the anaphor and its antecedent are located in the same sentence. On the other hand, intersentential anaphora is exhibited when the antecedent is in a different sentence from the anaphor. Reflexive pronouns are typical examples of intrasentential anaphors. Possessive pronouns can often be used as intrasentential anaphors too, and can even be located in the same clause as the anaphor. In contrast, personal pronouns and noun phrases acting as intrasentential anaphors usually have their antecedents located in the preceding clause(s) of the same complex sentence.

(1.45) Pop superstar *Robbie Williams* hid *his* secret heartbreak as *he* picked up three Brit awards last night. *He* was stunned to discover that *his*

ex-fiancée, All Saints beauty Nicole Appleton, is dating a New York rapper. *Robbie*, 25, was distraught after being dumped by the love of *his* life Nicole at Christmas.[66]

In the first sentence of (1.45) the anaphoric pronouns *his* and *he* are examples of intrasentential anaphors having their antecedent in the same sentence (the antecedent of *he* is in a preceding clause but still in the same sentence). On the other hand, *he* and *his* in the second sentence, and *Robbie* in the third sentence, act as intersentential anaphors since their antecedent is in a preceding sentence.

The distinction between intrasentential and intersentential anaphora is of practical importance for the design of an anaphora resolution algorithm. As pointed out in 5.3.1, 5.4 and 7.4.2, syntax constraints could play a key role in the resolution of intrasentential anaphors.

1.6 Indirect anaphora

Indirect anaphora[67] arises when a reference becomes part of the hearer's or reader's knowledge indirectly rather than by direct mention, as in (1.46):

(1.46) Although *the store* had only just opened, *the food hall* was busy and there were long queues at *the tills*.[68]

In (1.46) the noun phrase *the store* is regarded as antecedent of the indirect anaphors *the food hall* and *the tills*. It can be inferred that *the tills* make an indirect reference to *the store* because it is known that stores have tills and because *the store* has already been mentioned. Similarly, *the food hall* is understood to be part of the store. The inference may require more specialised 'domain' knowledge, however, and in the example:

(1.47) When *Take That* broke up, the critics gave *Robbie Williams* no chance of success.[69]

one must know that *Robbie Williams* was a member of the former pop group *Take That* in order to be able to infer the indirect reference.[70]

The above examples feature relationships such as *part-of* (1.46) and *set membership* (1.47) between the anaphor and its antecedent.[71] The latter includes the relationship *subset–set* between the anaphor and its antecedent as in (1.53) which are also instances of indirect anaphora. The distinction between direct and indirect anaphora is not clear-cut. Many definite descriptions can serve as examples of indirect anaphora and the amount of knowledge required to establish the antecedent may vary depending on whether the relation between the anaphor and the antecedent is that of generalisation, specialisation or even synonymy.[72] In example (1.25), for instance, some of the coreferential links can be established only on the basis of the knowledge that Roy Keane is Irish or that he is Manchester United's skipper. Hence some researchers (Vieira and Poesio 2000b) use the term **direct anaphora** to refer exclusively to the cases when the definite description and the antecedent have identical heads.

1.7 Identity-of-sense anaphora and identity-of-reference anaphora

In all preceding examples of pronominal and lexical noun phrase anaphora (except examples (1.15) and (1.31)) the anaphor and the antecedent have the same referent in the real world and are therefore coreferential. These examples demonstrate **identity-of-reference anaphora**, with the anaphor and the antecedent denoting the same entity. For example:

(1.48) In Barcombe, East Sussex, a family had to flee *their cottage* when *it* was hit by lightning.[73]

The anaphor *it* and *their cottage* have the same referent: the cottage that belonged to the family and that was hit by lightning. In (1.15), however:

(1.15) The man who gave his paycheck to his wife was wiser than the man who gave it to his mistress.

paycheck and *it* do not refer to the same entity but to one of a similar description. In particular, *it* refers to the paycheck of the second (less wise) man. Similarly, in (1.49)

(1.49) The physicians who had eaten strawberries were much happier than the physicians who had eaten egg sandwiches for lunch.[74]

the two mentions of *the physicians* are not coreferential.

This type of anaphora is called **identity-of-sense anaphora**. An identity-of-sense anaphor does not denote the same entity as its antecedent, but one of a similar description. Clearly identity-of-sense anaphora does not, by definition, trigger coreference because the anaphor and the antecedent do not have the same referent.

A further example of identity-of-sense anaphora is the sentence:

(1.50) The man who has his hair cut at the barber's is more sensible than the one who has it done at the hairdresser's. [75]

Note the identity-of-sense anaphors *it* and *the one*. The latter refers to an item of similar description (man) that is different from the man who has his hair cut at the barber's .

The following sentences supply yet more examples of identity-of-sense anaphora:

(1.51) George picked a *plum* from the tree. Vicky picked *one* too.
(1.52) Jenny ordered five *books*. Olivia ordered *several* too.

In (1.51) and (1.52) the anaphors *one* and *several*[76] refer to entities of a different description from their antecedents (Vicky picked a different plum from George; the books ordered by Olivia are different from those ordered by Jenny).

Note, on the other hand, that *several* in

(1.53) Jenny bought *10 apples*. *Several* were rotten.

is still an example of an identity-of-reference anaphor. In addition, (1.53) can be regarded as an instance of indirect anaphora since the discourse entity associated with the anaphor (several apples) is a subset of the discourse entity associated with the antecedent (10 apples).

Finally, it is worth mentioning that it is possible to come across anaphors that can be read either as identity-of-reference or as identity-of-sense, thus rendering the text ambiguous:

(1.54) John likes his hair short but Jenny likes it long.[77]

It can be either John's hair (identity-of-reference anaphora) or Jenny's hair (identity-of-sense anaphora).

1.8 Types of antecedents

This book, like most NLP projects, concentrates on anaphors whose antecedents are noun phrases. As already seen, however, even though these are the most common and best studied types of anaphors, they are not the only ones. An anaphor can replace/refer to a noun (example (1.31)), verb (1.32) and verb phrase (1.33). Also, the antecedent of a demonstrative pronoun[78] or the antecedent of the personal pronoun *it* can be a noun phrase, clause (1.55), sentence (1.56), or sequence of sentences (1.57).

(1.55) *Owen tried to help her with something*: *this* made indeed for disorder.[79]

(1.56) *They will probably win the match*. *That* will please my mother.[80]

(1.57) *Many years ago their wives quarrelled over some trivial matter, long forgotten*. *But one word led to another, and the quarrel developed into a permanent rupture between them*. *That*'s why the two men never visit each other's houses.[81]

In some cases, anaphors may have **coordinated** antecedents – two or more noun phrases coordinated by *and* or other conjunctions.[82] The anaphor in this case must be plural, even if each of the noun phrases is singular.

(1.58) The cliff rose high above *Paul and Clara* on their right hand. *They* stood against the tree in the watery silence.[83]

Similarly, a coordinated antecedent can arise when a list of noun phrases is separated by commas and/or a conjunction.

(1.59) Among the newspaper critics present, at that time unknown to each other and to James, were three men shortly destined to become the most celebrated writers of the age – *George Bernard Shaw, Arnold Bennett and H.G.Wells*. *They* appreciated James's intelligent dialogue. . . .[84]

1.9 Location of the antecedent

Information about the expected/possible distance between the anaphor and the closest antecedent[85] is not only interesting from the point of view of theoretical

linguistics, but can be very important practically and computationally in that it can narrow down the search scope of candidates for antecedents.[86] Empirical evidence suggests that the distance between a pronominal anaphor and its antecedent in most cases does not exceed 2–3 sentences. Hobbs (1978) found that 98% of the pronoun antecedents were in the same sentence as the pronoun or in the previous one. Pérez (1994) studied the SUSANNE manually tagged corpus[87] and reported that out of 269 personal pronouns, 83 had their antecedents in the same sentence, whereas 126 referred to an entity in the preceding sentence. Moreover, 16 pronouns had their antecedents two sentences back, whereas 44 pronouns had their antecedent three sentences back. A study based on 4681 anaphors from the UCREL Anaphoric Treebank corpus conducted by McEnery et al. (1997) established that in 85.64% of cases the antecedent was within a window of 3 sentences (current, previous and prior to the previous), whereas 94.91% of the antecedents were no further than 5 sentences away from the anaphor. Fraurud's (1988) study of novels, reports of court procedures and articles about technological innovations in Swedish found that in about 90% of the cases the antecedent was located in the same sentence as the anaphor or in the preceding one. Guindon (1988) obtained similar results for spoken dialogues as did Dahlbäck's (1992) findings for Swedish.

Both Fraurud and Guindon note that there is a small class of long-distance anaphors whose antecedents are not in the same or the preceding sentence. The greatest distance between a pronominal anaphor and its antecedent reported in Hobbs (1978) is 13 sentences and in Fraurud (1988) is 15 sentences. Fraurud's investigation also established that the animacy of the antecedent is a factor for long-distance pronominalisation: usually pronouns referring to humans can have their antecedents further away. This tendency was especially evident in the stories and it looks as if long-distance anaphors are more typical of certain genres. Biber et al. (1998) concluded that in news reportage and academic prose the distance between anaphors and their antecedents is greater than in conversation and public speeches.[88] Hitzeman and Poesio (1998) analysed a small corpus of oral descriptions of museum items and found that the long-distance pronouns comprise about 8.4% in this kind of data. However, for more conclusive results further analysis involving larger and more representative samples is needed. Hitzeman and Poesio's analysis looked at 83 pronouns only; Fraurud's findings were based on a sample consisting of 600 pronouns, and so cannot be regarded as definitive either.

Ariel (1990) conducted a corpus-based analysis and concluded that demonstrative anaphors[89] were normally longer-distance anaphors than pronouns, but the distance between definite descriptions or proper names and their antecedents may be even greater. In fact the present writer found it quite common for proper names to refer to antecedents which are 30 or more sentences away. For example, in one newspaper article[90] President Ronald Reagan's national security adviser *Robert McFarlane* was referred to by the proper name *McFarlane* 35 sentences (many of which were long and with complicated syntax) and 14 paragraphs after it was last mentioned.

For practical reasons most pronoun resolution systems restrict their search to the preceding 2–3 sentences when looking for an antecedent (see Kameyama 1997; Mitkov 1998b). On the other hand, since anaphoric definite noun phrases may have their antecedents further away, strategies for their resolution have involved the search of the 10 preceding sentences (Kameyama 1997).

1.10 Anaphora and cataphora

Cataphora arises when a reference is made to an entity mentioned subsequently in the text.

(1.60) *She* is now as famous as her ex-boyfriend. From the deserts of Kazakhstan to the south seas of Tonga, everyone knows *Monica Lewinsky*.[91]

In this example *she* refers to *Monica Lewinsky*, mentioned subsequently. Cataphora is similar to anaphora, the difference being the direction of the pointing (reference).

Where cataphora occurs, anaphoric reference is also possible and can be obtained by reversing the positions of the anaphor and the antecedent.[92] The new sentence is synonymous to the original one.[93]

(1.61) *Monica Lewinsky* is now as famous as her ex-boyfriend. From the deserts of Kazakhstan to the south seas of Tonga, everyone knows *her*.

Example (1.60) illustrates intersentential cataphora, but in English intrasentential cataphora is more usual.

(1.62) The elevator opened for *him* on the 14th floor, and *Alec* stepped out quickly.[94]

Typically, intrasentential cataphora occurs where the cataphoric pronoun is in a subordinate clause.[95]

(1.63) Lifting *his* feet high out of the sand, *Ralph* started to stroll past.[96]

Intrasentential cataphora is exhibited only by pronouns,[97] as opposed to intersentential cataphora which can be signalled by non-pronominal noun phrases too[98]:

(1.64) *The former White House intern* is now as famous as her ex-boyfriend. From the deserts of Kazakhstan to the south seas of Tonga, everyone knows *Monica Lewinsky*.

The nature of cataphora has been discussed and disputed by a number of researchers, both within the generative framework and outside it.[99] Some linguists such as Kuno (1972, 1975), Bolinger (1977) and Cornish (1996) argue against the genuine existence of cataphora, claiming that alleged cataphoric

pronouns must have, located in the previous text, corresponding coreferential items. Their observations are based on examples such as

(1.65) Though *her* party comprised 20 supporters, *Hillary* and a female colleague were the only two eating and the bill was $6.[100]

where even though the occurrence of *her* appears to be cataphoric, this is not the case if the extract is examined within the context (1.66) of the whole document, rather than in isolation (1.65).

(1.66) At about 10am, two men in suits appeared, asking to talk to the manager. It turned out they were Secret Service agents wanting to know if *Hillary Clinton* could pop in for breakfast [. . .] Though *her* party comprised 20 supporters, *Hillary* and a female colleague were the only two eating and the bill was $6.

On the other hand researchers such as Carden (1982) and Tanaka (2000) demonstrate that genuine cataphora does exist. Carden (1982) supports his argument with approximately 800 examples of cataphoric cases where such pronouns are, as he claims, the 'first mention of its referent in the discourse'.[101] Such a type of cataphora is described as 'first-mention' cataphora and counteracts the aforementioned scepticism that assumes that each pronoun acting cataphorically must possess a previously mentioned discourse referent.

The use of cataphoric references is typical in literary and journalistic writing and the following is an example of genuine cataphora.

(1.67) From the corner of the divan of Persian saddle-bags on which *he* was lying, smoking, as was *his* custom, innumerable cigarettes, *Lord Henry Wotton* could just catch the gleam of the honey-sweet blossoms of a laburnum . . .[102]

As this text occurs in the second paragraph of the first chapter of the book and there is no direct or indirect mention of Lord Henry Wotton in the first paragraph of this chapter, its title or the title of the book, it would not be possible to analyse the pronouns *he* and *his* as anything other than cataphoric.

1.11 Anaphora and deixis

In the example previously quoted

(1.24) He seems remarkably bright for a child of his age.

the pronoun *he* was not used anaphorically, but deictically: *he* did not refer to an item previously mentioned in the discourse, but pointed to a specific person in a given situation. The information that could have been derived from a potential antecedent was not necessary on this occasion and the statement was not dependent on information explicitly present in a text or discourse. However, if the above sentence had been preceded by the sentence *George is only 4 but can read and write in both English and Bulgarian*, the pronoun *he* would have been

interpreted anaphorically. **Deixis** is the linguistic phenomenon of picking out a person, object, place, etc. in a specific context or situation. The interpretation of the deictically used expression is determined in relation to certain features of the utterance act, such as the identity of the speaker and addressee together with the time and place at which it occurs (Huddleston, 1984). As an illustration, consider the utterance:

(1.68) I want you to be here now.

The deictic pronoun *I* refers to whoever is uttering the sentence and the pronoun *you* to whoever the addressee is. Similarly, the interpretations of *here* and *now* are associated respectively with the place and time of the utterance.

Among the words typically used in a deictic way are the personal pronouns *I*, *we*, *you* and their reflexive and possessive counterparts; the demonstratives *this* and *that*; the locatives *here* and *there* and a variety of temporal expressions such as *now*, *then*, *today*, *tomorrow*, *yesterday*, *next week*, *last month*, *next year*, *in the last decade*, *this century*, *last century*, *on Sunday*, etc.:

(1.69) *I* know that *you* will enjoy reading *this* chapter.
(1.70) *I* bet *you* were expecting *that* example.
(1.71) It was very fashionable to wear long hair *then*. (*then* deictic, e.g. uttered while watching a film)
(1.72) *Last century* has witnessed a real technological revolution. (*Last century* deictic, e.g. uttered at the beginning of the 21st century)

I have already shown that third person pronouns are usually anaphoric but sometimes they can be used deictically (1.24); on the other hand most uses of first and second person are not anaphoric. Demonstrative pronouns such as *that* are used both deictically (1.70) and anaphorically (When I used to ask my then[103] two-and-a-half-year-old son 'George, would you like to eat *a green pepper*?' he would reply 'I don't like *that*'). Similarly, adverbs such as *then* can be both deictic (1.71) and anaphoric (1.35).

Finally, there are uses that are simultaneously anaphoric and deictic:

(1.73) Maggie came[104] to England when she was four, and has lived *here* ever since.[105]

In (1.73) *here* is deictic in that it refers to the place where the utterance occurs but at the same time it is anaphoric to *England*, previously introduced in the text.

1.12 Anaphora and ambiguity

Many anaphors like *she* in (1.74)

(1.74) Jane told Mary she was in love.

are ambiguous – *she* could be either *Jane* or *Mary*. Equally ambiguous is the example

(1.75) Jane convinced Mary she was in love.

Often the level of ambiguity in similar examples depends on the semantics of the verb or other components in the sentence or discourse.

(1.76) Jane informed Mary she was in love.

In this example it is more likely that *Jane* was in love because if *Mary* were in love herself, perhaps she would not have needed to be informed of it.

Similarly,

(1.77) Jane told Mary she was in danger.

is ambiguous whereas in

(1.78) Jane warned Mary she was in danger.

Mary is by far the more probable antecedent because of the semantics of the verb *to warn* which focuses on the person being warned (and hence, the danger to the addressee).

In practice, however, some readings are much more probable than others:

(1.79) Jane told Sarah she was the nicest person she knew of.[106]

Even though this sentence is theoretically ambiguous (with four different meanings: each *she* can be either *Jane* or *Sarah*), in practice it is much more probable that Jane would praise somebody else rather than showing off so immodestly; therefore, *Sarah* would be the preferred antecedent of the first *she*. Similarly, *Jane* is inevitably the antecedent of the second *she*, since Jane cannot have 'inside knowledge' of what Sarah knows.

These examples illustrate that in many cases of ambiguous anaphors there is a probable, preferred or default antecedent,[107] which is taken as the correct one 'in the absence of contradicting context or knowledge' (Hirst 1981).

In many cases the preferred reading relies on extralinguistic knowledge such as

(1.80) Prime Minister Tony Blair had a fruitful meeting with President Yeltsin. The old man has just recovered from a heart attack.

The antecedent of *The old man* is most probably *President Yeltsin* who is known to be much older than Tony Blair and has poor health at the time of writing.

1.13 Anaphora and the resolution moment

The interpretation of anaphora may be delayed until other discourse elements intervene to elucidate the anaphoric reference. This becomes clear in the following example (Tanaka 2000: 221):

(1.81) Police officer David Cheshire went to Dillard's home. Putting his ear next to Dillard's head, Cheshire heard the music also.

The disambiguation moment of the pronoun *his* is the moment the reader processes *Dillard's head*. At this moment the reader would have no difficulty to

instantiate *David Cheshire* to the anaphor *his* instead of *Dillard*, since one cannot put one's ear next to one's own head. Therefore, the resolution moment is not that of the pronoun reading but a later one. Example (1.81) suggests that there is a distinction between the point when a reader encounters an anaphor and begins to interpret it (*initiation point*), and the point when the reader completes the interpretation of the pronoun (*completion point*). As Sanford and Garrod (1989) note, the gap between the two points can be almost nil, as in the case when a reader resolves a pronoun immediately after she/he encounters it. In other cases, the gap can be extended to the end of the phrase, clause, or sentence in which the pronoun is included. The problem of delayed resolution is also discussed in Cristea and Dima (2000).

1.14 Summary

This chapter introduces the linguistic phenomenon of *anaphora* (the act of pointing back to a previously mentioned item) and related phenomena and concepts.[108] I have shown that *anaphora* and *coreference* (the act of referring to the same referent in the real world) are not the same thing even though important classes of anaphora involve coreference. I have also outlined the related phenomena of *cataphora* (backwards anaphora) and *deixis* (non-textual reference in a specific situation). The classification of the varieties of anaphora proposed in this chapter aims to be simple enough for the purpose of Natural Language Processing (NLP).[109] I have pointed out that *nominal anaphora*, that is, anaphora exhibited by pronouns and lexical noun phrases[110] that refer to noun phrases, is the most crucial and best understood class in NLP. I have distinguished varieties of anaphora (i) according to the form of the anaphor (*pronominal*, *lexical noun phrase*, *noun*, *verb*, *zero* anaphora, etc.), (ii) according to the location of the anaphor and the antecedent (*intrasentential* as opposed to *intersentential*), (iii) according to the inference needed (*indirect* as opposed to *direct*) and (iv) according to whether the anaphor and the antecedent have the same referent in the real world or one of a similar description (*identity-of-reference* or *identity-of-sense* anaphora). Finally, I have briefly discussed the typical distance between the different varieties of nominal anaphora and their antecedents, and have alerted the reader to the fact that anaphors may be ambiguous.

Notes

1 Or in any other appropriate mode of communication such as gestural or more generally multimodal communication, sign language, etc.

2 Jane Austen, *Pride and Prejudice*, Ch. 6, p. 23. London: Penguin, 1995.

3 The etymology of the term *anaphora* goes back to Ancient Greek: *anaphora* (αναφορα) is a compound word consisting of the separate words *ana* (ανα), back, upstream, back in an upward direction, and *phora* (φορα), the act of carrying. Anaphora thus denoted the act of carrying back upstream.

4 Note that anaphora is not merely the act of referring to a previously mentioned item in a text: as will be seen later, not every type of anaphora is referential, that is, has a referring function (e.g. verb anaphora).

5 The 'pointing back' word (phrase) is also called a **referring expression** if it has a referential function.

6 As a matter of accuracy, note that *anaphora* is a linguistic phenomenon and not the plural of *anaphor* (the latter is the word/phrase pointing back), as it has been wrongly referred to as in some work on anaphora resolution so far.

7 In the literature both terms *anaphora resolution* and *anaphor resolution* have been used. Perhaps one can argue that *anaphor resolution* is a no less precise term since (i) it would be logical to say that the *anaphor* is resolved to its antecedent and (ii) it is acceptable to say *pronoun resolution* (which would be the 'parallel form' to *anaphor resolution*) but not *pronominalisation resolution* (the parallel form to *anaphora resolution*). However, *anaphora resolution* has established itself as a more widespread term and therefore has been adopted throughout this book.

8 In this example, both *this book* and *the book* can be regarded as antecedents of the anaphors *it* and *its* (see also section 1.3).

9 Ian MacMillan, *Light and Power Stories*, Story 5 'Idiot's Rebellion', p. 51. Columbia and London: University of Missouri Press, 1980.

10 The notion of coreference can be formally defined as a relation and the coreference chains can be described as equivalence classes. In particular, if we introduce the relation *t antecedes x* between an anaphor *x* and an antecedent *t* (note that this definition would apply to identity-of-reference anaphora only), then two discourse entities *x* and *t* are said to be coreferential (notated as *coref* (*x*, *t*)) if any of the following holds (Lappin and Leass 1994): (i) *t antecedes x*; (ii) *x antecedes t*; (iii) *s antecedes x* for some discourse entity *s* and *coref* (*s*, *t*) and (iv) *s antecedes t* for some *s* and *coref* (*s*, *x*). Also, *coref* (*x*, *x*) is true for any discourse entity *x*. The *coref* relation defines equivalent classes of discourse entities: each class corresponds to a coreferential chain *equiv (x)* = { *y* | *coref* (*x*, *y*) }.

11 Adapted from *Now*, 31 October 2001.

12 *The Times*, 16 May 2000, p. 7.

13 In addition to establishing coreference between two definite noun phrases in copular relation, another interpretation would be that the definite noun phrase after the verb *to be* has a predicative, rather than a referential function. See Lyons (1977), volume 2, p. 185 for related discussion.

14 *The Express*, 15 April 2000, p. 119.

15 *The Independent*, 28 November 2000, p. 1.

16 Example from Hirschman et al. (1997).

17 My interpretation is different from that adopted in the MUC (Message Understanding Conference) coreference task (see Chapter 6) where indefinite predicate nominals are regarded as coreferential with the NP they apply to.

18 In fact, since the indefinite NP designates an entire class of entities, it cannot properly have a referent (point made by Linda C. Van Guilder).

19 *Telegraph Magazine*, 8 April 2000, p. 26.

20 'Every man has his own destiny: the only imperative is to follow it, to accept it, no matter where it leads him'. Henry Miller, *The Wisdom of the Heart*. New York: New Directions, 1941.

21 Karttunen (1969).

22 There are other examples where the anaphor does not trigger coreference such as *My neighbour has a monster Harley 1200. They are really huge but gas-efficient bikes* (Sidner 1983). To account for such cases, Sidner introduces the relationship **co-specification**. She regards the

relationship anaphor-antecedent as kind of cognitive pointing to the same 'cognitive element', called **specification**. Co-specification allows one to construct abstract representations and define relationships between them which can be studied in a computational framework.

23 **Nominal anaphora** is the type of anaphora where the anaphor is a pronoun or a non-pronominal (lexical) definite noun phrase and the antecedent is a non-pronominal noun phrase; this class of anaphora is most crucial to Natural Language Processing (see sections 1.4.1 and 1.4.2).

24 Lexical noun phrases are non-pronominal noun phrases such as definite noun phrases and proper names (see section 1.4.2).

25 We can also speak about **anaphoric chains** as opposed to coreferential chains. In the case of identity-of-reference nominal anaphora the anaphoric chain would be a coreferential chain as well; however, there may be 'pure' anaphoric chains that are not coreferential (e.g. anaphoric chains featuring verb anaphora, noun (one-) anaphora, etc.). Such classes of anaphora are considered in more detail in 1.4.3 and 1.7 below.

26 Several coreference resolution approaches will be outlined in Chapter 5.

27 Esther Freud, 'Lessons in Inhaling'; in GRANTA 43 *Best of Young British Novelists*, ed. Bill Buford, Spring 1993, p. 71. London: Granta Publications, 1993.

28 S. Paretsky, *Indemnity Only*, p. 131. London: Penguin Books, 1982.

29 More formally, if the cardinal number of the set representing the discourse entity is greater than 1, then the reference can be made by a plural anaphor.

30 Gilles Neret, *Dalí*, Ch. 2, p. 23. Germany: Benedikt Taschen Verlag, 1994.

31 Gilles Neret, *Dalí*, Ch. 1, p. 8. Germany: Benedikt Taschen Verlag, 1994.

32 Gilles Neret, *Dalí*, Ch. 2, p. 26. Germany: Benedikt Taschen Verlag,1994.

33 Gilles Neret, *Dalí*, Ch. 2, p. 23. Germany: Benedikt Taschen Verlag, 1994.

34 Gilles Neret, *Dalí*, Ch. 1, p. 6. Germany: Benedikt Taschen Verlag, 1994.

35 Deictic words are those whose interpretation is derived from specific features of the context surrounding an utterance (e.g. who is the speaker, who is the addressee, where and when the utterance takes place) and not from previously introduced words, as is the case with anaphors. For a brief outline of deixis see section 1.11.

36 John Updike, *Brazil*, p. 34. London: Penguin Books, 1994.

37 P.D. James, *Original Sin*, Ch. 8, p. 6. London: Faber and Faber, 1995.

38 John Updike, *Brazil*, p. 7. London: Penguin Books, 1994.

39 Semantically empty.

40 Quirk et al. (1985).

41 Susan Sallis, *Come Rain or Shine*, Ch. 1, p. 9. London: Transworld Publishers, 1988.

42 As opposed to indefinite pronouns such as some, every, any, etc.

43 *The Sun*, 12 January 1999.

44 A number of authors restrict lexical noun phrase anaphora to references which have the same head as their antecedents, whereas references which have different heads are regarded as forms of substitution (Halliday and Hasan 1976). Others (Coulson 1995, Grishman 1986) regard substitution with coreferential noun phrases (see the above example) as lexical noun phrase anaphora and we are taking this line too. In fact substitution includes, among other things, the phenomenon identity-of-sense anaphora (see section 1.7) and anaphora realised by non-referring expressions (such as in the case of verb anaphora). For a detailed description of substitution and the distinction between co-reference and substitution see Quirk et al. (1985).

45 It should be noted that these are only the basic relationships between the anaphoric definite NP and the antecedent but not all.

46 It should be noted, however, that the distinction between proper names and definite descriptions can often be blurred. Whereas *Roy Keane* (1.25) is a 'pure' proper name, the

same cannot be said for *Irishman Keane* or for the noun phase *the great adventurer John Smith*.

47 Sarah Jackson, *Staying Alive*, Ch. 8, p. 8. London: Chamelon Books, 1996.

48 Jack London, *White Fang*, Ch. 1, p. 36. London: Parragon Book Service, 1994.

49 Enid Blyton, *The Famous Five and the Stately Homes Gang*, Ch. 19, p. 140. London: Knight Books, 1985.

50 N-bar in the X-bar notation, see Jackendoff (1977).

51 Paulina Simons, *Eleven Hours*, p. 5. Flamingo: Great Britain, 1999.

52 *Hotline*, Autumn 1999, p. 9.

53 Note that while *did* can be regarded as substitution, it does not have a referring function.

54 *The Sunday Times*, 14 May 2000, p. 20.

55 Alex Garland, *The Beach, Prisoners of the Sun*, p. 213. Penguin, 1997.

56 *The Sunday Times*, 14 May 2000, p. 8.

57 Note that pronouns belong to the syntactical category NP.

58 This is the view expressed by Coulson (1995).

59 Note that pronominal zero anaphora overlaps with 'zero noun phrase' anaphora. Since zero pronominal anaphora is realised by a missing pronominal constituent and since pronouns replace noun phrases, one could argue that the missing pronoun could well have been a missing noun phrase. As an illustration, the second clause of example (1.36) can be reconstructed as *it is nevertheless understood* but also as *the pronoun is nevertheless understood* and even as *this pronoun is nevertheless understood*. To describe cases such as (1.36)–(1.39), the terms **zero pronominal anaphora** or **zero pronoun** have been adopted extensively in the literature due probably to the fact that the pronoun would have been the most natural overt expression.

60 M. Magorian, *Goodnight Mister Tom*, p. 13. London: Penguin, 1981.

61 Many linguists (Foley and Van Valin 1984; Hinds 1978; Tsujimura 1996) highlight the difference between zero anaphora in Japanese which is controlled by inference (pragmatically controlled zero anaphora) and zero anaphora in Latin and Slavonic languages which is controlled by agreement. Nariyama (2000), however, argues that zero anaphora in Japanese is not controlled so much by inference but more importantly by the interaction of a number of different grammatical factors such as morphological agreement, syntax constraints and discourse topic.

62 The study was based on O'Henry's story 'The Last Leaf'. Note that in this case the original English text was translated into Japanese.

63 One has to bear in mind that the Japanese texts were translations from English. In non-translated Japanese texts the frequency of overt pronouns is typically much lower (personal communication, S. Nariyama).

64 *The Daily Mail*, 4 August 1999, p. 20.

65 The following would be an alternative interpretation: ∅ also acts as a zero anaphor with antecedent *car* and since it is coreferential with the anaphor *it*, *car* is regarded as the antecedent of *it*. Note that this would be a case of identity-of-sense anaphora.

66 *The Mirror*, 17 February 1999.

67 This class of anaphora is also known as **bridging** or **associative anaphora**.

68 Bill Bryson, *Notes from a Small Island*, Ch. 10, p. 135. BCA: England, 1995.

69 *The Mirror*, 17 February 1999.

70 Or alternatively, to know that musical bands are things that break up, have critics, have members who may or may not achieve success, etc.

71 Or more precisely between the discourse entities associated with the anaphor and the antecedent.

72 As mentioned earlier, this is not an exhaustive list of the possible relationships between a definite description and its antecedent.

73 *The Daily Mail*, 9 October 2001.
74 Jerome K. Jerome, *Three Men in a Boat*, Ch. 1, p. 8. London: Penguin, 1994.
75 Note the verb anaphor *done*.
76 Note that *one* and *several* act as noun anaphors; note also the zero noun anaphor after *several* (*apples* elliptically omitted) in (1.53).
77 Adapted from Hirst (1981).
78 See also section 1.4.1.
79 Henry James, *The Spoils of Poynton*, p. 139. London: Penguin, 1987.
80 Quirk et al. (1985).
81 Quirk et al. (1985).
82 Such antecedents are also referred to as **split** antecedents in the literature.
83 D.H. Lawrence, *Sons and Lovers*, p. 377. London: Penguin, 1973.
84 Introduction to Henry James, *The Spoils of Poynton* by David Lodge, p. 1. London: Penguin, 1987.
85 I use the term 'closest antecedent' because, as I explained in 1.2, each preceding coreferential non-pronominal entity is regarded as a possible antecedent.
86 See also section 2.2.2.
87 It consists of 130 000 words and is a subcorpus of Brown's *Corpus of American English*.
88 Biber measures the distance as the number of intervening NPs between anaphor and antecedent.
89 A thorough study of the distance between demonstrative anaphors and their antecedents is presented in Botley (1999).
90 'Captured warlord's cry for help fell on deaf US ears', *The Sunday Times*, 28 October 2001.
91 Adapted from *The Mirror*, 4 March 1999.
92 This does not necessarily apply to possessive pronouns: for example, reversing the positions of the anaphora and the antecedent in (1.63) would not produce a synonymous sentence.
93 However, the rhetorical effect is different.
94 John Burnham Schwartz, *Bicycle Days*, p. 13. London: Mandarin Paperbacks, 1989.
95 Or more generally, at a lower level of syntactic structure than the antecedent.
96 William Golding, *Lord of the Flies*, p. 164. London: Faber and Faber, 1974.
97 Including demonstrative pronouns such as in the case *He told me a story like this: 'Once upon a time . . .'* (Quirk et al. 1985).
98 I argue that in (1.64) *The former White House intern* is perceived to refer to a discourse entity (person) not yet introduced and is therefore viewed as cataphoric.
99 For a comprehensive account and update see Tanaka (2000).
100 *The Times*: Times 2, 21 March 2000.
101 Carden (1982: 366).
102 Oscar Wilde, *The Picture of Dorian Gray*, Ch. 1.
103 Note the deictic use of *then*.
104 Note the deictic function of the verb *to come* as opposed to the verb *to go*.
105 Adapted from Huddleston (1984).
106 Adapted from Hirst (1981).
107 Not all ambiguous anaphors have a default such as in examples (1.74), (1.75) and (1.77).
108 For detailed accounts (but not necessarily using the same terminology) see Brown and Yule (1983), Halliday and Hasan (1976), Huddleston (1974), Quirk et al. (1985) and Lyons (1977).
109 For alternative and more comprehensive classifications see Hirst (1981) and Quirk et al. (1985). Also see Cornish's (1986) classification of anaphora based on the type of antecedent.
110 Lexical noun phrases include definite descriptions and proper names but not pronouns.

CHAPTER TWO

The process of automatic anaphora resolution

This chapter discusses the sources of knowledge needed for anaphora resolution. It introduces the different phases of the pre-processing and resolution process and explains what tools and resources are necessary. Special attention is paid to the factors that form the basis of anaphora resolution algorithms. The chapter focuses on the computational treatment of anaphora and does not cover psycholinguistic issues.

2.1 Anaphora resolution and the knowledge required

The disambiguation of anaphors is a challenging task and considerable knowledge is required to support it – from low-level morphological and lexical information, to high-level semantic and pragmatic rules.

2.1.1 Morphological and lexical knowledge

Morphological and **lexical information** is required not only for identifying anaphoric pronouns, but also as input to further syntactic processing. Some anaphors are successfully resolved solely on the basis of lexical information such as gender and number. The fact that nominal anaphors usually match (the heads of) their antecedents in gender and number is sometimes sufficient for singling out a unique NP candidate, as in example (2.1):

(2.1) *Greene* had no letters from Catherine while in Switzerland and *he* feared the silence.[1]

Following the gender and number matching rule, the noun phrase *Greene* is selected as an antecedent of the pronominal anaphor *he* because the remaining candidates *Switzerland*, *Catherine* and *letters*[2] are discounted on the basis of a gender or number mismatch.

Similarly in the sentence:

(2.2) John Bradley spoke to Jane McCarthy and to the Browns about a forthcoming project. The businessman said this enterprise would cost millions.

the lexical noun phrase anaphor *the businessman* is resolved to *John Bradley*, the latter being the only possible gender and number match. In the same way, *this enterprise* is resolved to *a forthcoming project*.

Gender agreement is a useful criterion in English when the candidates for the anaphor are (i) proper female or male names such as *Geoffrey*, *Jade*, *John Bradley*, *Victoria Griffin*, etc., (ii) nouns referring to humans such as *man*, *woman*, *father*, *mother*, *son*, *daughter*, etc., (iii) nouns representing professions such as *teacher*, *doctor*, *singer*, *actor*, *actress* which cannot be referred to by *it*,[3] (iv) gendered animals such as *cow* or *bull* or (v) words such as *country* or *ship* which can be referred to by either *she* or *it*. Similarly, number agreement helps to filter out candidates that do not carry the same number as the anaphor. It is the number of the discourse entity associated with each candidate (and anaphor in the case of definite descriptions) which is taken into account and not the number of the NP head.[4] Coordinated antecedents such as *John and Mary* are referred to by plural pronouns, whereas collective nouns such as *committee*, *army*, *team* can be referred to by both *they* and *it*. Singular noun phrases that stand for a class of people, animals or objects[5] or that can be used to represent both male and female subjects, can also be referred to by plural pronouns in English, as in the following examples:

(2.3) The jungle was so thick. *An animal* may be five yards away and quite invisible, and half of the time *they* manage to dodge past the beaters.[6]
(2.4) Ask *another Macintosh user* about the problem you're having; *they* may have a solution (Macintosh Performa guide).
(2.5) You were called on the 30th of April at 21.38 hours. *The caller* withheld *their* number (BT standard message).
(2.6) If there is *a doctor* on board, could they please make *themselves* known to the crew (British Airways flight message).[7]

In some languages the plural pronouns mark the gender (e.g. *ils*, *elles* in French, *ellos*, *ellas* in Spanish) and when a coordinated antecedent features both masculine and feminine nouns or names, it is usually referred to by the masculine form of the plural pronoun (e.g. *ils* in French, *ellos* in Spanish). The above examples and the discussion so far show that it is vital for an anaphora resolution system to have information not only about the gender and number of common nouns, but also about the gender and number of proper names.

Since the vast majority of nouns in English are neuter, the gender and number agreement rule in English is not as discriminative as in languages such as German, Bulgarian or Russian, where nouns denoting inanimate objects are routinely marked for neuter, feminine or masculine gender. However, the gender filter is of little importance to languages that do not mark gender at all, such as Turkish.

The number agreement rule can be more discriminative when selecting the antecedent for languages which, in addition to singular, distinguish between dual and plural numbers. In Arabic, for instance, there are three plural anaphoric pronouns: *homa* which refers to a dual number (a set of two elements) of both masculine and feminine nouns; *hom* which refers to a plural number (a set of more than two elements) of masculine nouns; and *honna* which refers to a plural number of feminine nouns.

2.1.2 Syntactic knowledge

The previous examples demonstrate the importance of morphological and lexical knowledge for the resolution process. In addition and more significantly, they show the importance of **syntactic knowledge**. Thus, example (2.1) shows that *Greene*, *no letters*, *Switzerland* and *Catherine* should be identified as noun phrases. Similarly, in example (2.2) the candidates for antecedent are selected from the noun phrases preceding the lexical NP anaphor *the businessman*. Therefore, it becomes clear that syntactic information about the constituents of the sentences is essential.

Syntax is indispensable in anaphora resolution. In addition to providing information about the boundaries of the sentences, clauses and other constituents (e.g. NPs, PPs), syntax plays an important role in the formulation of the different rules used in the resolution process. As an illustration, consider the simplified rule stipulating that an anaphoric NP is only coreferential with the subject NP of the same simple sentence or clause when the anaphor is reflexive (2.7).[8] This rule, which relies on syntactic information about sentence and clause boundaries, along with information about the syntactic function of each word, would rule out *Jim* as antecedent of *him* in (2.8).

(2.7) *Jim* is running the business for *himself*.
(2.8) Jim is running the business for him.

Another syntactic constraint prohibits a pronoun in a main clause from coreferring to an NP in a subsequent subordinate clause (Hirst 1981)[9]:

(2.9) Because *Amanda* had saved hard, *she* was finally able to buy the car of her dreams.
(2.10) Because *she* had saved hard, *Amanda* was finally able to buy the car of her dreams.
(2.11) *Amanda* was finally able to buy the car of her dreams, because *she* had saved hard.
(2.12) She was finally able to buy the car of her dreams, because Amanda had saved hard.

In the sentences (2.9), (2.10) and (2.11) *she* and *Amanda* are coreferential. In (2.12), however, *she* cannot be coreferential to *Amanda* because of the above constraint.

In order to be able to apply this rule, an anaphora resolution program must have access to a fairly detailed parser identifying main and subordinate clauses.

Syntactic knowledge is used extensively in anaphora resolution[10] and together with morphological and lexical knowledge it plays a key role in the process of anaphora resolution.

2.1.3 Semantic knowledge

However important morphological, lexical and syntactic knowledge are, there are many cases where they alone cannot help to resolve anaphors. In the following example:

(2.13) *The petrified kitten* refused to come down from the tree. *It* gazed beseechingly at the onlookers below.

gender or number agreement rules can eliminate neither *the petrified kitten* nor *the tree* as a potential antecedent, because both candidates are gender neutral. The **selectional restrictions** of the verb *to gaze*[11] require that its agent (the subject in an active voice sentence) be animate; **semantic information** on the animacy of *kitten* would be crucial. In a computational system such information would reside in a knowledge base such as a dictionary or ontology.

In some cases the correct interpretation of anaphors may depend on the ability of a system to undertake semantic processing in order to identify the discourse entity that is associated with the antecedent. Consider the following examples:

(2.14) Each child ate *a biscuit*. *They* were delicious.
(2.15) *Each child* ate a biscuit. *They* were delighted.

In the first example the anaphor agrees with the number of the discourse entity associated with the antecedent *biscuit* (the biscuits that the children had). This plural discourse entity can be deduced from the quantifier structure of the sentence containing the antecedent. To this end, translation into logical form is necessary.[12]

The logical form of the sentence *Each child ate a biscuit* would be:

$(\forall\ c \in \text{children})\ (\exists\ b \in \text{biscuits})\ \text{ate}\ (c, b)$[13]

and the noun phrase *a biscuit* will give rise to the discourse entity

$\{b \in \text{biscuits} \mid (\exists\ c \in \text{children})\ \text{ate}\ (c, b)\}$

Semantic knowledge as to the permissible semantic attributes of the concepts *child* and *biscuit* would also be necessary in order to identify the discourse entity $\{b \in \text{biscuits} \mid (\exists\ c \in \text{children})\ \text{ate}\ (c, b)\}$ as the antecedent of *they* in the first sentence (e.g. the children cannot be delicious) and the discourse entity $\{c \mid c \in \text{children}\}$ as the antecedent of *they* in the second sentence.

Now consider the following example[14]:

(2.16) Mary bought several shirts at the shop. They cost £20.

In order for an NLU system to 'properly' understand this example, it would not be sufficient for the system to propose *several shirts* as the antecedent of *they* but to identify the associated discourse entity which is 'set of shirts which Mary bought at the shop'. This set description cannot be derived by syntactic means and should therefore be semantically computed from the logical form of the sentence.

In this way anaphora resolution can be regarded as a process of substitution: the anaphor is replaced by a more complete semantic description to permit the interpretation of the noun phrase in the subsequent stages of semantic processing (Grishman 1986). It would make sense for this substitution to take place after semantic analysis (translation into logical form) rather than after parsing. I could even argue that if discourse, pragmatic and real-world analysis were available (see below), the substitution would be done after the last stage of analysis.

The examples given above strongly suggest that a strategy of activating an anaphora resolution algorithm after semantic analysis rather than after syntactic analysis (parsing) will produce more accurate results. The majority of anaphora resolution systems, however – especially those operating in knowledge-poorer environments – have no means of performing complex semantic and further types of analysis.[15] Therefore such systems do not attempt to compute discourse entities, but rather work with surface constituents (i.e. noun phrases) and base their resolution strategies on the output of syntactic parsing, either partial or full.[16]

Semantic knowledge is of particular importance when interpreting lexical noun phrase anaphora, especially the indirect type. A strategy typically adopted is to search for conflicts between the semantic descriptions associated with the anaphoric noun phrase and those associated with the candidate noun phrases. A contradiction arises if the heads of the noun phrases are not in a synonymy, generalisation, specialisation or set membership relation.[17] A contradiction would also arise if the modifiers of the anaphor and of the candidate NP are semantically incompatible. For instance, *the first channel* and *the second channel* would be incompatible from the point of view of their modifiers; so would *the British bank* and *the French bank*. However, *the British bank* would be compatible with the *UK bank* or simply with *the bank*. In certain circumstances *the British bank* could be compatible with *the European bank* – e.g. if *the British bank* is referred to as *the European bank* in a remote non-European country. However, one could argue that this is not a trivial matter in that *the European Bank* may be taken to denote *the Central European Bank* in Frankfurt originally set up to support monetary union in the European Community. Therefore, considerable world knowledge and inferencing might be needed to determine the degree of compatibility of the modifiers in the preceding examples.

2.1.4 Discourse knowledge

Although the morphological, lexical, syntactic and semantic criteria for antecedent selection are very strong, they are still not always sufficient to distinguish among a set of possible candidates. Moreover, they serve more as filters to eliminate unsuitable candidates than as proposers of the most likely candidate. In the case of antecedent ambiguity, it is the most salient element among the candidates for antecedent that is usually the front-runner. This most salient element is referred to in computational linguistics as the **focus** (Grosz 1977a, b; Sidner 1979) or **center**[18] (Grosz et al. 1983; Joshi and Weinstein 1981; Grosz et al. 1995) although the terminology for this can be much more diverse (Hirst 1981; Mitkov 1995a).

As an illustration, neither machines nor humans would be confident in interpreting the anaphoric pronoun *it* in the sentence:

(2.17) Tilly tried on the dress over her skirt and ripped it.

However, if this sentence were part of a **discourse segment**,[19] which would make it possible to identify the most salient element, the situation would be different:

(2.18) Tilly's mother had agreed to make her a new dress for the party. She worked hard on the dress for weeks and finally it was ready for Tilly to try on. Impatient to see what it would look like, Tilly tried on the dress over her skirt and ripped it.

In this discourse segment, *dress* is the most salient entity and is the center of attention throughout the discourse segment.

The intuition behind theories of focus or center lies in the observation that discourse is normally structured around a central topic. This topic usually remains prominent for a few sentences before the focal point shifts to a new topic. The second key intuition has to do with the fact that the center of a sentence (or clause) is typically pronominalised. This hypothesis affects the interpretation of pronouns because once the center has been established, there is often a strong tendency for subsequent pronouns to refer to this center. Example:

(2.19) Tuesday morning had been like any other. Lisa had packed her schoolbag, teased her 12-year-old brother James and bossed her seven-year-old sister Christine. After breakfast at 8.25, she walked down the stairs of the family's first floor flat and shouted: 'I'm off to school now – bye Mum, bye Dad, I will see you later.'[20]

In this example the established center *Lisa* is referred to by the subsequent pronouns *her* and *she*. It is unlikely that any reader would associate *she* in the third line to her sister *Christine*, although this is the nearest potential antecedent.

It is now clear that very often when two or more candidates 'compete' for the antecedent role, the task of resolving the anaphor can be shifted to the task of tracking down the center/focus of the sentence or clause (see also center preference, section 2.2.3.2).

2.1.5 Real-world (common-sense) knowledge

Anaphora resolution offers an ideal illustration of the complexity of natural language understanding: the reader must already have perceived the difficulties involved in resolving anaphors, but there is yet another difficulty to consider.

An anaphora resolution system supplied with extensive morphological, lexical, syntactic, semantic and discourse knowledge may still find itself helpless when confronted with examples such as:

(2.20) The soldiers shot at *the women* and *they* fell.
(2.21) *The soldiers* shot at the women and *they* missed.[21]

The resolution of the above pronominal anaphors would only be possible if further **world** (**common-sense**) **knowledge**, for example in the form of the following rules, were available.

- *Rule 1* If X shoots at Y and if Z ($Z \in \{X, Y\}$) falls, then it is more likely for Z to be Y.
- *Rule 2* If X shoots at Y and if Z ($Z \in \{X, Y\}$) misses, then it is more likely for Z to be X.

The following pronominal anaphors are no easier to deal with:

(2.22) *The council* prohibited the demonstration of the women because *they* feared violence.

(2.23) The FBI's role is to ensure *our country's freedom* and be ever watchful of those who threaten *it*.[22]

Many real-life examples of anaphors require world knowledge[23] for their resolution. While reading a British Home Office document, the following text struck me:

(2.24) If the applicant has been represented by a solicitor in connection with his application he is not empowered to administer the oath to the applicant.

In this example where the adjacent pronominal anaphors *his* and *he* are not coreferential, it is only the knowledge that an applicant cannot administer an oath to himself/herself, and that an oath is usually administered by a solicitor, that helps to resolve the anaphoric ambiguity.

Finally, applying real-world knowledge without performing additional reasoning or verifying additional conditions may lead to erroneous results. Consider (2.25):

(2.25) If Peter Mandelson had been in Tony Blair's shoes he would have demanded his resignation the day the Prime Minister forced him to leave the Cabinet.[24]

A common-sense rule would stipulate that if *X* demands *Y*'s resignation, then it is most likely that *X* and *Y* are distinct and therefore in (2.25) the anaphors *he* and *his* should not refer to the same person. In this particular case, however, the first *he* refers to Peter Mandelson acting in Tony Blair's role and *Y* to Peter Mandelson himself (acting in Peter Mandelson's role), and therefore coreference between *X* and *Y* should be regarded as perfectly normal.[25]

Incorporating extensive real-world knowledge into a practical anaphora resolution system is a very labour-intensive and time-consuming task. Consequently, the vast majority of systems simply do not have access to such extralinguistic knowledge (apart from 'toy' systems operating in very narrow domains). Therefore anaphors requiring real-world knowledge for their resolution stand the least chance of being resolved successfully.

2.2 Anaphora resolution in practice

The automatic resolution of anaphors consists of the following main stages: (1) identification of anaphors, (2) location of the candidates for antecedents and (3) selection of the antecedent from the set of candidates on the basis of anaphora resolution factors.

2.2.1 Identification of anaphors

The first step in the process of automatic anaphora resolution is the **identification of the anaphors** whose antecedents have to be tracked down. The automatic

identification of anaphoric words or phrases, at least as far as English is concerned, is not a trivial task.[26]

2.2.1.1 IDENTIFICATION OF ANAPHORIC PRONOUNS

In pronoun resolution only the anaphoric pronouns have to be processed further, therefore non-anaphoric occurrences of the pronoun *it* as in (2.26) and (2.27) have to be recognised by the program.

(2.26) It must be stated that Oskar behaved impeccably.[27]
(2.27) It was a limpid black night, hung as in a basket from a single dull star.[28]

When a pronoun *it* does not refer to anything specific, it is termed **pleonastic**.[29] Therefore, grammatical information as to whether a certain word is a third person pronoun would not be sufficient: each occurrence of *it* has to be checked in order to find out if it is referential or not.

Several algorithms for identification of pleonastic pronouns have been reported in the literature. Lappin and Leass (1994) consider an occurrence of *it* pleonastic if it appears in constructions such as the following, where *ModalAdj* denotes modal adjectives (*important*, *imperative*, *necessary*, etc.) and *CogV* denotes cognitive verbs (*think*, *believe*, *recommend*, etc.): 'It is ModalAdj that S', 'It is ModalAdj (for NP) to VP', 'It is CogV-ed that S', 'It seems/appears/means/follows (that) S', or in syntactic variants such as 'It is not/may be ModalAdj', 'Wouldn't it be ModalAdj', etc.

Denber's (1998) algorithm is a modification of Lappin and Leass's algorithm. It also operates on simple pattern recognition, but in addition to the non-anaphoric use of *it* signalled by modal adjectives and cognitive verbs, the algorithm also recognises pleonastic *it* in constructions describing weather conditions such as *It is cloudy*, *It is snowing* and in temporal constructions such as *It's three o'clock*, *It's almost time to go*.

The most detailed algorithms for identification of pleonastic pronouns, both from the point of view of description and evaluation, are those of Paice and Husk (1987) and Evans (2000, 2001). Paice and Husk's approach proposes a number of patterns based on data from the LOB corpus[30] and prior grammatical description of *it*. Unlike the approaches proposed in Lappin and Leass (1994) and Denber (1998), it applies constraints during the pattern-matching process. As an illustration, one pattern identifies *it* as non-referential if it occurs in the sequence '*it . . . that*'. This rule is prevented from over-applying by setting some constraints on the text between *it* and *that*. For instance, no more than 25 words may lie between them and there are limits on the appearance of punctuation symbols. Another constraint states that pleonastic uses of *it* are never immediately preceded by some prepositions such as *beside*, *to* and *upon*. Paice and Husk (1987) report a very high accuracy of 93.9% in classifying *it* as pleonastic or not.[31]

Evans (2000, 2001) describes an approach that identifies not only pleonastic pronouns but any non-nominal occurrences of *it*.[32] An occurrence of *it* is represented as a sequence (vector) of 35 features that classify *it* as pleonastic,

non-nominal or NP anaphoric. These features are extracted from the output of the FDG[33] tagger, and include the location of the pronoun as well as features related to the surrounding material in the text, for instance the proximity and form of NPs, adjectives, gerunds, prepositions and complementisers. The approach benefits from training data extracted from the BNC[34] and Susanne corpora consisting of approximately 3100 occurrences of *it*, 1025 of which were non-nominal, annotated for these features. The TiMBL's memory-based learning algorithm (Daelemans et al. 1999) maps each pronoun *it* into a vector of feature values, computes similarity between these and the feature values of the occurrences in the training data and classifies the pronoun accordingly. The author reports an accuracy of 78.68%, compared with of 78.71% for Paice and Husk's method over the same texts.

In other languages too, the identification of anaphoric pronouns is not always straightforward. In French, for instance, the words *le* and *la* could be both definite articles as in *J'ai lu le livre* (I read the book) and anaphoric pronouns as in *Je l'ai lu* (I read it). Therefore, some partial syntactic analysis (e.g. part-of-speech tagging) may be necessary to identify their class. Similar problems are experienced in Spanish.

In addition, even though most uses of first and second person pronouns are not anaphoric, their anaphoric use in reported speech or dialogue is not uncommon. Example (2.28) illustrates anaphoric uses of both *I* (referring to *Old Boggles*) and *you* (referring to *Dr. Rhinehart*). Simple rules for the identification of anaphoric first and second person pronouns include recognising the text as reported speech or dialogue, and gender and number matching applied to potential anaphors or antecedents.

(2.28) Old Boggles had his overcoat on now and with a toothy grimace was backing toward the door. 'Good day, Dr. Rhinehart, I hope you're better soon' he said.[35]

2.2.1.2 IDENTIFICATION OF ANAPHORIC NOUN PHRASES

The search for anaphoric noun phrases can be even more problematic. Definite noun phrases (definite descriptions) are potentially anaphoric, often referring back to preceding noun phrases, as *The Queen* does in (2.29):

(2.29) *Queen Elizabeth* attended the ceremony. *The Queen* delivered a speech.

It is important to bear in mind that not every definite noun phrase is necessarily anaphoric. In (2.30) the NP *The Duchess of York* is not anaphoric and does not refer to *the Queen*.

(2.30) The Queen attended the ceremony. The Duchess of York was there too.

Typical examples of definite noun phrases that are not anaphoric include definite descriptions that describe a specific, unique entity (as *The Duchess of York* in 2.30) or definite descriptions used in a generic way (as *the wheel* or *the piano* in (1.29) and (1.30)).

It would be equally wrong to regard all noun phrases lacking articles or demonstratives as non-anaphoric. In the genre of technical manuals or cooking instructions, where it is typical to omit definite articles, it is common to have such noun phrases referring to previously mentioned items and therefore these constructs should be regarded as potentially anaphoric.

(2.31) To oven cook *naan bread*: remove wrapper and place *bread* directly onto the oven shelf in a pre-heated oven 190°C/375°F/Gas Mark 5 for 5 minutes.[36]

Similarly to the automatic recognition of pleonastic pronouns, it is important for an anaphora resolution program to be able to identify those definite descriptions that are not anaphoric. Bean and Riloff (1999) describe a corpus-based approach for identification of non-anaphoric definite descriptions. Their algorithm generates a list of non-anaphoric noun phrases and NP patterns from a corpus and uses them to recognise non-anaphoric noun phrases in new texts. Four different heuristics support the extraction of non-anaphoric NPs. The *syntactic heuristic* looks for structural clues of 'restrictive pre-modification' such as *the U.S. president* and of 'restrictive post-modification' such as *the president of the United States* which signal non-anaphoric definite descriptions or attempts to identify referential NPs such as *the 12 men*. The *sentence one heuristic* assumes that if a definite NP occurs in the first sentence in a text, then the NP is not anaphoric. The so-called *existential head patterns* indicate that head nouns in certain NP patterns represent non-anaphoric entities when pre-modified (e.g. *the Salvadoran Government*, *the Guatemalan Government*). Finally, the *definite-only list* heuristic stipulates that some non-anaphoric NPs never appear in indefinite constructions (e.g. *the F.B.I.*, *the contrary*, etc.). When all these heuristics are employed simultaneously, Bean and Riloff's approach extracts non-anaphoric NPs with a recall of 77.7% and precision 86.6%.

Vieira and Poesio's (2000b) algorithm for identification of non-anaphoric definite descriptions draws on the work by Hawkins (1978) who identified a number of correlations between certain types of syntactic structure and discourse-new descriptions, particularly those which he called 'unfamiliar' definites.[37] The algorithm is based on syntactic and lexical features of the noun phrase which include the presence of special predicates (e.g. the occurrence of pre-modifiers such as *first* or *best* when accompanied by full relatives as in the case of *the first person to sail to America*), restrictive modification (*the inequities of the current land-ownership system*), definites that behave like proper names (*the United Kingdom*), definites that have proper nouns in their pre-modification (*the Iran–Iraq war*) and definites referring to time (*the morning*). Vieira and Poesio (2000b) report a recall of 69% and a precision of 72% in the identification of discourse-new descriptions.

Muñoz (2001) proposes a method for classifying definite descriptions as anaphoric or non-anaphoric based on the generation of a semantic network from WordNet for Spanish. For each definite description a list of possible antecedents is produced which consists of all noun phrases preceding the definite description under consideration. The noun phrases that have a head different from that

of the definite description and that are not in a semantically compatible relation with it, such as synonymy, hyperonymy or hyponymy, are declared non-anaphoric. In addition, the modifiers of the heads of the definite description and the candidates are checked for compatibility (e.g. anaphoric items cannot be in an antonymy relation). A word sense disambiguation module is used for obtaining the correct sense of the head nouns.

Finally, proper names are regarded as potentially anaphoric to preceding proper names that match in terms of first or last names (e.g. John White . . . John . . . Mr White).

2.2.1.3 TOOLS AND RESOURCES FOR THE IDENTIFICATION OF ANAPHORS

Morphological or lexical information is usually provided by a *morphological analyser*, *part-of-speech tagger* or *dictionary*. The advantage of a POS tagger is that it can disambiguate words that can be assigned more than one lexical category (e.g. *button* as a noun and *button* as a verb). However, there are a number of languages for which there are no POS taggers available (e.g. there are none for Bulgarian or Arabic at the time of writing). Therefore, programs for anaphora resolution in such languages have no choice but to use enhanced morphological analysers (e.g. Tanev and Mitkov 2000) which are often, but not always, capable of carrying out lexical disambiguation.

A *program for recognising pleonastic pronouns* or one for *identifying non-anaphoric definite descriptions* is needed to locate anaphors in English. Pleonastic recognisers based on constructs featuring modal adjectives or cognitive verbs will either need to identify these or maintain an explicit list of all such words. In addition, morphological and syntactic analysis will have to be employed for identifying the past participle of cognitive verbs or for recognising the syntactic variants of the rules listed above and therefore a *parser* will be essential. Alternatively, machine learning techniques may require large *annotated corpora*.

In French a dictionary or a morphological analyser would be unable to distinguish between *le* or *la* as articles and *le* or *la* as anaphoric pronouns. Therefore, in the case of French, a POS tagger is needed. In Spanish too, a POS tagger would be needed to distinguish between *la* definite article and *la* pronoun.

The detection of NP anaphors requires at least partial parsing in the form of *NP extraction*. A *named entity recogniser*, and in particular a *program for identifying proper names*, could be of great help at this stage. Zero anaphor identification requires more complete *parsing*, which reconstructs elliptically omitted items. As seen in examples (2.29) and (2.30), sometimes domain or world knowledge is necessary in order to distinguish anaphoric from non-anaphoric noun phrases and, therefore, *ontologies* may be useful. One such ontology is *WordNet* (see section 2.2.3.5), which has been successfully used in a number of NLP projects.

2.2.2 *Location of the candidates for antecedents*

Once the anaphors have been detected, the program has to identify the possible candidates for their antecedents. The vast majority of systems only handle

nominal anaphora since processing anaphors whose antecedents are verb phrases, clauses, sentences or sequences of sentences is a more complicated task. Typically in such systems all noun phrases preceding an anaphor within a certain **search scope** are initially regarded as candidates for antecedents.

2.2.2.1 THE SEARCH SCOPE OF CANDIDATES FOR ANTECEDENT

The search scope takes a different form depending on the processing model adopted and may vary in size depending on the type of anaphor. Since anaphoric relations often operate within or are limited to a discourse segment, the search scope is often set to the discourse segment that contains the anaphor (Kennedy and Boguraev 1996). Anaphora resolution systems which have no means of identifying the discourse segment boundaries usually set the search scope to the current and *N* preceding sentences, with *N* depending on the type of the anaphor. For pronominal anaphors, the search scope is usually limited to the current and two or three preceding sentences (Mitkov 1998b). Definite noun phrases, however, can refer further back in the text and for such anaphors the search scope is normally longer (Kameyama 1997 uses a window of 10 sentences).[38] Approaches that search the current or the linearly preceding units to locate candidates for antecedents are referred to by Cristea et al. (2000) as *linear models*. The alternative is *hierarchical models*, which consider candidates from the current or the hierarchically preceding discourse units, such as the discourse-VT model based on the Veins Theory (Cristea et al. 1998).[39] Cristea et al. (2000) show that, compared with linear models, the search scope of the discourse-VT model is smaller, making it computationally less expensive, and possibly more accurate in picking out the potential candidates. However, the automatic identification of the structural units underlying the Veins model (veins) cannot be performed with satisfactory accuracy and therefore this model remains unattractive for practical anaphora resolution developments.

Once all noun phrases in the search scope have been identified, different anaphora resolution factors are employed to track down the correct antecedent (see section 2.2.3).

2.2.2.2 TOOLS AND RESOURCES NEEDED FOR THE LOCATION OF POTENTIAL CANDIDATES

A full *parser* can be used for identifying both noun phrases and sentence boundaries. However, it is possible to make do with simpler tools, such as a *sentence splitter* to single out consecutive sentences,[40] and a *noun phrase extractor* to retrieve potential candidates for antecedents. A *tokeniser* is responsible for detecting (the boundaries of) independent tokens in the text, such as words, digits and punctuation marks. Several knowledge-poor approaches use *part-of-speech (POS) taggers*[41] and simple noun phrase grammars to identify noun phrases (Baldwin 1997; Ferrández et al. 1997; Mitkov 1996, 1998b). An *unknown word guesser*[42] would also be very helpful to tackle words that are not in the dictionary or that cannot be identified by the POS tagger, especially proper names.

Parser-free approaches operating on clauses rather than sentences (Mitkov 1998b) may ideally require a *clause splitter* to divide complex sentences into separate clauses. It should be pointed out that in practice some of the tools are incorporated in others: tokenisers are included in sentence splitters, sentence splitters are often incorporated in POS taggers, NP extractors use POS taggers and the NP extractors are part of parsers.

A point worth noting is that the identification of discourse entities requires a *semantic analyser* capable of arriving at the logical form of each sentence on the basis of its parse trees. However, this is too ambitious for most current NLP research.

Approaches that set their search scope to a discourse segment must be able to identify discourse segment boundaries. The design and implementation of a *discourse segmentation algorithm* is a difficult task. Also, algorithms for discourse segmentation (similarly to center tracking algorithms) are often based on prior information about anaphoric relations and therefore may not be usable as discourse pre-processing tools for anaphora resolution. However, discourse segmentation has been tackled by means of corpus-based, statistical methods (Hearst 1994, 1997; Crowe 1996). On the other hand, several approaches use the simple (but not always accurate) heuristics of approximating a discourse segment to a paragraph (Baldwin 1997; Mitkov 1998b).

A *proper name recogniser* plays an important role for identifying proper name candidates. The task of recognising proper names itself is a rather challenging one. Lexical databases consisting of thousands of proper names have been automatically constructed and used (Muñoz et al. 1998) to address this problem.[43] A dictionary of proper names may be a starting point but nouns which can be both proper names and common names pose a problem, as do proper names which are not in the dictionary. The disambiguation should normally be carried out by a POS tagger but, as will be seen later, this task is far from trivial and what is in fact needed is a task-oriented proper name recogniser.[44]

There are additional difficulties related to the processing of proper names. For instance, there is an overlap between girls' names and flower names (daisy, heather, ivy, rose, etc.). Also, artistic names (pseudonyms) can be anything such as Frank Zappa's daughter *Moon Unit* or the artist formerly known as *Prince* (Denber 1998). One must also take into account the fact that some proper names can be ambiguous in gender (*Chris*, *Lesley* or *Robin* in English, *Claude* in French). In addition, some names can differ in gender across languages, such as *Jean* which is a female name in English but a male name in French. In general, the occurrence of foreign names in a text could make things more complicated. There is a further ambiguity between the names of persons and other proper nouns such as place names, names of organisations, names of products or even names of months. For example, *Troy* can be both a boy's first name and a city (in fact one of several different cities); *June* is both a female name and a month of the year. Finally, the number of proper names is open-ended: it can be argued that any combination of letters, pronounceable or not, is a potential proper name.

Proper names are definite noun phrases which can be simple and can contain only one name (*Tony*) or a sequence of names and titles (*The Right Honourable*

Tony Blair MP). For more complex constructions a proper name grammar might be helpful.[45] Such a grammar should be able to recognise *George Washington* as an animate, masculine 'complex' proper name, whereas *George Washington Bridge* should be recognised as an inanimate, neuter name. The selectional restrictions associated with proper names represent an additional problem. Names of state capitals such as *Washington*, *Moscow*, *London*, etc., can act as human agents when standing for governments of countries.

The identification of proper names has attracted considerable attention over the last few years; it has also featured as a separate task (Named Entity Recognition task) at the Message Understanding Conferences.[46] For a more detailed discussion see Grishman (2002).

2.2.3 *The resolution algorithm: factors in anaphora resolution*

Once the anaphors have been detected, the program will attempt to resolve them by selecting their antecedents from the identified sets of candidates. The resolution rules based on the different sources of knowledge and used in the resolution process (as part of the anaphora resolution algorithm), are usually referred to as *anaphora resolution factors*. Factors frequently used in the resolution process include gender and number agreement, c-command constraints (see Chapter 3, section 3.2), semantic (selectional) restrictions,[47] syntactic parallelism, semantic parallelism, salience, proximity, etc. These factors can be 'eliminating', i.e. discarding certain noun phrases from the set of possible candidates, such as in the case of gender and number constraints, c-command constraints and selectional restrictions. The factors can also be 'preferential', giving more preference to certain candidates over others, such as salience (center of attention), parallelism or proximity. The computational linguistics literature uses diverse terminology for these factors. For example, whereas Rich and LuperFoy (1988) refer to the 'eliminating' factors as *constraints*, and to the preferential ones as *proposers*, Carbonell and Brown (1988) use the terms *constraints* and *preferences*. Other authors (e.g. Mitkov 1997a) argue that all factors should be regarded as preferential, giving higher preference to more restrictive factors and lower preference to less 'absolute' ones, calling them simply *factors* (Preuß et al. 1994), *attributes* (Pérez 1994), *symptoms* (Mitkov 1995b) or *indicators* (Mitkov 1996, 1998b).

2.2.3.1 CONSTRAINTS

Constraints are considered to be obligatory conditions that are imposed on the relation between the anaphor and its antecedent. Therefore, their strength lies in discounting candidates that do not satisfy these conditions; unlike preferences, they do not propose any candidates.

Gender and number agreement

This constraint requires that anaphors and their antecedents must agree in number and gender.[48] For example:

(2.32) As it emerged that *Jo Moore* had also tried to launch a 'dirty tricks' campaign against London transport supremo Bob Kiley, Downing Street pointedly refused to support *her*.[49]

In nominal anaphora this agreement usually occurs at the level of NP heads, but in the case of complex noun phrases that contain noun phrases as constituents, reference can also be made to a noun phrase that is not the head of the complex noun phrase. In complex possessive noun phrases, for instance, the noun phrase that represents the possessor, and whose possessive form acts as modifier to the head of the whole construction, can equally be referred to:

(2.33) *Arsene Wenger's* human rights campaign took a dramatic turn yesterday when *he* told the Football Association that it can shut *him* up only by throwing *him* into jail.[50]

In the above example the head of the complex possessive noun phrase *Arsene Wenger's human rights campaign* is *campaign* but the antecedent is the noun phrase *Arsene Wenger*.

C-command constraints

In intrasentential anaphora resolution, constraints imposed by the **c-command** relation[51] play an important role in discounting impossible candidates for antecedents of anaphors that are not reflexive pronouns and in selecting antecedents of reflexive anaphors.[52] As an illustration, consider the application of the c-command constraint that a non-pronominal NP cannot corefer with an NP that c-commands it to the example

(2.34) She almost wanted Hera to know about the affair.[53]

In this example *she* c-commands *Hera* and therefore, coreference between *she* and *Hera* is impossible.

The notions of c-command and local domain constraints are discussed in greater detail in Chapter 3, section 3.2. Such types of constraints are often referred to in the literature as **configurational** constraints (Carter 1987a).

Selectional restrictions

This constraint stipulates that the **selectional** (**semantic**) **restrictions** that apply to the anaphor should apply to the antecedent as well. Therefore in (2.35) the antecedent should be an object which can be disconnected (the computer, but not the disk), whereas in (2.36) the antecedent should be an object which can be copied (the disk, but not the computer).

(2.35) George removed the disk from *the computer* and then disconnected *it*.
(2.36) George removed *the disk* from the computer and then copied *it*.

In section 2.2.3.4 below it will be argued that selectional restrictions, as other constraints, should not be regarded as absolute conditions.

2.2.3.2 PREFERENCES

Preferences, unlike constraints, are not obligatory conditions[54] and therefore do not always hold. For instance, there is a general (but weak) preference for *the most recent NP* matching the anaphor in gender and number to be the antecedent as in example (2.37), but this is not always the case as shown by (2.38).

(2.37) Most weekend newspapers these days contain *colour supplements full of rubbish*. It's a waste of time reading *them*.[55]

(2.38) *Most weekend newspapers* these days are full of advertisements. It's a waste of time reading *them*.[56]

Other examples include the preference for *candidates in the main clause* over those in the subordinate clause, preference for *NPs which are positioned higher in the parse tree* over those that have a lower position[57] and preference for *candidates in non-adjunct phrases* over those in adjunct phrases. In some cases these preferences may be strong enough to interfere with the expected logical interpretation. For example:

(2.39) Jack drank the wine on the table. It was brown and round.[58]

Even though semantic constraints clearly suggest that only the table can be *brown and round*, some people would still find it difficult to assign *the table* as the antecedent of *it* (and thus perceive the text as odd) since *it* appears to refer to *the wine* given the preference for entities in non-adjunct phrases.[59]

Two more types of preference will be illustrated: syntactic parallelism and center of attention.

Syntactic parallelism

Syntactic parallelism can be helpful when other constraints or preferences are not in a position to propose an unambiguous antecedent. This preference is given to noun phrases that have the *same syntactic function* as the anaphor.

(2.40) The programmer successfully combined *Prolog* with C, but he had combined *it* with Pascal last time.

(2.41) The programmer successfully combined Prolog with *C*, but he had combined Pascal with *it* last time.

Syntactic parallelism is a preference and not a constraint as it is relatively easy to find an example that does not follow this preference:

(2.42) *The program* successfully combined Prolog with C, but Jack wanted to improve *it* further.

In this example the anaphor *it* and its antecedent *the program* have different syntactic functions, whereas *it* and *Prolog* have the same syntactic function (direct object). Example (2.35) is another illustration that syntactic parallelism is a preference and not a constraint.

Center preference

In a coherent discourse it is the *most salient* and central element in a current clause or sentence that is likely to be pronominalised in a subsequent clause or sentence. The **center preference** is very strong in pronoun resolution, and it would not be inaccurate to say that in most cases it is the center of the previous clause or sentence[60] which is the antecedent of a pronominal anaphor.[61] In (2.18) for instance, there are two syntactically and semantically acceptable candidate antecedents (*dress* and *skirt*) for the pronoun *it*, but the antecedent is *skirt*, being the center of the previous clause (as Tilly tried on the dress over her skirt).

The center is still a matter of preference, however, so there are cases in which the anaphor does not refer to the center of the previous clause/sentence. As an illustration, consider the following example:

(2.43) It was Oliver who persuaded *Joan* to borrow the car. *She* was unaware of the repercussions that later followed.

In this example *Oliver* is the center of the first sentence and, therefore, one would expect it to be pronominalised in the subsequent sentence. However, the anaphor in the second sentence must refer to *Joan* because of gender constraints.[62]

The center of the previous clause (sentence) is the most likely antecedent of an anaphor under consideration. This explains why the following English sentences sound odd and humorous: it takes longer for the reader to process the actual meaning given that, contrary to the 'natural' expectation of the hearer/reader, the centering preference has not been observed.

(2.44) If the baby does not thrive on raw milk, boil it.
(2.45) If an incendiary bomb drops near you, don't loose your head. Put it in a bucket and cover it with sand.[63]

In (2.44) the noun phrase *the baby* is a prime candidate for pronominalisation in the following clause, being more salient than the noun phrase *raw milk*. However, it is the less salient noun phrase (*raw milk*) that is pronominalised in the following clause. In (2.45) *your head* is the center of the clause prior to the anaphor *it*, but the reference is to *incendiary bomb*. In both (2.44) and (2.45) the preference for the most salient candidate is overridden by common-sense constraints.

Subject preference

Some anaphora resolution approaches give preference to the candidate that is the **subject** of the sentence. This preference sometimes overlaps with center preference since in English the subjects are the favoured sentence centers. For example:

(2.46) *The customer* lost patience and called the waiter. *He* ordered two 12-inch pizzas.

However, subject preference is not strong enough and can be easily overruled by common-sense constraints or preferences:

(2.47) The customer lost patience and called *the waiter*. *He* apologised, and said *he* had been delayed by other orders.

Algorithms that have no information about the syntactic functions of the words may give preference to the first noun phrase in non-imperative sentences, thus approximating it to the subject in subject-first languages like English (Mitkov 1998b).

Chapter 3 will discuss more centering preferences (3.1) and syntactic constraints (3.2). Also, constraints and preferences will be discussed in Chapters 4 and 5 where different approaches to anaphora resolution will be outlined.

2.2.3.3 EXAMPLE OF ANAPHORA RESOLUTION BASED ON SIMPLE FACTORS

As an illustration, consider a simple model using the gender and number agreement constraint, the c-command constraint that a non-pronominal NP cannot corefer with an NP that c-commands it, and the center preference. First the constraints are applied and if the antecedent still cannot be determined, the center preference is activated. It is assumed that analysis has taken place and that all the necessary information about the morphological features of each word, the syntactic structure of the sentences and the center of each clause is available and that all anaphors have been identified. Consider the application of this model to the following text.

(2.48) How poignant that one of the television tributes paid to Jill Dando shows her interviewing people just before the funeral of Diana Princess of Wales. Some of the words she used to describe the late princess could equally have applied to her.[64]

This discourse segment features four anaphors: *her* (first sentence), *she*, *the late princess* and *her* (second sentence). The resolution takes place from left to right. Initially all noun phrases preceding the first anaphor *her* are considered potential candidates for antecedents: *one of the television tributes*, *the television tributes* and *Jill Dando*. The number agreement constraint discounts *the television tributes*, whereas gender agreement rejects *one of the television tributes* proposing *Jill Dando* unambiguously as the antecedent of *her*. Next, the anaphor *she* has to be interpreted. The initial candidates are again all preceding NPs: *one of the television tributes*, *the television tributes*, *Jill Dando*, *people*, *the funeral of Diana Princess of Wales*, *the funeral*, *Diana Princess of Wales*, *some of the words*, *the words*, but the gender and number filter eliminate all candidates but *Jill Dando* and *Diana Princess of Wales*. Now center preference is taken into account, proposing the center of the preceding clause *Jill Dando* as the antecedent. Due to gender and number mismatch, the anaphor *the late princess* can be resolved only to *Jill Dando* or *Diana Princess of Wales*.[65] Next, the c-command constraint is activated. Since *she* has been already instantiated to *Jill Dando*, and since *she* c-commands *the late princess*, coreference between *Jill Dando* and *the late princess* is impossible. Therefore, *Diana Princess of*

Wales is the antecedent of *the late princess*. Finally, the anaphor *her* in the second sentence has to be resolved between *the late princess/Diana Princess of Wales* and *her/Jill Dando*. The center of the clause prior to the one containing the anaphor is *she* (*Jill Dando*), therefore *Jill Dando* is the preferred antecedent.[66]

2.2.3.4 COMBINATION AND INTERACTION OF CONSTRAINTS AND PREFERENCES

Usually constraints and preferences work in combination towards the goal of identifying the antecedent. Applying a specific constraint or preference alone may not result in the tracking down of the antecedent.

It should also be noted that constraints and preferences usually do not act independently but interact with other factors. This interaction could make a specific constraint or preference look stronger or weaker. Consider again the earlier examples:

(2.35) George removed the disk from *the computer* and then disconnected *it*.

(2.36) George removed *the disk* from the computer and then copied *it*.

The semantic restriction in (2.36) favours *the disk* as an antecedent of *it* and the decision is enhanced by the syntactic parallelism preference which would single out *the disk* as well. In addition, the chances of the NP *the computer* being picked as an antecedent are weakened by the fact that it is an indirect object as opposed to the NP *the disk*, which is a direct object.[67]

On the other hand in (2.35) both the syntactic parallelism and the direct object preference work against the NP *the computer*, yet they cannot override the selectional (semantic) restriction. It is this interaction between constraints and preferences that suggests that perhaps *the computer* in (2.35) is not so much of an unambiguous antecedent as *the disk* in (2.36). And yet it is worth pointing out that even in (2.36) there is not an absolute restriction on 'copying computers', which can be seen from the following example:[68]

(2.49) The Chinese have been copying American computers and producing them at less than a quarter of the cost.

The examples above suggest that the borderline between constraints and preferences is sufficiently blurred as to encourage a growing number of authors to regard all factors as preferences rather than as absolute constraints (Mitkov 1995b, 1997a). I believe that treating certain factors in an 'absolute' way may often be too risky. Consider the number agreement constraint for English. Unless an exhaustive list of rules or exceptions describing when singular nouns can be referred to by plural anaphors is available, discounting candidates on the basis of number agreement could increase a system's error rate. This is particularly important for algorithms that do not include semantic analysis and that are not able to generate a correct logical form, since the grammatical number of the anaphor matches the number of the discourse entity and not that of the NP associated with it. A preference-based system, on the other hand, takes as its starting

point the equal consideration of all the candidates and in turn considers all cases of preference, and typically assigns a numerical score for each NP candidate.

The previous examples demonstrate that real-world (common-sense) knowledge appears to be an especially privileged factor that can override others. In fact, this seems to be the factor that human readers use to judge what the antecedent 'really' is, and whether other factors lead to erroneous results.

The impact of different factors and/or their coordination have also been investigated by Carter (1990). He argues that a flexible control structure based on numerical scores assigned to preferences allows greater cooperation between factors as opposed to a more limited depth-first architecture. His discussion is grounded in comparisons between two different implemented systems – SPAR (Carter 1987a, 1987b) and the SRI Core Language Engine (Alshawi 1992).

In addition to the impact of each factor on the resolution process, some factors may have an impact on other independent factors. An issue which needs further attention is the '(mutual) dependence' of factors. Dependence/mutual dependence of factors is defined in the following way (Mitkov 1997a). Given the factors x and y, y is taken to be *dependent* on factor x to the extent that the presence of x implies y. Two factors will be termed mutually dependent if each depends on the other.[69]

The phenomenon of (mutual) dependence has not yet been fully investigated, but I believe that it can play an important role in the process of anaphora resolution, especially in algorithms based on the ranking of preferences. Information on the degree of dependence would be especially useful in a comprehensive probabilistic model and is expected to lead to more precise results.

More research is needed to give precise answers to questions such as: 'Do factors hold good for all genres?' (i.e. Which factors are genre specific and which are language general?) and 'Do factors hold good for all languages?' (i.e. Which factors seem to be multilingual and which are restricted to a specific language only?). One tenable position is that factors have general applicability to languages, but that languages will differ in the relative importance of factors, and therefore on their relative weights in the optimal resolution algorithm.[70] For some discussion on these topics, see Mitkov (1997a) and Mitkov et al. (1998).

Finally, while a number of approaches use a similar set of factors, the 'computational strategies' for the application of these factors may differ. The term 'computational strategy' refers here to the way factors are employed, i.e. the formulae for their application, interaction, weights, etc. Consider a system where candidates are assigned scores with the application of each preference and the candidate with the highest composite score is proposed as the antecedent. The composite score may be a simple adding of the scores associated with each factor (Mitkov 1998b) or a 'normalised' score obtained by dividing the composite score by a confidence value (Rich and LuperFoy 1988). The score may also be calculated on the basis of more sophisticated techniques such as uncertainty reasoning (Mitkov 1995b). I showed (Mitkov 1997a) that it is not only the optimal selection of factors that matters but also the optimal choice of computational strategy. Another important factor concerning the choice of a computational strategy for preference-based approaches is the optimisation of the score of each factor (see more on optimisation in Chapter 7).

2.2.3.5 TOOLS AND RESOURCES NEEDED FOR IMPLEMENTING ANAPHORA RESOLUTION FACTORS

The factors employed by anaphora resolution algorithms are based on rules which rely on different types of knowledge, so different tools and resources may be needed to enable their operation.

The gender and number filters require information on the gender and number of the anaphor and its candidates. Therefore, *dictionaries*, *morphological analysers* or *part-of-speech taggers*[71] are needed but they are far from sufficient. As mentioned earlier (section 2.1.1), English is not so gender discriminate as other languages but in addition to the vast majority of neuter words, a number of nouns are feminine or masculine or both feminine and masculine, and failing to identify the gender of such words can easily lead to errors in the interpretation of anaphors. The gender of proper names can be another tough problem (see 2.2.2.2). A program identifying animate entities could provide essential support in employing the gender constraints. Denber (1998) and Cardie and Wagstaff (1999) use WordNet to recognise animacy. Evans and Orasan (2000) propose a method combining the FDG shallow parser, WordNet, a first name gazetteer and a small set of heuristic rules to identify animate entities in English texts. Their study features extensive evaluation and provides empirical evidence that in supporting the application of agreement constraints, animate entity recognition contributes to better performance in anaphora resolution.[72] Automatic identification of gender has been addressed by Orasan and Evans (2001) in a method involving the use of WordNet and machine learning techniques, and by Ge et al. (1998) in a method involving unsupervised learning of gender information.

The c-command constraints require access to the tree structure of the sentence and therefore a *full parser* is needed to capture these factors. A parser would also be helpful for the implementation of the syntactic parallelism preference. Part-of-speech taggers or *shallow parsers* might be sufficient for implementing this preference as many of them (e.g. Lingsoft's ENGCG, FDG, Xerox shallow parser) mark the syntactic function (subject, object, etc.) of most words.

Semantic knowledge can be provided by *WordNet*, an ontology which is widely used by researchers in NLP. For instance, from WordNet a number of semantic relations (between words) such as synonymy, antonymy, hypernymy ('is-a', 'is-a-kind-of'), hyponymy ('subsumes'), meronymy (part–whole relation) and familiarity (rare/uncommon/common) can be retrieved.[73] Also, some semantic information can be obtained from verb selectional restrictions if supplied in dictionary entries. On the other hand, word sense disambiguation (e.g. to distinguish different senses such as *bank* (river bank) and *bank* (financial institution)) may be necessary before applying selectional restrictions. Certain approaches employing further semantic constraints or preferences (e.g. semantic parallelism) may need *deeper semantic analysis* (e.g. performed by case grammars).

A *center* or *focus tracking program* is needed for approaches employing center preference. Some approaches use a simplified centering model and approximate the center of a sentence to its subject. Subject identification can be performed by a full or shallow parser.

2.3 Summary

This chapter has shown that anaphora resolution is a complex task which requires different forms of knowledge and which can be regarded as a three-stage process: identification of anaphors, location of candidates for antecedents and selection of antecedent. The last stage is performed through a resolution algorithm based on the interaction of various factors. Some of these factors, termed constraints, appear to be more restrictive in discounting improbable candidates, whereas others, called preferences, impose fewer restrictions and only point to a preferred antecedent. The chapter has outlined the tools and resources needed for each of these stages in anaphora resolution.

Notes

1 Norman Sherry, *The Life of Graham Greene*, Vol. 2, p. 264. Penguin Books.
2 Note that we are focusing on nominal anaphora and that only NPs preceding the anaphor are regarded as candidates for antecedents.
3 Note that *teacher*, *doctor* or *singer* can be referred to by both *he* and *she*.
4 See example (1.17), Chapter 1.
5 See also Sidner's example in note 21, section 1.2, Chapter 1.
6 G. Orwell, *The Complete Novels (Burmese Days)*, p. 171. Penguin, 2000.
7 These are not the only examples of gender or number mismatch between the anaphor and the antecedent. For instance, there are cases of indirect anaphora where a singular anaphor can point to a plural antecedent as in 'In the newsagents there were only *two newspapers* left. *One* was a right-wing tabloid.' In addition, cases of indirect anaphora can be encountered where the anaphor and the antecedent may differ in gender as in '*The car* was going nearly eighty miles an hour. *He* did not see the curve in time' (Smith 1991). It becomes clear that gender or number agreement should not be regarded as absolute constraints. For further discussion the reader is referred to Barlow (1998).
8 This is an approximation of a more general rule stated in Chapter 3. Note that whereas this rule will work in most of the cases, it would not be helpful in examples such as 'Jenny feared the man next to her'.
9 This rule is an approximation of a more general rule which is to be stated in the section on Binding Theory in Chapter 3.
10 See c-command constraints and syntactic parallelism in section 2.2.3; see also Lappin and Leass's approach (Chapter 5, section 5.3.1) which employs various syntactic constraints.
11 Note that the morphological analysis will have to identify *gazed* as the past tense of the verb *to gaze*.
12 For more on logical form, see Grishman (1986), Chapter 3, section 3.2.
13 Equivalent also to 'Every child ate some biscuit'.
14 Adapted from Grishman (1986).
15 Substantial semantic analysis is especially unrealistic for systems that process unrestricted texts.
16 See for instance Lappin and Leass (1994), Kennedy and Boguraev (1996), Baldwin (1997), Kameyama (1997), Mitkov (1996; 1998b) or Chapter 5 of this book. The benefits of shallow analysis for practical applications such as information retrieval and question answering have also been noted in Vicedo and Ferrández (2000).

17 For examples of these relations see 1.4.2 and 1.6. Information about such relations can be automatically derived from an ontology or lexicon.
18 Center and focus are close, but not identical concepts. The reader is referred to Walker et al. (1998) or Grosz et al. (1995); centering theory is discussed in greater detail in Chapter 3.
19 Discourse segments are stretches of discourse in which the sentences are addressing the same topic (Allen 1995).
20 *The Guardian*, 23 January 1999.
21 Hutchins and Somers (1992).
22 Hirst (1981).
23 Often the distinction between 'semantic' and 'real-world' knowledge is unclear. I assume that semantic knowledge is limited to cases where simple semantic attributes, such as animacy, can help disambiguate a specific anaphor (e.g. 'The monkeys ate the bananas because they were hungry' as opposed to 'The monkeys ate the bananas because they were ripe'). Real-world knowledge, on the other hand, is based on real-world norms, common sense and inference rules such as the rules following examples (2.20) and (2.21).
24 *The Independent*, 10 March 2001, p. 6.
25 In fact, an extended rule could be formulated as follows: If *X* demands *Y*'s resignation, *X* and *Y* should be distinct unless *X* acts in *Z*'s role, $Z \neq Y$.
26 As stated in Chapter 1, this book focuses on the most widespread and central class of anaphora to NLP applications – that of nominal anaphora.
27 Thomas Keneally, *Schindler's List*, p. 165. London: BCA, 1994.
28 F. Scott Fitzgerald, *Tender is the Night*, p. 49. London: Penguin, 1986.
29 See section 1.4.1 for more discussion on pleonastic pronouns.
30 LOB stands for Lancaster–Oslo–Bergen.
31 However, on a different text a re-implemented version of this algorithm by R. Evans was evaluated to perform with an accuracy of 78.71% (Evans 2000).
32 These include instances of *it* whose antecedents are constituents other than noun phrases such as verb phrases, sentences, etc.
33 FDG stands for Functional Dependency Grammar.
34 British National Corpus.
35 Luke Rhinehart, *The Dice Man*, Ch. 5, p. 49. London: Harper Collins, 1972.
36 Preparation guidelines, Tesco's *Chicken Balti Rogan Josh with Naan*, 1999.
37 Definite descriptions whose existence cannot be expected to be known on the basis of generally shared knowledge.
38 See also the discussion in Chapter 1, section 1.9.
39 The Veins Theory (VT) extends and formalises the relation between discourse structure and reference as proposed by Fox (1987). It identifies 'veins' – chains of elementary discourse units over discourse structure trees that are built in compliance with the Rhetorical Structure Theory (Mann and Thompson 1988).
40 Periods cannot serve as reliable sentence boundaries since in a number of cases (e.g. abbreviations) a period does not signal the end of a sentence.
41 A number of POS taggers such as Brill's tagger, Lancaster's CLAWS4, Lingsoft's ENGCG, Connexor's ENGCG-2, Itpos, etc., perform sentence splitting too.
42 An unknown word guesser is a program which predicts the lexical class of a word if it cannot be accounted for by the POS tagger. Many POS taggers incorporate unknown guessers.
43 Muñoz et al. (1998) used a dictionary of 4337 first names and 4657 second names retrieved from a university student registry as well as a place name dictionary with 53 000 entries provided by the post office.
44 For languages using an alphabetic writing system, a rule stating that all proper names begin with a capital first letter is far from sufficient in cases where the name is the first word

in a sentence and therefore begins with a capital letter too. In English, other words such as some adjectives and common nouns also conventionally begin with a capital (e.g. French, Frenchman). In German, all nouns are spelt with a capital. Other problems arise with words in headings and words entirely in upper case.

45 Since the development of proper name grammar is usually a time-consuming job, alternative techniques such as machine learning have been recently explored.

46 The MUCs are U.S. Government-sponsored evaluations which rank the performance of Information Extraction systems according to the following tasks: Name Entity, Coreference, Template Element, Template Relation and Scenario Template.

47 The terms *restriction* and *constraint* are used interchangeably in this chapter.

48 As pointed out in 2.1.1, certain collective nouns in English do not necessarily agree in number with their antecedents and should be exempted from the agreement test. For instance, *government*, *team*, *army*, *parliament*, etc., can be referred to by *they*; equally some plural nouns such as *data* and *media* can be referred to by *it*. In English antecedents usually agree with the anaphors in gender, whereas this is not always the case in other languages such as German where *sie* (she, female) can agree with *Mädchen* (girl, neuter). Barlow (1998) shows that such gender and number mismatches are not uncommon across languages.

49 *The Daily Mail*, p. 1, 12 October 2001.

50 *The Daily Mail*, 9 January 1999.

51 A node *A* c-commands a node *B* if and only if (i) *A* does not dominate *B*, (ii) *B* does not dominate *A*, (iii) the first branching node dominating *A* also dominates *B* (Haegeman 1994). See section 3.2 for more details on Binding Theory.

52 Within the so-called local domain which for our purposes here can be broadly defined as a finite clause or a complex possessive construction (see sections 3.2.1 and 3.2.2).

53 Victoria Griffin, *The Mistress*, Ch. 6, p. 73. Bloomsbury: London, 1999.

54 It will be seen on the basis of a number of examples to follow that constraints can hardly be absolute: almost always there will be exceptions.

55 Example suggested by Geoffrey Leech.

56 Example suggested by Geoffrey Leech.

57 This preference can be regarded as fairly general since it often covers subject preference, the preference for candidates in the main clause, the preference for candidates in non-adjunct phrases and gives preferential status to NP in cleft constructions as example (2.43).

58 This sentence is an adaptation by Allen (1995) of the original sentence proposed by Wilks (1975b).

59 Another way of explaining this anomaly is that 'the wine' is the most salient NP of the first sentence and is a prime candidate for pronominalisation in a subsequent sentence (see also the examples related to centering).

60 Or more accurately previous 'utterance' (see Chapter 3) which for practical reasons is taken to be a clause (in complex sentences) or a sentence.

61 Provided there are no other anaphors in the sentence. See also Rule 1 in centering (Chapter 3, section 3.1).

62 In effect, the speaker is moving on to a new topic here.

63 This sentence is of an 'obscure' origin but is believed to be from a British Second World War anti-raid leaflet.

64 *The Mirror*, 30 April 1999, p. 4.

65 Note that this model does not use any semantic knowledge or inferencing which could help find that *the late princess* refers to *Diana Princess of Wales* on the basis that the previous sentence reports her funeral; also, this model does not use any matching rule suggesting (not always correctly, however) that NPs with identical heads are coreferential and therefore cannot establish a coreferential link between *the late princess* and *Diana Princess of Wales*.

66 See Chapter 3 for more on (rules in) centering.

67 NPs which are indirect objects are usually less salient than NPs which are direct objects (see also Lappin and Leass 1994, Mitkov 1995b and Mitkov 1998b); the centering theory also prefers direct object to indirect object.

68 Point made and example suggested by G. Leech (personal communication).

69 In order to clarify the notion of (mutual) dependence, it would be helpful to view the factors as 'symptoms' or 'indicators' observed to be 'present' or 'absent' with the candidate in a certain discourse situation. For instance, if *gender agreement* holds between a candidate for an anaphor and the anaphor itself, I shall say that the symptom or indicator *gender agreement* is present with this candidate. Similarly, if the candidate is in a subject position, I shall say that the symptom *subjecthood* is present. As an illustration consider the example 'Mary invited John to the party. He was delighted to accept.' In this discourse the symptoms subjecthood, number agreement and entities in non-adjunct phrases are present (among others) with the candidate *Mary*; the symptoms gender agreement, number agreement and entities in non-adjunct phrases are observed with the candidate *John*; and finally number agreement and recency are present with the candidate *the party*.

70 If a specific factor is not applicable to a language, then its importance or weight for this language will be 0.

71 Note that POS taggers for languages which mark gender such as French, Spanish and German usually return gender information.

72 The experiment was carried out on the pronoun resolution system MARS when applied to reports from Amnesty International that had a political register. The system's success rate was increased by 5% when animacy recognition was used to support gender agreement between pronouns and competing candidates. MARS is outlined in Chapter 7, section 7.4.

73 The original version of WordNet is for English (Fellbaum 1998) but the recent project EuroWordNet produced French, Spanish, German, Italian, Dutch, Czech and Estonian versions of the ontology. Whereas WordNet was developed originally for lexicographers, some of the Euro WordNet design principles are more directed towards NLP.

CHAPTER THREE

Theories and formalisms used in anaphora resolution

This chapter outlines some of the theories and formalisms that have been successfully used in anaphora resolution. Centering theory and binding theory are introduced in order to demonstrate how relevant rules and constraints may be applied to the interpretation of anaphors. Finally, other theories such as focusing and the Discourse Representation Theory (DRT) are briefly sketched too.

3.1 Centering

Centering is a theory about discourse coherence and is based on the idea that each **utterance** features a topically most prominent entity called the **center**. Centering regards utterances[1] which continue the topic of preceding utterances as more coherent than utterances which feature (or flag up an impending) topic shift.

The main idea of centering theory (Grosz et al. 1983; Grosz et al. 1995) is that certain entities mentioned in an utterance are more central than others and this imposes certain constraints on the use of referring expressions and in particular on the use of pronouns. It is argued that the coherence of a discourse depends on the extent to which the choice of the referring expressions conforms to the centering properties.

As an illustration, consider the following examples:

Discourse A
(3.1) John works at Barclays Bank.
(3.2) He works with Lisa.
(3.3) John is going to marry Lisa.
(3.4) He is looking forward to the wedding.

Discourse B
(3.1) John works at Barclays Bank.
(3.2) He works with Lisa.
(3.3) John is going to marry Lisa.
(3.5) She is looking forward to the wedding.

Centering predicts that Discourse B is less coherent than Discourse A. In both examples the discourse entity realised by John is the center in utterances (3.2) and (3.3),[2] but while in (3.4) the center remains the same, utterance (3.5) shifts the center to the discourse entity realised by Lisa. The shift in center and the use of a pronominal form to realise the new center contribute to making B less coherent than A: in fact, in utterance (3.4), unlike (3.5), it is the center of utterances (3.2) and (3.3) which has been pronominalised.

Discourses consist of continuous discourse segments. A **discourse segment** D consists of a sequence of utterances $U_1, U_2, \ldots U_N$. Each utterance U in D is assigned a set of potential next centers known as **forward-looking centers** Cf(U, D)[3] which correspond to the discourse entities evoked by the utterance. Each utterance (other than the first) in a segment is assigned a single center defined in the centering theory as the **backward-looking center**[4] Cb(U). The backward-looking center Cb(U) is a member of the set Cf(U) and is the discourse entity the utterance U is about. The Cb entity connects the current utterance to the previous discourse: it focuses on an entity that has already been introduced. A central claim of centering is that each utterance has exactly one backward-looking center.[5]

The set of forward-looking centers Cf(U) is partially ordered according to their discourse salience. The highest-ranked element in Cf(U) is called *the preferred center* Cp(U) (Brennan et al. 1987). The preferred center in a current utterance U_N (denoted as Cp(U_N)) is the most likely backward-looking center of the following utterance (denoted as Cb(U_{N+1})). Discourse entities in subject position are preferred over those in object position, which are preferred over discourse entities in subordinate clauses or those performing other grammatical functions.[6]

Grosz et al. (1995) define three types of transition relations across pairs of utterances.

1. Center **continuation**: Cb(U_{N+1}) = Cb(U_N), i.e. the backward-looking center of the utterance U_{N+1} is the same as the backward-looking center in the utterance U_N and this entity is the preferred center of Cf(U_{N+1}). In this case Cb(U_{N+1}) is the most likely candidate for Cb(U_{N+2}).
2. Center **retaining**: Cb(U_{N+1}) = Cb(U_N), but this entity is not the most highly ranked element in Cf(U_{N+1}). In this case therefore, Cb(U_{N+1}) is not the preferred candidate for Cb(U_{N+2}) and although it is retained as Cb in U_{N+1}, it is not likely to fill that role in U_{N+2}.
3. Center **shifting**[7]: Cb(U_{N+1}) ≠ Cb(U_N).

To exemplify the theory, here are two very simple discourses differing in their last sentences from Discourses A and B:

Discourse C
(3.1) John works at Barclays Bank.
(3.2) He works with Lisa.
(3.3) John is going to marry Lisa.
(3.6) Lisa has known him for two years.

Discourse D
(3.1) John works at Barclays Bank.
(3.2) He works with Lisa.
(3.3) John is going to marry Lisa.
(3.7) She has known John for two years.

Sentence (3.3) exhibits center continuation; the backward-looking centers of (3.2) and (3.3) and the forward-looking centers of sentences (3.1), (3.2) and (3.3) are listed as follows:

(3.1) John works at Barclays Bank.
Cb unspecified[8]
Cf = {John, Barclays Bank}
(3.2) He works with Lisa.
Cb = John
Cf = {John, Lisa}
(3.3) John is going to marry Lisa.
Cb = John
Cf = {John, Lisa}

In sentence (3.6) we have center retaining,[9]

(3.6) Lisa has known him for two years.
Cb = John
Cf = {Lisa, John}

whereas in (3.7) we have a center shift

(3.7) She has known John for two years.
Cb = Lisa
Cf = {Lisa, John}

Centering includes two rules which state:

> *Rule 1* If some element of Cf(U_N) is realised as a pronoun in U_{N+1}, then Cb(U_{N+1}) must also be realised as a pronoun.
>
> *Rule 2* Transition states are ordered. The *Continue* transition is preferred to the *Retain* transition, which is preferred to the *Shift* transition.[10]

Rule 1 stipulates that if there is only one pronoun in an utterance, then this pronoun should be the (backward-looking) center. It is reasonable to assume that if the next sentence also contains a single pronoun, then the two pronouns corefer. The center is the most preferred discourse entity in the local context which is to be referred to by a pronoun[11] (see also Chapter 2, section 2.2.3.2). The use of a pronoun to realise the backward-looking center indicates that the speaker/writer is talking/writing about the same thing. Psycholinguistic research (Gordon et al. 1993; Hudson-D'Zmura 1988) and cross-linguistic research (Di Eugenio 1990; Kameyama 1985, 1986, 1988; Walker et al. 1994) have validated that Cb is preferentially realised by a pronoun (e.g. in English) or by equivalent forms such as zero pronouns in other languages (e.g. Japanese).

Rule 2 provides an underlying principle for coherence of discourse. Frequent shifts detract from local coherence, whereas continuation contributes to coherence. Maximally coherent segments are those which do not feature changes of center, concentrate on one main discourse entity (topic) only and therefore require less processing effort.

Rule 2 is used as a preference in anaphora resolution (Brennan et al. 1987; Walker 1989; see also Chapter 4, section 4.6). As an illustration, consider the following discourse:

Discourse E
(3.8) Although Jenny was in a hurry, she was glad to bump into Kate.
(3.9) She told her some exciting news.

This discourse segment consists of the following utterances:

U_1 = Jenny was in a hurry
U_2 = she was glad to bump into Kate
U_3 = She told her some exciting news

The discourse entity 'Jenny' is both the backward-looking center of the second utterance Cb(U_2) and the preferred center Cp(U_2) on the list of forward-looking centers. Since continuation is preferred over retaining, centering favours 'Jenny' as both Cb(U_3) and Cp(U_3), therefore predicting *she* as 'Jenny' and *her* as 'Kate' (the instantiations *she* = 'Kate' and *her* = 'Jenny' would have signalled retaining since in this case we would have had Cp(U_3) = 'Kate', Cb(U_3) = 'Jenny').[12]

Centering has proved to be a powerful tool in accounting for discourse coherence and has been used successfully in anaphora resolution; however, as with every theory in linguistics, it has its limitations (see also Kehler 1997a). For instance, the original centering model only accounts for local coherence of discourse. In an anaphora resolution context, when the candidates for the antecedent of an anaphor in the current utterance U_K have to be identified, centering proposes that the discourse entities in the immediately preceding utterance U_{K-1} be considered. Centering, however, does not offer a solution for resolving anaphors in U_K whose antecedents can be found only in U_{K-2} (or even further back in the discourse). To overcome this restriction, Hahn and Strube (1997) put forward an alternative centering model that extends the search space for antecedents.

Walker (1998) goes even further and argues that the restriction of centering to operate within a discourse segment should be abandoned in favour of a new model integrating centering and the global discourse structure. To this end it is proposed that a model of attentional state, the so-called *cache model*, be integrated with the centering algorithm.

Strube (1998) proposes an alternative framework by replacing the backward-looking center and the centering transitions with an ordered list of salient discourse entities (referred to as *S-list*). The S-list ranking gives preference to *hearer-old* over *hearer-new* discourse entities (Prince 1981) and can account for the difference in salience between definite NPs (usually hearer-old) and indefinite NPs (usually hearer-new). In contrast to centering, Strube's model can also handle intrasentential anaphora.

Kibble (2001) discusses a reformulation of the centering transitions. Instead of defining a total preference ordering, the author argues that a partial ordering emerges from the interaction between 'cohesion' (maintaining the same center), 'salience' (realising the center as subject) and Strube and Hahn's notion of 'cheapness' (realising the anticipated center of a following utterance as subject).

A recent corpus-based study (Poesio et al. 2000) investigating the validity of the claim that each utterance has exactly one backward-looking center (apart from the first utterance in the discourse segment) and of the claim stating that if any Cf(U_N) is pronominalised in U_{N+1}, then Cb(U_{N+1}) must also be pronominalised, found that both these claims are subject to frequent violation. The authors experimented with different definitions of utterances (Kameyama 1998; Suri and McCoy 1994) such as sentences or finite clauses, and also treating adjuncts as embedded utterances. They also allowed a discourse entity to serve as a Cb of an utterance even if it was only indirectly referred to by a bridging reference. This led to fewer violations of the first claim but to more of the second. The study concludes that texts can be coherent even if the above claims do not hold since coherence can be achieved by other means such as rhetorical relations.

For practical examples of the use of centering rules in anaphora resolution, the reader may refer to the work of Brennan et al. (1987), Hahn and Strube (1997), Strube and Hahn (1996), Tetreault (1999) and Walker (1989). See also Chapter 4, section 4.6 and Chapter 5, section 5.10.

For further information on the various methods that have been proposed for center/focus tracking the reader is referred to Abraços and Lopes (1994), Brennan et al. (1987), Dahl and Ball (1990), Mitkov (1994b), Sidner (1983), Strube and Hahn (1996), Stys and Zemke (1995) and Walker et al. (1994).

3.2 Binding theory

The **binding theory** is part of the principles and parameters theory (Chomsky 1981, 1995) and, among other accomplishments, imposes important syntactic constraints as to how noun phrases may corefer. It accounts for the interpretation of anaphors including *reflexive pronouns* (hereafter referred to as *reflexives*), *personal pronouns* and *lexical noun phrases*.[13] The binding theory regards reflexives in English as short-distance anaphors and requires that reflexive anaphors refer to antecedents that are in a so-called *local domain*. Since reflexives are 'bound'[14] by their antecedents in this local domain, they are often called *bound anaphors*.[15] In contrast, personal pronouns are 'free' anaphors with respect to the same local domain – they are long-distance anaphors which permit antecedents to come only outside their local domain. Arriving at a useful definition of this local domain in structural terms has been an active area of research.

As an illustration, consider the following examples[16]:

(3.10) Victoria believed George had seen herself.
(3.11) Victoria believed George had seen him.

In (3.10) the noun phrases *Victoria* and *herself* do not corefer because the reflexive is too far away: a reflexive pronoun must corefer with a noun phrase in the same local domain. On the other hand, in (3.11) *George* and *him* cannot corefer because they are too close: a non-reflexive pronoun cannot corefer with the noun phrase in the same local domain.

Consider now the following examples:

(3.12) *Sylvia* believed *she* was the most diligent student.
(3.13) Sylvia believed he was the most diligent student.
(3.14) She believed Sylvia was the most diligent student.

In (3.12) *Sylvia* and *she* can be coreferential (although need not) but in (3.13) coreference between *Sylvia* and *he* is not possible because the anaphor and the antecedent must agree in gender and number (see also *constraints*, section 2.2.3.1). In (3.14) coreference between *she* and *Sylvia* does not hold and on this occasion one may be tempted to conclude that this is because the antecedent does not precede the anaphor. However, it is well known that in the case of cataphora (section 1.10), the anaphor may precede the antecedent:

(3.15) As *she* was always late for everything, *Jenny* could hardly be described as reliable.

The explanation as to why in (3.14) no coreference is possible will be provided later by the constraint introduced in section 3.2.3. The same constraint will also explain why in some cases coreference would be possible if a pronoun were used, as opposed to a lexical noun phrase such as *the young model* in example (3.16):

(3.16) Sylvia believes the young model is the most beautiful girl.

Before turning to the interpretation of reflexives, pronouns and lexical NPs, I shall introduce the structural relation of c-command which plays an important role in the constraints formulated in the next sections.

A node *A* c-commands a node *B* if and only if (Haegeman 1994):

(i) *A* does not dominate *B*
(ii) *B* does not dominate *A*
(iii) the first branching node dominating *A* also dominates *B*.

In Figure 3.1, which illustrates the notion of c-command, it can be seen for example (not exhaustive) that:

B c-commands *C* and every node that *C* dominates.
C c-commands *B* and every node that *B* dominates.
D c-commands *E* and *J*, but not *C*, or any of the nodes that *C* dominates.
H c-commands *I* and no other node.

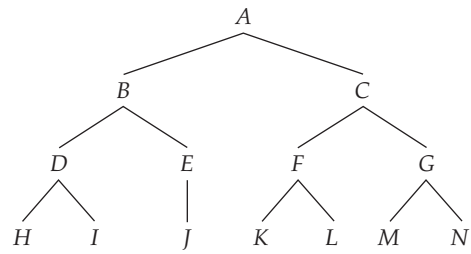

Figure 3.1

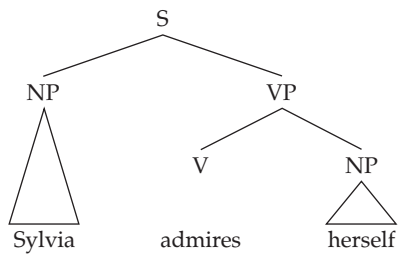

Figure 3.2

3.2.1 Interpretation of reflexives

The interpretation of reflexive anaphors is associated with factors such as grammatical agreement, c-command relation and local domain. To start, a reflexive anaphor must agree in person, gender and number with its antecedent.

Another key constraint that delimits the interpretation of reflexives states that

A reflexive anaphor must be c-commanded by its antecedent.

A close examination of the examples (see Figures 3.2, 3.3 and 3.4)[17]

(3.17) *Sylvia* admires *herself*.
(3.18) *Sylvia* likes the photo of *herself*.
(3.19) *Sylvia* believes *herself* to be the most beautiful girl.

confirms that *herself* is c-commanded by *Sylvia* in all three sentences.

Before attempting to identify the antecedent, it is helpful to determine the maximum extent of the search scope. The establishment of the exact local domain in which the reflexive anaphor must be bound is not a trivial matter.

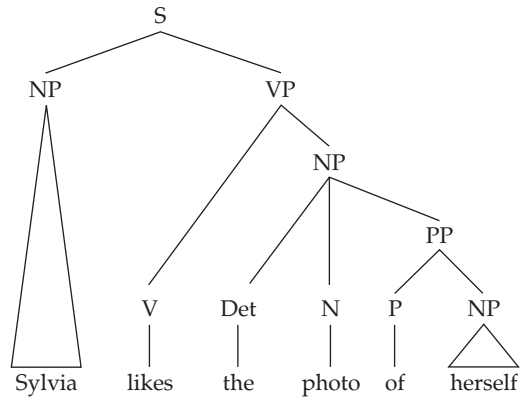

Figure 3.3

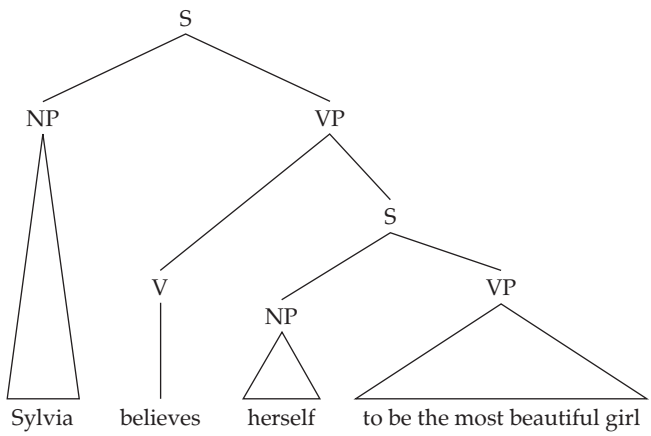

Figure 3.4

For the sake of simplicity, I shall describe the local domain very loosely as a clause or a complex NP (e.g. possessive constructions), though the presence of the subject and governor[18] is relevant too.[19] The following examples demonstrate the scope of the local domain, which is denoted by square brackets:

(3.20) [*George* hurt *himself*].
(3.21) Nelly thinks that [*George* hurt *himself*].
(3.22) Vicky admires [*Elitza*'s picture of *herself*].[20]

The constraint that the antecedent of a reflexive anaphor must c-command it within the local domain[21] has already been used in computational systems (Ingria and Stallard 1989) to assign possible antecedents of bound anaphors. As an illustration, in the following example

(3.23) Jane thought the yellow looked better on Sylvia but Sylvia wanted to choose for herself.

Sylvia would be assigned unambiguously as an antecedent of *herself*. Here *Sylvia* c-commands *herself* and, as opposed to *Jane*, is in the local domain of the reflexive anaphor.

3.2.2 Interpretation of personal pronouns

The interpretation of non-reflexive pronominal anaphors differs from that of reflexives. From the examples

(3.17) *Sylvia* admires *herself*.
(3.24) *Sylvia* admires *her*.

it is clear that whereas *herself* is bound and refers to *Sylvia*, the pronoun *her*, which is in the same syntactic position as *herself*, is free within the domain defined by the sentence and must refer to an entity different from *Sylvia* and outside this domain. Note that *Sylvia* c-commands both *herself* and *her*.

The domain in which pronominal anaphors are free is the same as the domain in which reflexives are bound (see Haegeman 1994). The antecedent of a reflexive lies within the local domain of the reflexive anaphor and c-commands it. On the other hand, a noun phrase and a pronominal anaphor cannot be coreferential if the noun phrase is situated in the local domain of the anaphor and c-commands it.

The main constraint in the interpretation of pronouns stipulates that

A pronoun cannot refer to a c-commanding NP within the same local domain.

This constraint has been used in automatic anaphora resolution (Ingria and Stallard 1989) to narrow down the search scope of candidates for antecedents. For instance, applying this constraint to the examples

(3.24) Sylvia admires her.
(3.25) Sylvia likes the photograph of her.
(3.26) Sylvia told Jane about her.

would rule out *Sylvia* in (3.24) and (3.25), and *Sylvia* and *Jane* in (3.26), as possible antecedents (note that *her* is c-commanded by *Sylvia* in (3.24) and (3.25), and c-commanded by both *Sylvia* and *Jane* in (3.26); *Sylvia* (and *Jane*) lie in the local domain of *her*). Finally in the sentence

(3.27) *Sylvia* listened to Jane's song about *her*.

the pronoun *her* can refer to the NP *Sylvia* because although *Sylvia* c-commands *her*, it is not in the local domain of the pronoun (the local domain is *Jane's song*

about her – note the role of *Jane* as the 'subject' of the possessive NP construction *Jane's song about her*). For the anaphor *her* to corefer with *Jane*, it would have to be reflexive (*herself*).

3.2.3 *Interpretation of lexical noun phrases*

Lexical noun phrases are the class of noun phrases which are not pronouns (including reflexive or reciprocal pronouns), such as *Sylvia* or *the young model*. These types of noun phrases, also referred to as *referential expressions*, are (as their name suggests) inherently referential, select their reference from the universe of discourse and therefore have independent reference. In contrast to reflexive pronouns which must be bound locally, or non-reflexive pronouns which must be free locally but may be bound outside their local domain, referential expressions must be free everywhere (Haegeman 1994), that is, they cannot be bound by an antecedent within or outside their local domain.

(3.28) *Michelle* asked if *the manageress* believed that *Sarah* knew that *the young model* was leaving.

This example shows that no matter how far away the lexical NP is, there is no 'obligation' for it to corefer with another NP within or outside a certain domain.

An important constraint delimiting the interpretation of lexical noun phrases states that

A non-pronominal NP cannot corefer with an NP that c-commands it.

This constraint has been used in anaphora resolution systems (Ingria and Stallard 1989) to discount coreference in examples such as

(3.29) She admires Sylvia.
(3.30) She likes a photograph of Sylvia.
(3.31) Sylvia said the young model was the most beautiful girl.

In these examples the non-pronominal noun phrases *Sylvia* and *the young model* (and *the most beautiful girl*) are c-commanded by the NPs *She* and *Sylvia* respectively and, therefore, cannot be coreferential with them.

The binding theory is helpful in determining impossible antecedents of pronominal anaphors and in assigning possible antecedents to bound anaphors, and some of the constraints outlined above have been used for automatic anaphora resolution (Ingria and Stallard 1989; Carvalho 1996). However, the theory is still an active area of research in syntax and is not yet fully developed: there are still a number of cases that cannot be accounted for. For a useful introduction to later developments in this theory, see Harbert (1995).

3.3 Other related work

Centering is compatible with the **theory of discourse structure** proposed by Grosz and Sidner (1986). Grosz and Sidner suggest that discourse structure is

based on three components: a *linguistic structure*, an *intentional structure* and an *attentional state*. At the level of linguistic structure, discourses divide into constituent discourse segments; an embedding relationship may hold between two segments (Grosz et al. 1995).

Previous research on **focusing** provides the background for centering theory. Grosz (1977a, b) explained that there are two levels of focusing in discourse: *global* and *immediate* (or *local*). Entities that are most relevant and central throughout the discourse are termed globally focused, whereas those that are the most important and central to a specific utterance within the discourse are said to be immediately or locally focused.

Sidner (1979) offered a detailed analysis of **local focusing**. Sidner assumes that at a given point, a well-formed discourse is 'about' some entity mentioned in it. This entity is called the *focus of discourse* or *discourse focus* (Sidner 1979, 1983), which she further assumes can be identified by the hearer/reader. Similarly to the assumptions of centering theory, as the discourse progresses, the speaker may maintain the same focus or re-focus on another entity. Also, the change of focus, or the lack thereof, is signalled by the linguistic choices of the speaker and in particular by the use of anaphoric expressions.

Sidner's apparatus is as follows.[22] The state of focus at a given point in the text is represented by the contents of six focus registers. The *discourse focus* (DF) and *actor focus* (AF) registers each contain the representation of a single entity mentioned in the text; the *potential discourse focus* (PDF), *potential actor focus* (PAF), *discourse focus stack* (DFS) and *actor focus stack* (AFS) registers each contain a list of zero or more entities. The entities mentioned in the sentence (by noun phrases or clauses) other than the discourse focus are called potential discourse foci.[23] Sidner argues that the actor focus is needed to account for the behaviour of pronouns. It is defined as the agent of the most recent sentence that has an agent. Other animate expressions in the most recent sentence are regarded as potential actor foci.

Sidner proposed a method for assigning antecedents of definite pronouns[24] and definite full noun phrases based on her algorithm for tracking the discourse focus. The algorithm makes an initial prediction of the focus after the first sentence; this selection is called the *expected focus*. The choice of the expected focus depends on syntactic and semantic criteria. The syntactic criteria which point to the expected focus include the *subject* of a sentence if the sentence is an *is-a* or *there*-insertion sentence (e.g. There was once a prince who was changed into a frog) or *cleft constructions* (e.g. It was George who ate the whole chocolate). In the absence of syntactic pointers, the semantic category *theme* is given preference (Sidner 1983). The DF register is set to the expected focus and the PDF register to the potential discourse foci.

For non-initial sentences, an anaphora interpretation algorithm is applied to each anaphor.[25] Each rule in the algorithm appropriate to the anaphor suggests one or more antecedents[26] according to what the focus registers contain. The proposed antecedent is assessed by an inference mechanism (which Sidner assumes to exist) which looks for any resulting contradictions. The first proposal not giving rise to a contradiction is accepted.

Next, a *focus update algorithm* updates the focus registers, taking the results of anaphor interpretation into account. If the DF changes, the old DF is pushed on to the DFS, or if the new DF is already in the DFS, the DFS is popped. Whether the DF changes or not, the PDF list consists of representations of every entity mentioned in the current sentence other than the DF itself. The AF, PAF and AFS registers are updated analogously, except that they can only hold animate entities.

Once the focus registers have been updated, the next sentence is passed on to the anaphora interpretation algorithm for processing.

In Sidner's theory, definite anaphors are regarded as signals that tell the hearer what elements are in focus and in which registers. On the other hand, the focus state, as defined by the six focus registers, partly determines the interpretation of definite anaphors. Focusing reduces the inferencing load necessary to resolve anaphors and, as a consequence, a number of algorithms have used Sidner's original or a modified model of focusing (Carter 1986, 1987a; Dahl 1986; Azzam et al. 1998b). Sidner's theory, however, does not specify how candidates which are in the same sentence as the pronoun should be considered and does not take into account any possible interaction between the applications of the rules to different anaphors in a sentence. These problems, which need to be addressed in a practical system, are largely solved by Carter (1987a).

Joshi and Kuhn (1979) and Joshi and Weinstein (1981) were the first to discuss the connection between changes in immediate focus and the complexity of semantic inferences.[27] To avoid confusion with previous uses of the term 'focus', they introduced the concept of centering. Their notions of 'forward-looking' and 'backward-looking' centers (see section 3.1) correspond roughly to Sidner's potential foci and local focus.

Another theory successfully used for anaphora resolution is the **Discourse Representation Theory** (Kamp and Reyle 1993). According to DRT, semantic interpretation is a matter of incorporating the content of a sentence into the existing context (Poesio 2000). The context is described as a set of *discourse representation structures* (DRSs) derived systematically from the syntactic structure of the sentences of a discourse. Apart from representing the meaning of discourse, these structures impose constraints on pronoun resolution. A DRS is a pair consisting of a set of discourse referents[28] together with a set of conditions expressing properties of these referents. Each DRS is represented as a diagram, with the discourse referents displayed at the top of the diagram and the conditions below them.

As an illustration, the sentence 'John loves Lisa' would correspond to the DRS-diagram on Figure 3.5:

Note that this DRS includes a discourse referent for John (x) as well as for Lisa (y). DRSs have well-specified semantics and the DRS in Figure 3.5 is semantically equivalent to the first-order logic formula

$$\exists\, x, y \text{ John } (x) \wedge \text{Lisa } (y) \wedge \text{loves } (x, y)$$

Similarly the discourse 'John loves Lisa. He adores her' will be represented as in Figure 3.6.

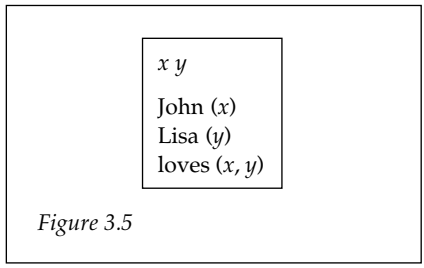

Figure 3.5

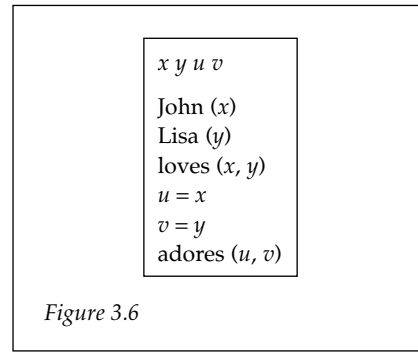

Figure 3.6

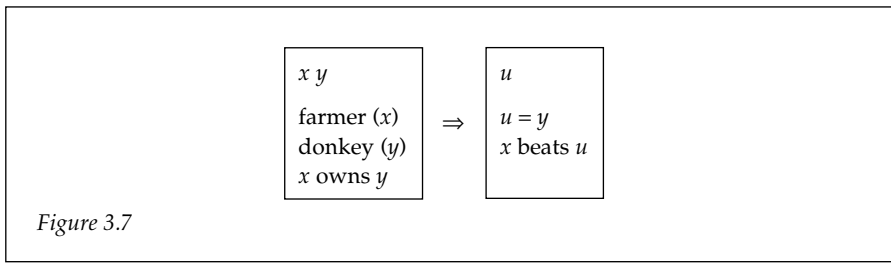

Figure 3.7

In DRT discourse referents are made accessible beyond the clause in which they are introduced because the semantic interpretation procedure in DRT always begins by adding the syntactic interpretation of a new sentence to the existing DRS. One of the basic premises of DRT is that indefinite NPs introduce new discourse variables into the discourse. Definite NPs, on the other hand, update the state of existing discourse variables.

DRS can represent conditionals and quantifiers as more complex DRS-diagrams. As an illustration, the sentence 'Every farmer who owns a donkey beats it' can be expressed by Figure 3.7 which allows the discourse referent *y* to be accessible to the position occupied by *it*.[29]

Discourse Representation Theory has had an important impact on the research in anaphora resolution and has been adopted by a number of researchers (e.g. Günther and Lehmann 1983; Abraços and Lopes 1994; Carvalho 1996). Some researchers have combined DRT and focusing (e.g. Cormack 1993; Abraços and Lopes 1994). Cormack (1998) also highlighted some shortcomings of the DRT model and proposed modifications to the original theory. Some imperfections of the DRT (e.g. redundancy in the representation of discourse referents in the universe of the DRS) have recently been pointed out by Cornish (1999).

Other theories or formalisms which have been used successfully in anaphora resolution include **Webber's formalism** (1979) and the **veins theory** (Cristea et al. 1998).

3.4 Summary

Theories such as centering, binding theory, focusing and DRT (discourse representation theory) have been employed successfully in anaphora resolution. The main idea of centering theory is that in an utterance a certain entity called the *center* is more prominent than others. This imposes constraints on the use of pronouns in that if a discourse entity is pronominalised in the following utterance, then the center is pronominalised too. As a consequence, the center of the preceding utterance is the preferred candidate for antecedent of a pronominal anaphor under consideration. Centering also defines a set of transitions, each one of which has a different impact on the coherence of the discourse. The transitions are ranked in preferential order and can be used as preferences in the resolution process. The binding theory imposes important syntactic constraints as to how noun phrases may corefer. It accounts for the interpretation of anaphors that are reflexive pronouns, personal pronouns and lexical noun phrases. In particular, the antecedent of reflexive pronouns must be in the so-called local domain, whereas the antecedent of personal pronouns must be outside this domain.

Notes

1 In very broad terms, we can think of an utterance as a finite clause or a sentence.
2 Centering does not assign a center to the first utterance of a discourse segment.
3 To simplify notation, I shall drop *D* which denotes the discourse segment of which the utterance is part.
4 The backward-looking center is often referred to simply as the center. However, the qualification 'backward-looking' is in line with the requirement that the backward-looking center of a current utterance establishes a link to the previous utterance and must be on its list of forward-looking centers.
5 Apart from the initial utterance of a discourse segment.
6 This statement is valid for English and for a number of other languages.
7 Brennan et al. (1987) distinguish between *smooth-shift* or *shifting – 1* (if $\mathrm{Cb}(U_{N+1}) = \mathrm{Cp}(U_{N+1})$) and *rough-shift* or simply *shifting* (if $\mathrm{Cb}(U_{N+1}) \neq \mathrm{Cp}(U_{N+1})$).
8 According to Grosz et al. (1995), the first utterance in a discourse segment is not assigned a center. It could be argued that there are cases where the most salient element is clearly identifiable even in the first utterance (e.g. with cleft constructions).
9 Note that if there is one pronoun, it realises the center (see below Rule 1).
10 As defined by Brennan et al. (1987), smooth-shift is preferred to rough-shift (see also Chapter 4, section 4.6).
11 Deleted as a zero pronoun in languages exhibiting extensive use of zero pronouns such as Japanese, Italian, Spanish and Bulgarian.
12 Note that *she* and *her* in U_3 cannot be coreferential (see section 3.2.2).
13 The binding theory addresses the interpretation of reciprocals too (Haegeman 1994), but they will not be discussed here. Chomsky restricts the term *anaphor* to reflexives and reciprocals.
14 To be more precise, the binding theory defines *binding* in terms of c-command as follows: *x binds y* if and only if (i) *x* c-commands *y* (see the definition of c-command) and (ii) *x* and

y are coindexed (Haegeman 1994). The latter means that *x* and *y* are coreferential: a reflexive cannot have an independent reference but depends for its reference on the 'binder'.

15 Note the alternative use of the term *bound anaphor* outside the binding theory to denote anaphors which have as their antecedent quantifying NPs such as *every student*, *most readers* (e.g. see Chapter 1, section 1.2).

16 Many of the examples of this section are based on or adapted from Haegeman (1994).

17 Note that the trees represented in these diagrams are rather simplified.

18 The governor is the element whose presence imposes a requirement upon a second element, the governed category. Usually, all heads (e.g. main verb in a sentence) are regarded as potential governors.

19 For a more detailed and precise description of 'local domain', c-command and government see Haegeman (1994).

20 A slightly more precise but still simplified and not complete procedure for finding the local domain can be described as follows: (i) find the governor of the reflexive, (ii) find the closest subject. The smallest finite clause or noun phrase containing these two elements will be the binding domain in which the reflexive must be bound with a c-commanding and agreeing antecedent (Haegeman 1994). This definition explains why the local domain of '*George* believes *himself* to be the best' is the whole sentence (the reflexive *himself* is governed by the verb *believe*). Also, the NP *Elitza* is regarded as the subject (in Haegeman's terminology) of the complex NP *Elitza's picture of herself* (consider the semantically equivalent form *Elitza pictured herself*).

21 It should be noted that a number of counterexamples to the original statements of binding theory can be found. For instance in 'No composer enjoyed a better family background than *Mozart*. Like *himself*, both his father and sister were remarkable musicians' (Quirk et al. 1985) a cross-sentential reference is possible.

22 To a great extent, the outline here follows that of Carter (1987a).

23 The term *potential focus* refers to any new item in the discourse. According to Sidner, potential foci have a short lifetime. If a potential focus does not become the focus after the interpretation of the sentence following the one in which the potential is seen, it is dropped as a potential focus (Sidner 1983). As shown in other work (Abraços and Lopes 1994), however, potential foci can be re-activated later in the discourse even if they do not become the focus in a subsequent sentence.

24 As opposed to indefinite pronouns such as *some*, *few*, etc.

25 Sidner (1979) proposes seven algorithms for anaphors of various types and in various roles in the sentence representation.

26 'Specifications' in Sidner's original terminology.

27 Inferences required to integrate a representation of the meaning of an individual utterance into a representation of the meaning of the discourse of which it was part.

28 The set of discourse referents is called the universe of the DRS.

29 For more details on accessibility see Kamp and Reyle (1993).

CHAPTER FOUR

The past: work in the 1960s, 1970s and 1980s

4.1 Early work in anaphora resolution

This chapter covers work on anaphora resolution from the 1960s, 1970s and 1980s, outlining the most important research of this period as reported by Bobrow (1964), Winograd (1972), Woods et al. (1972), Hobbs (1976, 1978), Carter (1986, 1987a), Rich and LuperFoy (1988) and Carbonell and Brown (1988). The early work typically relied on heuristic rules and did not exploit full linguistic analysis, as exemplified by Bobrow's STUDENT system or Winograd's SHRDLU (the latter being much more sophisticated and featuring a set of clever heuristics). However, it did not take long before the research evolved into the development of approaches benefiting from a variety of knowledge sources. For instance, Hobbs's naïve approach (Hobbs 1976) was primarily based on syntax, whereas LUNAR and Wilks's approach mainly exploited semantics. The late 1970s saw the first discourse-oriented work (Sidner 1979; Webber 1979); later approaches went even further, resorting to some form of real-world knowledge (Carter 1986; Carbonell and Brown 1988; Rich and LuperFoy 1988).

As with many NLP tasks in the 1970s and 1980s, anaphora resolution was more theoretically-oriented and rather ambitious in terms of the types of anaphora handled.[1] In the 1990s, however, the rising awareness of the formidable complexity of anaphora resolution and the pressing need for working systems encouraged more practical and down-to-earth research, often limiting the treatment of anaphora to a specific genre, but offering working and robust solutions in exchange.

It is worth noting that much of the early work is difficult to compare with recent methods (e.g. in terms of resolution success rate) since many of the early systems were not implemented or evaluated. Those evaluated were usually manually tested and focused on the resolution algorithm only: the texts were syntactically and semantically analysed by humans, thus offering the algorithm the advantage of operating on a perfectly pre-processed input.

In the following sections, some of the most significant projects on anaphora resolution in the 1960s, 1970s and 1980s will be briefly outlined.[2] Where appropriate, I have sought to provide a brief description of the resolution methods and techniques used, hoping that this will help the reader to better understand how automatic resolution of anaphora works.

4.2 STUDENT

One of the earliest attempts to resolve anaphors by a computer program is reported in STUDENT (Bobrow 1964), a high-school algebra problem-answering system. STUDENT tries to pattern-match anaphors and antecedents. For example, it can successfully tackle the following text.

(4.1) The number of soldiers *the Russians* have is half the number of guns they have. The number of guns is 7000. What is the number of soldiers *they* have?

The system identifies the antecedent of *they* as *the Russians* by matching up *the number of soldiers the Russians have* and *the number of soldiers they have*.

Bobrow's heuristics include a rule saying that phrases containing *this* refer to preceding 'similar' phrases and in (4.2) *this price* is taken to refer to *the price*.

(4.2) *The price* of a radio is 69.70 dollars. *This price* is 15% less than the market price.

In fact, STUDENT only relies on limited heuristics and apart from simple matching techniques, the sentences are not parsed and no real resolution process takes place. As Hirst (1981) points out, the following two references to sailors would not be matched up:

(4.3) The number of soldiers the Russians have is twice the number of *sailors* they have. The number of soldiers is 7000. How many *sailors* do the Russians have?

4.3 SHRDLU

Winograd (1972) was the first to develop 'real' procedures for pronoun resolution in his SHRDLU system, which maintained dialogues about a microworld of shapes such as blocks and pyramids. His heuristics are much more complex than those of STUDENT, thus providing an impressive and (especially for its time) sophisticated treatment of anaphors. SHRDLU is able to handle references to earlier parts of the conversation between the program and its user.

Winograd's algorithm checks previous noun phrases for possible antecedents and does not consider only the first likely candidate but examines all the possibilities in the preceding text. Plausibility is rated on the basis of syntactic position: subject is favoured over object and both are favoured over the object of a preposition. In addition, 'focus' elements are favoured, the focus being determined from the answers to wh-questions and from indefinite noun phrases in yes–no questions. If none of the candidates for antecedents stands out clearly as an antecedent, the user is asked to help in the selection between the best candidates.

SHRDLU has a number of practical heuristics which today, almost 30 years on, are still very relevant to pronoun resolution systems. For instance, if *it* or *they* occurs twice in the same sentence, or in two adjacent sentences, the occurrences are assumed to be coreferential.[3]

SHRDLU can resolve some references to events as in

(4.4) Why did you do it?

by remembering the last event referred to.

It is also worth pointing out that the system can handle some contrastive uses of *one* as in

(4.5) a big green pyramid and a little one

A list of pairs of words such as *big* and *little* which are used contrastively is employed to work out that *little one* here means *little green pyramid* and not *little pyramid* or *little big green pyramid*.

Finally, the SHRDLU can handle some zero anaphors as in

(4.6) Find the red blocks and stack up three.

by identifying the elliptically omitted reference as *red blocks*.

4.4 LUNAR

The LUNAR Sciences Natural Language Information System (Woods et al. 1972) uses an ATN parser (Woods 1970) and a semantic interpreter based on the principles of procedural semantics (Woods 1968). Anaphora resolution is performed within the semantic interpreter which distinguishes two classes of anaphors: partial and complete. Anaphors that have complete NPs as antecedents are regarded as *complete*, while those which refer to parts of preceding NPs are termed *partial*. The following examples show a complete and a partial anaphor, respectively:

(4.7) Which *coarse-grained rocks* have been analysed for cobalt? Which *ones* have been analysed for strontium?

(4.8) Give me all *analyses of sample 10046* for hydrogen. Give me *them* for oxygen.

In (4.8) the antecedent of *them* is *analyses of sample 10046* and not the complete NP *all analyses of sample 10046 for hydrogen*. Such partial anaphors are signalled by the presence of a relative clause or prepositional phrase modifying the pronoun (in this case *for oxygen*). It is clear that complete anaphors are identity-of-reference anaphors, whereas partial anaphors are identity-of-sense anaphors.

The resolution strategy for partial anaphors is to search for an antecedent which occurs in a syntactic (and semantic) structure parallel to that of the anaphor. In (4.8) for instance, the parallel structures *analyses of sample 10046 for hydrogen* and *them for oxygen* are established (both being 'NP + prepositional phrase' structures) and the prepositional phrase (PP) *for oxygen* is substituted for

the PP *for hydrogen*. Thus the system derives the meaning of the anaphor as *analyses of sample 10046 for oxygen*.

Unlike Bobrow's system, this approach operates at syntactic and semantic levels rather than at the lower level of lexical matching with a little added syntax. It suffers, however, from the same limitation as STUDENT: LUNAR can resolve only anaphors where the antecedent is of a similar structure. As an illustration, LUNAR would not be able to resolve the anaphors *ones* and *those* in (4.9), (4.10) or (4.11):

(4.9) Give me all *analyses of sample 10046* for hydrogen. Give me the oxygen *ones*.
(4.10) Give me all *analyses of sample 10046* for hydrogen. Give me the *ones* carried out for oxygen.
(4.11) Give me all *analyses of sample 10046* for hydrogen. Give me *those* that have been done for oxygen.

Three different methods are used for complete anaphoric references depending on the form of the anaphor. The first method applies to lexical NP anaphors of the form 'Demonstrative pronoun + Noun':

(4.12) Do any *breccias* contain aluminium? What are *those breccias*?

The technique used here is to look for a preceding noun phrase whose head is *breccias* and propose this noun phrase as an antecedent. LUNAR would not be able to resolve anaphors whose heads are different from the heads of their antecedents as in cases of hypernymy or synonymy and would also fail to track down the antecedent in (4.13):

(4.13) Do any *breccias* contain aluminium? What are *those samples*?

The second class of anaphors LUNAR deals with are pronouns such as those in (4.14):

(4.14) How much titanium is in *type B rocks*? How much silicon is in *them*?

In order to identify *type B rocks* as antecedent of *them*, the system uses semantic and real-world knowledge that *silicon* is an element, that elements are contained in samples and that *type B rocks* are samples.

The third type of complete anaphors LUNAR can handle are *one*-anaphors as in (4.7). These are resolved either with or without modifiers like *too* and *also*. Note that the presence of *too* or *also* will completely change the meaning/reference:

(4.15) Which *coarse-grained rocks* have been analysed for cobalt? Which *ones* have been analysed for strontium too?

The resolution of this type of anaphor is based on similar selectional restriction rules to those used for pronouns.

The primary limitation of LUNAR is that it cannot handle intrasentential anaphors.

4.5 Hobbs's naïve approach

Hobbs proposed two approaches to pronoun resolution: one *syntactic* operating on syntactic trees and another using *semantic* knowledge (Hobbs 1976, 1978). In the following, the section will focus on his syntactic treatment, often referred to as *Hobbs's naïve approach*, which has attracted considerable attention in the research community and is still one of the most successful algorithms: recent comparisons show that it is still on a par with the vast majority of modern resolution systems.

Hobbs's algorithm operates on surface parse trees and on the assumption that these represent the correct grammatical structure of the sentence with all adjunct phrases properly attached, and that they feature 'syntactically recoverable omitted elements' such as elided verb phrases and other types of zero anaphors or zero antecedents. Hobbs also assumes that an NP node has an N-bar node below it, with N-bar denoting a noun phrase without its determiner. Truly adjunctive prepositional phrases are attached to the NP node. This assumption, according to Hobbs, is necessary to distinguish between the following two sentences:

(4.17a) Mr. Smith saw a *driver* in *his* truck.
(4.17b) Mr. Smith saw a driver of his truck.

In (4.17a) *his* may refer to the *driver*, but in (4.17b) it may not. The structures to be assumed for the relevant noun phrases in (a) and (b) are shown in Figure 4.1.

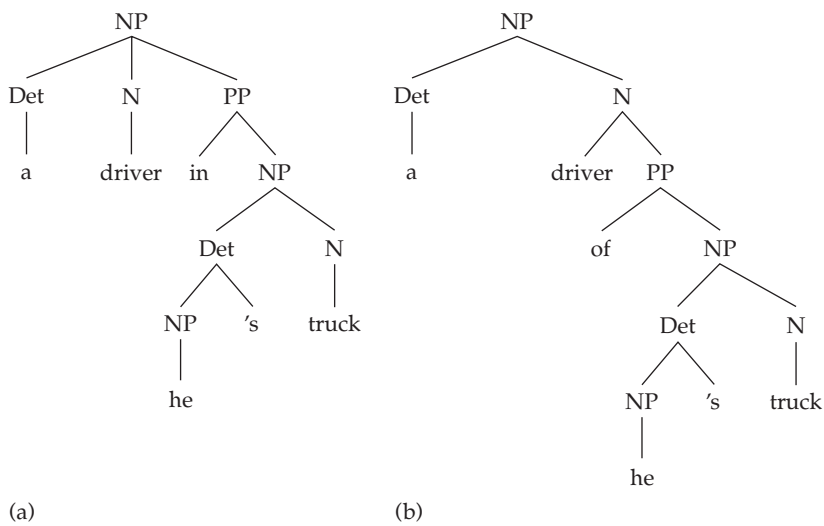

Figure 4.1 Syntactic structures corresponding to (4.17a) and (4.17b).

1. Begin at the NP node immediately dominating the pronoun in the parse tree of the sentence S.
2. Go up the tree to the first NP or S node encountered. Call this node *X*, and call the path used to reach it *p*.
3. Traverse all branches below node *X* to the left of path *p* in a left-to-right, breadth-first fashion.[4] Propose as the antecedent any NP node encountered that has an NP or S node between it and *X*.
4. If the node *X* is the highest S node in the sentence, traverse the surface parse trees of previous sentences in the text in order of recency, the most recent first; each tree is traversed in a left-to-right, breadth-first manner, and when an NP node is encountered, it is proposed as antecedent. If *X* is not the highest node in the sentence, proceed to step 5.
5. From node *X*, go up the tree to the first NP or S node encountered. Call this node *X* and call the path traversed to reach it *p*.
6. If X is an NP node and if the path *p* to *X* did not pass through the N-bar node that *X* immediately dominates, propose *X* as the antecedent.
7. Traverse all branches below the node *X* to the left of path *p* in a left-to-right, breadth-first manner. Propose any NP node encountered as the antecedent.
8. If *X* is S node, traverse all branches of node *X* to the right of path *p* in a left-to-right, breadth-first manner, but do not go below any NP or S node encountered. Propose any NP node encountered as the antecedent.
9. Go to step 4.

Figure 4.2 Hobbs's algorithm.

4.5.1 The algorithm

Hobbs's algorithm traverses the surface parse tree in a particular order looking for a noun phrase of the correct gender and number. The traversal order is detailed in Figure 4.2 (Hobbs 1976, 1978).

Steps 2 and 3 of the algorithm take care of the level in the tree where a reflexive pronoun would be used. Steps 5–9 cycle up the tree through S and NP nodes. Step 4 searches the previous sentences in the text.

As an illustration, Hobbs chooses the following context-free grammar to generate surface structures of a fragment of English (parentheses indicate optionality; asterisks mean 0 or more occurrences):

S → NP VP
NP → (Det) N-bar (PP/Rel)*
NP → pronoun
Det → article/NPs
N-bar → noun (PP)*
PP → preposition NP
Rel → wh-word S
VP → verb NP (PP)*

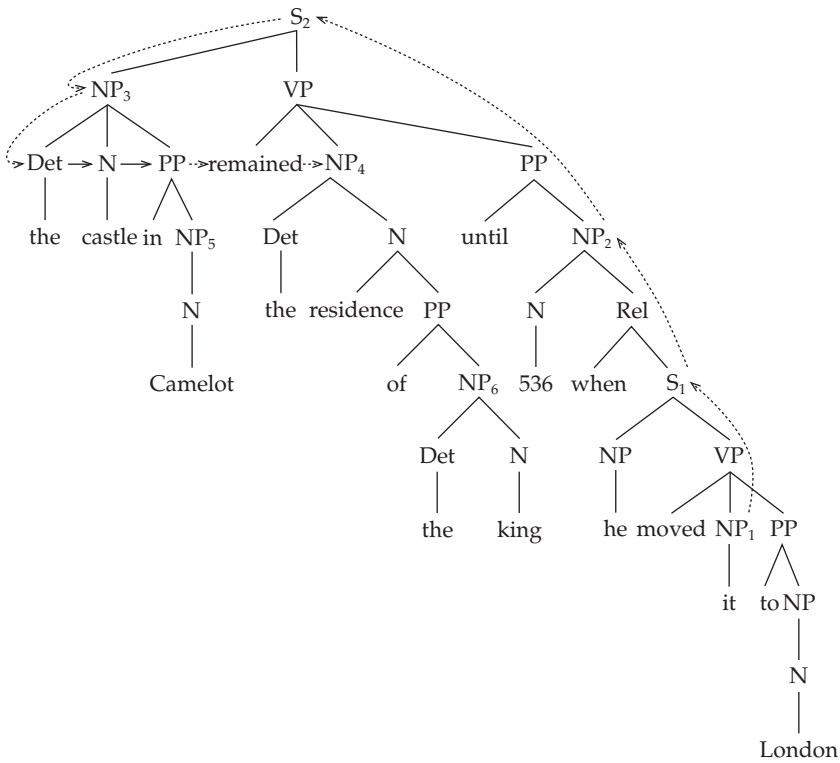

Figure 4.3 Structure of (4.18) and illustration of the work of the algorithm.

Figure 4.3 illustrates the algorithm working on the sentence

(4.18) The castle in Camelot remained the residence of the king until 536 when he moved it to London.

In Figure 4.3, Node NP_1 labels the starting point of step 1 of the algorithm. Step 2 takes us to the node S_1; this node is called *X*. The path *p* is marked with a dashed line. Step 3 searches to the left of *p* below *X* but finds no eligible NP node. Step 4 does not apply. Step 5 rises to NP_2. Step 6 proposes NP_2 as antecedent. Therefore, at this stage *536* is proposed as antecedent.[5]

Simple selectional constraints such as 'dates don't move', 'places can't move' or 'large objects don't move' can help rule out *536* as an antecedent.[6]

After NP_2 is rejected, steps 7 and 8 bring nothing, and control is returned to step 4 which does not apply. Step 5 rises to S_2, where step 6 does not apply. In step 7, the breadth-first search first recommends NP_3 (*the castle*), which is rejected by selectional constraints. The search then continues to NP_4 to correctly propose *the residence* as antecedent of *it* (Hobbs 1976, 1978).

If the algorithm were tracking down the antecedent of *he*, the search would continue, first turning down NP_5 (*Camelot*) because of gender mismatch and then correctly settling upon NP_6, *the king*.

Hobbs notes that when attempting to resolve *they*, his algorithm considers plural and collective singular noun phrases and selects semantically compatible entities. In the example

(4.19) John sat on the sofa. Mary sat by the fireplace. *They* faced each other.

the algorithm would propose *Mary and John*, rather than *Mary*, *the fireplace* or *the sofa*. When two plurals are conjoined, the conjunction is preferred over either plural.

(4.20) *Human bones and relics* were found at this site. *They* were associated with elephant tusks.

Hobbs also adopts two syntactic constraints proposed by Langacker (1969).[7] The first constraint is that a non-reflexive pronoun and its antecedent may not occur in the same simple sentence. As an illustration, consider (4.21) and (4.22)

(4.21) John likes him.
(4.22) John's portrait of him.

John and *him* cannot be coreferential (in English, a coreferential pronoun here would have to be reflexive: *John* likes *himself*). This constraint is accommodated by steps 2 and 3 of Hobbs's algorithm.[8]

The second rule, proposed by Langacker (1969), states that the antecedent of a pronoun must precede or command the pronoun. A node NP_1 is said to command node NP_2 if neither NP_1 nor NP_2 dominates the other and if the S node which most immediately dominates NP_1 dominates but does not immediately dominate NP_2.[9] The command relation was proposed by Langacker to account for backward pronominalisation:

(4.23) After *he* robbed the bank, *John* left town.
(4.24) That *he* was elected chairman surprised *John*.

Step 8 of the algorithm handles such cases.[10]

4.5.2 *Evaluation of Hobbs's algorithm*

Hobbs evaluated his algorithm on 300 pronouns from three different texts: 100 of these pronouns were from William Watson's *Early Civilization in China*, 100 were from the first chapter of Arthur Haley's novel *Wheels* and 100 from the 7 July 1975 edition of *Newsweek*. The pronouns were *he*, *she*, *it* and *they*[11]; *it* was not counted when referring to a syntactically recoverable 'that' clause or when pleonastic.[12]

As Hobbs pointed out, significant differences were noted among the texts. *Early Civilization in China* is characterised by long, grammatically complex sentences requiring every step of the algorithm. *Wheels*, on the other hand, is highly colloquial. The sentences are generally short and simple, often comprising nothing more than an exclamation, and with dialogue prevailing. Finally, *Newsweek*

has a very rich verbal structure, which mixes grammatical complexities and colloquialisms.

Hobbs investigated the distribution of pronouns and their antecedents in the aforementioned texts. To this end, he defined the following *candidate sets* C_0, C_1 . . . C_N with C_0 being a subset of C_1, C_1 of C_2, etc.:

C_0 = (a) the set of entities in the current sentence and the previous sentence if the pronoun comes before the main verb, or (b) the set of entities only in the current sentence if the pronoun comes after the main verb
C_1 = the set of entities in the current sentence and the previous sentence
C_N = the set of entities in the current sentence and the previous N sentences

Hobbs found that 90% of all antecedents were in C_0 while 98% were in C_1. This observation motivated him to propose that in most cases Klapholz and Lockman's (1975) hypothesis, stating that 'the antecedent is always found within the last N sentences, for some small N', worked (Charniak 1972 was more explicit and proposed, with reservations, $N = 5$). Hobbs (1976) noted, however, that 'there is no useful absolute limit on how far back one need look for the antecedent'. In one of his examples the antecedent occurred nine sentences before the pronoun. He also found out that the pronoun *it*, especially in technical writing, could have a very large number of plausible antecedents[13] in one sentence, and one example in *Early Civilization in China* had 13. Therefore, he noted that 'any absolute limit we impose might therefore have dozens of plausible antecedents and would hardly be of practical value' (Hobbs 1976).

Hobbs also tested the heuristic Winograd used in his micro-world blocks system, stating that 'if the same pronoun occurs twice in the same sentence or in two consecutive sentences, the occurrences are coreferential' (Winograd 1972; see also section 4.3). The heuristic performed less successfully than expected. It was applicable 48 times (out of the 132 'conflicts') and returned the correct antecedent only 28 times, or 58.3%. On the *Early Civilization in China* technical text it worked only 9 times out of 20 (45%), but it did better on the highly colloquial *Wheels* – 10 times out of 12 (83%). The fact that this heuristic worked better on colloquial texts featuring predominantly dialogues was not surprising, this genre being closer to the genre covered by Winograd's system which focused on maintaining dialogues.

Hobbs's algorithm worked in 88.3% of the cases. The version employing selectional constraints worked 91.7% of the time. Hobbs commented that these success rates were somewhat deceptive since in more than half of the cases there was only one plausible antecedent. For that reason, he separately calculated the success rate of the algorithm on the examples in which more than one plausible antecedent occurred in the candidate set. Of 132 such examples, 12 were resolved by selectional restrictions, and 96 of the remaining 120 were resolved by the algorithm. Thus, 81.8% of these 'conflicts' were resolved by a combination of the algorithm and the selectional restrictions.

Hobbs concluded that whether the success rate was 92%, 91.7% or 81.8%, the results showed that the naïve approach was very good. He expressed the view that 'it will be a long time before a semantically based algorithm is sophisticated

enough to perform as well', and correctly pointed out that 'these results set a very high standard for any other approach to aim for' (Hobbs 1976).

In its original form, Hobbs's algorithm was simulated manually.[14] As a consequence, it operated on 'perfectly' analysed sentences and the success rates of 88.3% and 91.7% given by Hobbs should be regarded as ideal. An anaphora resolution program would have certainly added some errors due, for instance, to incorrect syntactic analysis, lexical (POS) tagging or named entity recognition, and thus could have possibly degraded the success rate.[15]

Jerry Hobbs's approach remains one of the most influential works in the field and frequently serves as a 'classical' benchmark for evaluating current proposals (e.g. Baldwin 1997; Mitkov 1998a; Walker 1989). Recently some researchers (Tetreault 1999)[16] have implemented the algorithm with a view to carrying out comparative evaluation.

4.6 The BFP algorithm

The BFP algorithm for pronoun resolution (Brennan et al. 1987; Walker 1989) stemmed from Brennan, Friedman[17] and Pollard's extended centering model (see also Chapter 3, section 3.1). The extension of the original centering framework proposed in Grosz et al. (1986) consisted of fine-tuning the transitions in centering.[18] Brennan and colleagues distinguish between *smooth-shift*[19] and *rough-shift*.[20] *Smooth-shift* occurs when the center Cb(U_{N-1}) shifts to a new center Cb(U_N) but the backward-looking center Cb(U_N) is the same as the preferred center Cp(U_N). In contrast, *rough-shift* arises when the center Cb(U_{N-1}) changes to a new center Cb(U_N) with the backward-looking center Cb(U_N) being different from the preferred center Cp(U_N). *Rough-shift* is claimed to be less coherent than *smooth-shift*. In both cases the speaker shifts the center to a different discourse entity but while in the *smooth-shift* transition he/she indicates the intention to continue talking about the shifted-to entity (by realising this entity in a highly ranked Cf(U_N) position such as subject), no such intention is signalled in the *rough-shift* transition. Transition states are ordered: *continue* is preferred to *retain* which is preferred to *smooth-shift*, which in turn is preferred to *rough-shift*.

The BFP algorithm adds the so-called 'contra-indexing' constraints to the centering framework. These syntax constraints are similar to the ones based on the notions of c-command and minimal domain described in section 3.2 and are adopted from an earlier work by Reinhart (1976). The authors illustrate these constraints by the example *He likes him* where the pronouns *he* and *him* cannot be coreferential. The algorithm assumes that comprehensive syntactic analysis can compute whether these constraints hold and also that parsing can identify the syntactic functions of subject, object and indirect object, which play an important role in ranking the preferred center.

Another assumption the algorithm makes is that it is possible to structure both written texts and task-oriented dialogues in segments.[21] To this end, the authors propose a procedure using criteria such as orthography, distribution of

anaphors, cue words and task structure. For instance, they assume that in published texts a paragraph is a new segment unless the first sentence contains a pronoun in subject position or the paragraph contains a pronoun with which none of the preceding internal noun phrases agrees.

The algorithm consists of three basic phases.

1. Generate possible Cb–Cf[22] combinations (pairs).[23]
2. Filter by constraints (rules).
3. Rank by transition orderings.

To start with, the referring expressions are identified and ordered by grammatical function (e.g. subject, object, etc.) in U_N. Then a set of possible pairs of lists of forward-looking centers Cf and backward-looking centers Cb is generated.[24]

The second phase of the algorithm applies three filters to each Cb–Cf pair. If a pair passes through the filters, it is still under consideration, otherwise it is removed from the list of Cb–Cf combinations. The first filter checks for 'contra-indexing'. If a referring expression in a Cb–Cf pair is proposed to be resolved to a discourse entity with which it is contra-indexed, then this pair is removed. The second filter uses the constraint that 'Cb(U_N) is the highest-ranked element of Cf(U_{N-1}) that is realised in U_N'. For example, if the proposed Cb of the pair does not equal the first element on its Cf(U_{N-1}) list, then this pair is rejected. The third filter applies the rule that 'if some element of Cf(U_N) is realised as a pronoun in U_N, then so is Cb(U_N)' (see also Chapter 3, section 3.1). Therefore, if none of the pronouns in the proposed Cf(U_N) equals the proposed Cb, then the pair is eliminated.

In the third phase each remaining pair is classified as one of the transitions: *continuing*, *retaining*, *smooth-shift* and *rough-shift* by taking U_{N-1} to be the previous utterance and U_N to be the utterance currently being worked on. Finally, the pairs Cb–Cf are ranked on the basis of preference in the above order.

The authors illustrate their algorithm on the following discourse:

(4.25) Brennan drives an Alfa Romeo.
(4.26) She drives too fast.
(4.27) Friedman races her on weekends.
(4.28) She often beats her.

More details as to how the possible Cb–Cf combinations (pairs) are constructed, filtered, and ranked can be found in Brennan et al. (1987). The preference of *Friedman* over *Brennan* for *she* in utterance (4.28) is due to the fact that *smooth-shift* (with this transition in $U_{4.28}$ *she* would be referring to the new center *Friedman*[25] and in this case Cb($U_{4.28}$) = Cp($U_{4.28}$) = Friedman) is favoured over *rough-shift* (with this transition in $U_{4.28}$ *she* would be referring to *Brennan* and in this case Cb($U_{4.28}$) ≠ Cp($U_{4.28}$) = Brennan).

In a later work, Walker (1989) evaluated the BFP algorithm and compared its performance with Hobbs's naïve algorithm. The evaluation was based on a hand simulation of both algorithms which implies that both algorithms operated in an 'ideal environment' and were free from any pre-processing errors. Three types of data were used to analyse the performance of the BFP algorithm. Two of the

samples were those used previously by Hobbs to evaluate his algorithm: an excerpt from Arthur Hailey's novel *Wheels* and the 7 July 1975 edition of *Newsweek* (see 4.5.2) – each containing 100 pronouns. The third sample was a set of five human–human, keyboard-mediated and task-oriented dialogues about the assembly of a plastic water pump which contained 81 occurrences of *it* and no other anaphoric pronouns (Cohen 1984). The BFP algorithm resolved correctly 90 pronouns from the novel, whereas Hobbs's algorithm succeeded in 88 cases. The naïve algorithm outperformed the BFP on the *Newsweek* text tracking down the correct antecedent for 89 of the pronouns, as opposed to the BFP proposing 79 correct antecedents.[26] Hobbs's algorithm had a slight edge over the BFP on the task-oriented dialogues too, with 51 correct outputs as opposed to 49.

Walker concludes that the comparison of the two algorithms on each dataset individually and an overall analysis of the three datasets combined does not suggest any significant difference in the performance of the two algorithms.[27]

Walker's extensive evaluation covers error chaining analysis,[28] analysis of the performance of both algorithms on each type of anaphoric pronoun (*he*, *she*, *it* and *they*), error analysis[29] of each algorithm and an analysis of the cases in which both algorithms fail (for more details, see Walker 1989). She discovered that every case in which Hobbs's algorithm successfully obtained the correct antecedent, but the BFP did not, could be attributed to Hobbs's favouring of intrasentential antecedents. With a view to improving the BFP, she proposed a potential modification based on Carter's extension of Sidner's algorithm for local focusing. Carter (1986) argued that intrasentential candidates should be preferred over candidates from previous sentences only in the cases where no discourse center[30] has been established or where the discourse center is rejected for syntactic or selectional reasons. The addition of Carter's rule to BFP would raise the number of correctly resolved anaphors to 93 in the *Wheels* sample, to 84 in the *Newsweek* text and to 64 in the task-oriented dialogues, which would represent a significant improvement.

The BFP has been extensively cited in the anaphora resolution literature and has been used on a number of occasions as a benchmark for comparative evaluation (e.g. Tetreault 1999).

4.7 Carter's shallow processing approach

Carter describes in his Ph.D. thesis and later in his book (see Carter 1986, 1987a, 1987b) a 'shallow processing' approach which exploits knowledge of syntax, semantics and local focusing as heavily as possible without relying on large amounts of world or domain knowledge.[31] His algorithm is restricted to nominal anaphora.

Carter's approach is implemented in a program called SPAR (Shallow Processing Anaphor Resolver) which resolves anaphora[32] in simple English stories and generates sentence-by-sentence paraphrases corresponding to the interpretations selected. The program combines and develops several existing theories, most notably Sidner's (1979) theory of local focusing and Wilks's

(1975a) 'preference semantics' theory of semantics and common-sense inference. Carter describes SPAR as a Sidnerian anaphor resolver which uses Wilksian semantics and common-sense inference (CSI) to do the job of Sidner's 'normal mode' and 'special mode' inference respectively.[33] The result is one of the highest success rates obtained by anaphora resolution programs so far. In fact, SPAR's performance supports Carter's *shallow processing hypothesis*:

> A story processing system which exploits linguistic knowledge, particularly knowledge about local focusing, as heavily as possible and has access only to limited quantities of world knowledge which it invokes only when absolutely necessary, can usually choose an appropriate antecedent for an anaphor even in cases where the common-sense inference mechanism by itself cannot do so.
>
> (Carter 1987b: 238)

SPAR works on initial sentence interpretations produced by Boguraev's (1979) English analyser – a system that employs syntactic knowledge encoded as an augmented transition network and a modified form of Wilksian semantics. This analyser resolves most word senses and structural ambiguities but does not handle anaphoric ambiguities.

SPAR resolves the anaphors in the dependency structures and, while doing so, it resolves any remaining word sense or structural ambiguity. When a sentence has been fully processed (including the resolution of anaphoric reference), a paraphrase is produced. For instance, the sentence

(4.29) John promised Bill that he would mend his car.

is paraphrased after anaphora resolution as

(4.29a) John promised Bill that John would mend Bill's car.

SPAR acts on the dependency structure(s) as follows. First, the semantic formula for each word sense in a dependency structure is matched with the surrounding parts of the structure. This provides a measure of 'semantic density' (strong agreement is a ground for preferring a reading associated with it) and constrains the semantic ranges of pronouns. As an illustration, the formula for *drink* specifies a liquid object, so in the sentence *He drank it*, the anaphor *it* would be restricted to match only a liquid antecedent. Note that the semantic formulae trigger selectional restrictions as defined in 2.2.3.1.

Next, Sidner's pronoun interpretation (PI) rules are applied to each pronoun in a sentence while other focus-based rules are applied to lexical noun phrase anaphors. The PI rules normally propose a single candidate antecedent for each pronoun, according to the contents of a set of focus registers which have been set during processing of earlier sentences.[34] If the proposed candidate passes agreement filters, it is matched with the pronoun, using Wilksian semantic formulae (and any restrictions imposed in the first stage of processing). Carter explains that this matching corresponds roughly to invoking Sidner's 'normal mode' inference, since most contradictions resulting from temporary binding take the form of semantic clashes. If the match succeeds, a firm prediction that the pronoun and candidate corefer is returned by the PI rules. Otherwise the rules suggest other candidates.

If the PI rules propose more than one candidate, each of them is matched semantically with the pronoun. If several survive, CSI is not invoked immediately as in Sidner's original framework; instead alternative predictions are returned which are to be adjudicated later.

The original PI rules do not explain how or when candidates from the same sentence as the pronoun should be considered. Carter alleviates this problem by augmenting the focus registers with intrasentential candidates, ordered approximately as specified by Hobbs's algorithm, and the PI rules can then pick them up as they do contextual candidates. The consequence is that there are fewer cases when only one antecedent is proposed. It becomes more common for several candidates to be suggested together, but as explained above, in such cases, CSI is not invoked immediately.

After applying the PI to each anaphor in the sentence, configurational constraints (similar to the local domain constraint in 3.2.2) are employed to discount the inconsistent predictions. As Carter points out, this may remove the need for invoking CSI. As an illustration, consider (4.30):

(4.30) I took my dog to the vet on Friday. He bit him on the hand.

The PI rules, together with the semantic matcher, predict that *he* can be either the dog or the vet, whereas *him* can only be the vet (since *hand* is defined as part of a person, not of a dog). The configurational constraints bar *he* and *him* from coreferring and SPAR concludes that since *him* refers to *the vet*, the *he* can only be *the dog*. This example shows that it is not always necessary to invoke CSI when the PI rules suggest two plausible candidates.

If configurational constraints detect a clash between two firm predictions, the PI rules are reapplied so that further plausible candidates can be found.[35] CSI is only called upon if some anaphors remain unresolved after the application of configurational constraints. If CSI still cannot propose antecedents, then three 'weaker' heuristics are activated. Carter reports that even though there are many counterexamples, these heuristics usually point to the correct interpretations when they apply. When they do not apply, other, still weaker preferences associated with Sidner's PI rules are employed.

The first and most useful heuristic is that 'repetitions' should be preferred. For instance, if a pronoun and one of its remaining candidates have the same role in two semantically similar events in the story, that candidate is preferred. The second heuristic favours interpretations in which the discourse focus (as defined by Sidner's rules) remains unchanged. The third heuristic prefers NPs which c-command the pronoun. The usefulness of these heuristics is evident from examples (4.31)–(4.39) below, taken from Carter (1987b).

(4.31) John promised Bill that he would mend his car.
(4.32) He took it to his friend's garage.
(4.33) He tried to persuade his friend that he should lend him tools.
(4.34) His friend said that he was not allowed to lend tools.
(4.35) John asked his friend to suggest someone from whom he could borrow tools.

(4.36) His friend did not answer.
(4.37) Fulfilling his promises was important to John.
(4.38) He was angry.
(4.39) He left.

In sentence (4.31) neither of the anaphoric pronouns can be resolved easily and CSI is needed to choose between the candidates *John* and *Bill*. The correct choice is now made using rules that people tend to make promises about their own deliberate actions rather than other people's, and that people tend to want their own possessions to work. In (4.32), *his* is resolved without CSI, since PI rules and semantic matching recommend *John* without doubt. CSI is now invoked to select between *John*, the actor focus[36] and *Bill*, the potential actor focus, as candidates for *he*. It makes use of the formula for *garage*, which says that a garage is a place where people mend things, and decides that *he* is taking *it* to the garage so that someone can mend *it*. Both *John* and *Bill* are predicted as antecedents of *he* since both of them are expected to want the car to work (John having made a promise and Bill being the owner of the car); *it* is bound to *the car*. The third weak heuristic (preferring pronouns to be c-commanded by coreferring phrases) correctly selects John as the antecedent.

According to the PI rules in (4.33) both occurrences of *he* and the *him* are ambiguous between *John* and *his friend*. The first occurrence of *he*, however, is resolved to *John* because the configurational constraints block the alternative. These constraints also forbid the second *he* and *him* to corefer, but since both pronouns are still ambiguous, no alternatives can be ruled out. Now CSI is invoked but at this stage no reasoning is performed that indicates that *him* is *John* because John is likely to want tools to mend the car. Instead, CSI simply binds *him* to the first *he* using a shallower, more general CSI stating that people are more likely to possess things themselves rather than to want other people to possess them. Since the first *he* is *John*, *him* is also set to *John* and therefore the second *he* is identified as *his friend* after applying configurational constraints again. Therefore, in this sentence focusing (incorporated in the PI), CSI and syntax are all vital for the correct resolution of pronouns.

In the remainder of the story, CSI fails to provide any solution. However, the repetition heuristic is helpful here. This heuristic recognises that since sentence (4.33) mentions the friend lending John tools, *he* in (4.34) is the friend and *he* in (4.35) is John (on the basis of the obvious semantic relationship between borrowing and lending). Also, it realises that *his* in (4.37) is John and not the friend, as John was mentioned as making promises in (4.31) and the friend is not associated with any promises.

Examples (4.31)–(4.39) show that even though SPAR does not use large amounts of world or domain information, knowledge of syntax, semantics, local focusing and common-sense inference are exploited as heavily as possible.

SPAR was tested on 60 stories covering a variety of topics. The stories were grouped in two categories. The first category consisted of 40 texts, of two or three sentences each, which were specially written/selected to test SPAR. All of the 65 pronouns of this category were correctly resolved.

The second category consisted of over 20 stories written by people with little or no knowledge of SPAR's way of working; many of these texts were originally written for other language-processing systems. These stories were on the average 9 sentences long, the longest being 23 sentences. SPAR resolved 226 out of the 242 pronouns (93%).[37] Carter (1986) points out that this figure could go up to 96% (232 correctly resolved anaphors) if an error recovery procedure were implemented.

Carter noted that the contribution of CSI to this performance was in only 29 (12%) of the cases when CSI inference chains were used (each time correctly) to propose the antecedent. On many other occasions CSI inference chains were formed but they either confirmed the decisions already made or were rejected as incompatible with the predictions of the other components of the system.

4.8 Rich and LuperFoy's distributed architecture

Elaine Rich and Susann LuperFoy (1988) describe the pronominal anaphora resolution module of Lucy, a portable English understanding system. The preprocessing is done by a parser which generates a feature graph representing the syntactic properties of the constituents of the sentence and by a semantic processor which produces as its output a list of discourse referents and a set of assertions about them. The anaphora resolution module augments the assertion set with additional assertions regarding coreference relations between discourse referents. For instance, the semantic processing of the simple discourse

(4.40) *Dave* created *a file*. *He* printed *it*.

identifies *create* and *print* as predicates of each of the sentences, the discourse referents *Dave* and x_1(= He) as agents and the discourse referents *a file* and x_2(= it) as patients.[38] The job of the anaphora resolution module is to establish that *Dave* and *he*, as well as *a file* and *it*, are coreferential.

The authors explain that the design of their pronoun resolution module is motivated by the observation that even though 'there exists no single, coherent theory upon which an anaphora resolution system can be built, there are many partial theories each of which accounts for a subset of phenomena that influence the use and interpretation of pronominal anaphora' (Rich and LuperFoy 1988). In line with this observation, Rich and LuperFoy encode each 'partial theory' as a separate module into a 'distributed architecture' designed to cover a wide range of pronominal anaphora cases (Figure 4.4). These modules interact to propose candidate antecedents and to evaluate each other's proposals; an oversight module, called the *handler*, mediates and resolves differences in opinion. According to the authors, the ovals in the figure 'represent an implementation of one of the partial theories' and are referred to as *constraint sources* since each one of them can be viewed as imposing a set of constraints on the choice of the antecedent. Note that the modules (constraint sources) correspond roughly to the *factors* introduced in section 2.2.3.[39]

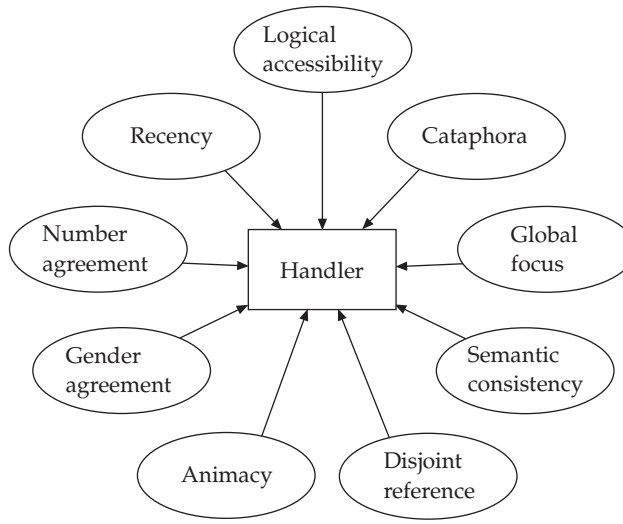

Figure 4.4 Rich and LuperFoy's 'distributed architecture'.

One of the important contributions of Rich and LuperFoy's work is their analysis as to how factors can interact and influence the decision on the antecedent. In their algorithm the selection of the antecedent from among a set of candidates is made on the basis of a combined score resulting from the examination of each candidate by the entire set of factors.

The initial implementation uses a score between –5 and 5 for each factor, with the handler averaging the individual scores to form a composite score. The authors explain that there is a drawback in this scoring strategy in that there is no way to account for factors which 'have no opinion'. Also, the initial scoring procedure does not allow for factors which have an opinion but are very (or not at all) confident of it. To remedy these problems, Rich and LuperFoy propose a scoring formula in which each factor provides both a score and a confidence measure. The score is a number in the range –5 to 5, the confidence number is in the range 0 to 1, and the function which combines a set of n (score, confidence) pairs is:

$$\text{running score} = \frac{\sum_{i=1}^{n} \text{score}(i) \times \text{confidence}(i)}{\sum_{i=1}^{n} \text{confidence}(i)}$$

This function computes an average which is weighted not by the number of distinct scores, but by the total confidence expressed for the scores. A factor that wishes to offer no opinion can simply suggest a confidence of 0 to its opinion which, in turn, will have no effect on the running score of a candidate.

The factors implemented are classified by Rich and LuperFoy (1988) as falling into one of the following four categories.

1. *Finite set generators* are factors which propose a fixed set of candidates. They assign all candidates the same score, the latter being a function of the number of competing candidates. An example of such a factor is *disjoint reference* (see further below):

Number of candidates to propose	*Score*	*Confidence*
1	5	1
2	4	1
3	3	1

   These factors never evaluate: when they are asked to do so, they return a confidence of 0.

2. *Fading infinite set generators* are factors that can keep on proposing the same candidates, but with lower scores as the text progresses. As an illustration, *recency* is such a type of factor:

Sentence	*Score*	*Confidence*
n (current)	1	0.5
n – 1	2	0.5
n – 2	0	0.5

   These factors, similar to the finite set generators, never evaluate.

3. *Filters* are factors which never propose candidates. They only filter out candidates which do not satisfy specific (almost obligatory) conditions. Examples of filters are the requirements for gender and number agreement between the anaphor and the antecedent. Filters use the following values for score and confidence when evaluating candidates:

	Score	*Confidence*
pass	0	0
fail	–5	0.9

   Pass means that since the confidence level is 0, the score does not matter and does not have any effect on the overall decision regarding this candidate: the latter has passed the test, but has not been given any special (preferential or non-preferential) treatment. The candidate's score is insensitive to the number of filters it passes and the evaluator will be called to make the final decision. *Fail* means that a candidate has not passed the test for conforming to specific requirements; its composite score will drop below the minimum threshold and will eventually be eliminated from any further consideration.

4. Finally, preferences such as *semantic content consistency* (see below) are factors which impose preferences rather than absolute opinions on a set of candidates. These factors are said to exploit the full range of (score, confidence) values.

Rich and LuperFoy admit that the scoring scheme is not perfect in that it does not capture cases where numbers are used to represent uncertainty. A few years on, an uncertainty reasoning approach (Mitkov 1995b) is independently proposed which regards the factors' values as uncertainty.

The following factors[40] are implemented in Lucy: recency, number agreement, gender agreement, animacy, disjoint reference, semantic type consistency, global focus, cataphora, logical accessibility, local focus, rhetorical structure, set generation and rhetorical 'they'.[41] *Recency* proposes candidates occurring in the recently preceding discourse but has no opinion with regard to proposals from other factors. *Number agreement* knows that anaphors and antecedents should match in number. This factor does not propose any antecedents; it only acts as a filter on candidates proposed by other factors. *Gender agreement* functions similarly to number agreement, filtering candidates on the basis of the obligatory gender agreement between the antecedent and the anaphor. *Animacy*, which also serves as filter, knows that neuter pronouns refer to inanimate things, whereas masculine and feminine pronouns refer to animate things. *Disjoint reference* makes use of syntax-based restrictions which apply to reflexive and non-reflexive pronouns (Reinhart 1983a; see also sections 3.2.1 and 3.2.2). This factor proposes antecedents for reflexive pronouns as in the sentence '*George* saw *himself*', but functions as a filter for non-reflexive pronouns discounting coreference in sentences such as 'George saw him'. *Semantic type consistency* acts as a filter and restricts antecedents only to candidates which satisfy the type constraints imposed by the semantically acceptable interpretation of the sentence. As an illustration of this factor, Rich and LuperFoy offer the discourse 'The system created an error log. It printed it.' Assuming that the interpretation of *print* imposes the following type constraints[42] on the semantic roles agent and patient[43]:

agent: human $\vee$ computer
patient: information structure

this factor would reject *an error log* as the antecedent of the first occurrence of *it* given that the type hierarchy does not include *log* as a subclass of either *human* or *computer*. This factor would discount *the system* as the antecedent of the second occurrence of *it* since the type hierarchy does not feature *system* as a subclass of *information-structure*.

Among the other factors employed is *global focus* which proposes as antecedents discourse entities that are in global focus. *Cataphora* knows about a class of syntactic constructions in which a pronoun can precede the full lexical NP with which it corefers. This factor will propose *George* as a candidate antecedent for *he* in the sentence 'When *he* is happy, *George* sings'. The *logical accessibility* factor imposes constraints on the accessibility of referents such as function quantifiers and negation (Kamp 1981; see also section 3.3, Discourse Representation Theory) and would rule out *a donkey* as the antecedent for *it* in the sentence 'If a farmer doesn't own a donkey, he beats it'. *Semantic content consistency* exploits semantic knowledge about context-dependent phenomena as opposed to simply applying 'static' type hierarchy constraints. Rich and LuperFoy say that the boundary between semantic type consistency and semantic content consistency is fuzzy,[44] the key difference being that while accessing a type hierarchy is fast, there are cases in which applying semantic content consistency would need a lot of reasoning. Therefore, this factor appears to be very similar to the real-world/common-sense knowledge factor which can be

illustrated by the example: '*Your car* is parked next to a fire hydrant. You'll have to move *it*.' Even though the mention of *hydrant* is more recent, the antecedent of the pronoun *it* is *your car* because common-sense reasoning gives a higher likelihood for cars to be movable. *Local focus* tracks objects which are locally in focus in the discourse and *rhetorical structure* segments and organises the discourse as a set of plans for fulfilling conversational roles. Another factor used is *set generation* which would create sets of referents acting collectively as antecedents for plural pronouns. For instance, this factor would propose *George and Elitza* as the antecedent for *they* in the discourse '*George* phoned *Elitza*. *They* had a long chat'. Finally, the *generic they* factor knows about salient individuals or groups and proposes *them* as the referent of *they* in sentences such as 'Why don't they ever fix the roads?'[45]

The implementation of the anaphora resolution program includes tracing tools which display information such as which NPs are recognised as anaphors, which constraint sources (factors) are consulted and in what order, and what effect each factor has on the overall rating assigned to each proposed antecedent. During test runs this information could be very helpful to the developers with a view to improving the algorithm further.

Rich and LuperFoy's paper does not report any evaluation results.

4.9 Carbonell and Brown's multi-strategy approach

Carbonell and Brown's main philosophy, like that of Rich and LuperFoy, adheres to the principle that an integrated approach exploiting different knowledge sources performs better than a monolithic method. They propose a general framework for intersentential anaphora resolution based on a combination of multiple knowledge sources: sentential syntax, case-frame semantics, dialogue structure and general world knowledge (Carbonell and Brown 1988). These are expressed in various constraints or preferences which are used in the resolution process.

Constraints employed relate to local syntax agreement, case-role semantics and pre-conditions/post-conditions. *Local anaphor constraints* basically check if the candidates match the anaphors in gender and number (see also section 2.2.3.1) and eliminate those that do not. *Case-role semantic constraints* require that if fulfilled by the anaphor, these constraints should be satisfied by the antecedent too; candidates which violate constraints on the case role[46] occupied by the anaphor are eliminated from further consideration. These constraints, which correspond to the 'semantic type consistency' filter used in Rich and LuperFoy (1988 – see previous section) and are also known as 'selectional restrictions' (see 2.2.3.1), would discount *the table* (tables are not edible) from being the antecedent in (4.41) and would rule out *the cake* (cakes are not washable) as antecedent in (4.42).

(4.41) John took *the cake* from the table and ate *it*.
(4.42) John took the cake from *the table* and washed *it*.

Pre-condition/post-condition constraints use real-world knowledge and pragmatics and apply to cases where a specific candidate cannot be the antecedent of an anaphor because some action occurring between the candidate and the anaphor invalidates the assumption that they denote the same object or event. These constraints are exemplified by (4.43):

(4.43) George gave *Martin* an apple. *He* ate the apple.

Here *he* refers to *Martin*, as *George* no longer has the apple. The post-condition on *give* is that the agent no longer has the object being given. This conflicts with the precondition on *eat* that the agent has the item being eaten, if it is assumed that the agent is *George*.

These constraints eliminate from consideration all candidates involved in actions whose post-conditions violate the pre-conditions of the action containing the anaphor. As Carbonell and Brown note, simple though it is, the pre-condition/post-condition strategy requires a huge amount of knowledge to be successful for a wide range of cases.

The preferences used are semantic parallelism, semantic alignment, syntactic parallelism, syntactic topicalisation and intersentential recency. *Semantic parallelism*[47] is applied to candidates which satisfy all constraints and favours NPs from an earlier utterance which fill the same semantic case role as the anaphor as in (4.44):

(4.44) Elitza gave a sweet to *Tina*. George also gave *her* a chocolate.

In this example both *Tina* and *her* map into the same semantic case, *recipient*.

Carbonell and Brown exemplify their *semantic alignment* preference by the following discourses:

(4.45) Elitza drove from the park to *the club*. George went *there* too.
(4.46) Elitza drove from *the park* to the club. George left *there* too.

The second sentence in discourse (4.45) 'aligns semantically'[48] with the 'destination (goal) part' of the first sentence, whereas in (4.46) the second sentence 'aligns' with the source part of the first sentence. This preference is similar to semantic parallelism but in addition states that if the clause in which the anaphor is located 'aligns semantically' with a previous clause or with part of a previous clause, candidates from that previous clause should be searched first for antecedents.

The *syntactic parallelism* preference plays an important role if two clauses are directly contrasted in a coordinate structure or by means of explicit discourse markers. As an illustration of syntactic parallelism,[49] consider again examples (2.40) and (2.41) from Chapter 2.

(2.40) The programmer successfully combined *Prolog* with C, but he had combined *it* with Pascal last time.
(2.41) The programmer successfully combined Prolog with *C*, but he had combined Pascal with *it* last time.

This factor searches for coordinated clauses, adjacent sentences or explicitly contrasted sentences and prefers the candidate that preserves the syntactic function

of the anaphor. For instance in (2.40) both the anaphor *it* and the antecedent *Prolog* are direct objects.

The *syntactic topicalisation* preference favours topicalised candidates and proposes them as antecedents if they do not violate any constraint. The syntactic topicalisation is indicated through linguistic devices such as fronting (*As for Alexander . . .*) and cleft constructions (*It was Alexander who . . .*).

The *intersentential recency* preference advocates searching sentences in reverse chronological order. If there are no good candidates in the previous sentence, then the one before that is considered, and so on.

Carbonell and Brown distinguish between *constraints,* which cannot be violated, and *preferences,* which discriminate among candidates satisfying all constraints. They propose that the latter be ranked in partial order as goal trees (Carbonell 1980), or be offered a voting scheme where the stronger preferences are assigned more votes. Conflicting preferences of equal voting power indicate true ambiguity.

The resolution strategy applies the constraints first with a view to reducing the number of candidates for antecedent. Next, the preferences are applied to each remaining candidate. If more than one preference applies and favours different candidates, then the anaphor is considered to have an ambiguous antecedent. Note that this approach is different from robust approaches such as Baldwin (1997) or Mitkov (1996, 1998b) which always propose the most likely antecedent on the basis of rules or aggregate scores.

The practical implementation of Carbonell and Brown's anaphora resolution framework includes local constraints, semantic constraints, pre-/post-condition constraints, semantic parallelism, intersentential recency preference and syntactic topicalisation preference. The implementation is part of the Universal Parser (UP) project at the Centre for Machine Translation, Carnegie-Mellon University. The UP uses a modified version of lexical-functional grammar which employs syntactic and semantic knowledge sources to produce a full analysis of each sentence. The input to the anaphor resolver is a set of semantic roles and a syntactic tree associated with each sentence. The noun phrases extracted from the most recent sentences serve as candidates for antecedents. Each preference is given an individual weight, but the latter is not specified. In addition to eliminating candidates, semantic and local anaphora constraints may cast votes for eligible candidates most closely matched to the anaphor and can trigger preference in the absence of hard constraints. For example, the gender constraint would prefer a candidate of feminine gender over indeterminate gender when resolving a feminine gender anaphor while ruling out all candidates of masculine gender. After applying all preferences, the most preferred candidate adopts the gender of the anaphor to restrict further searches. For example, if *she* was established to refer to *doctor*, all future anaphoric references to *doctor* would have to be feminine or neuter.[50]

In addition to resolving personal pronouns whose antecedents are noun phrases, Carbonell and Brown's approach handles lexical NP anaphors which refer to noun phrases. The heads and the modifiers of the candidate NPs are checked for agreement with the lexical anaphor. One rule used is that the head

noun of the candidate must be the same as the head noun of the anaphor. For the remaining modifiers of the candidate it suffices that they are present as modifiers of the anaphor or simply missing. Note that the requirement for the head nouns of the anaphor and the antecedent to be the same would fall short of successfully tackling lexical NP anaphors whose head is different from that of the antecedent but represents a semantically close concept, as in the case of synonyms or superordinates (see section 1.4.2).

Evaluation was reported on a sample of 31 anaphors of which 27 were pronouns and 4 lexical NP anaphors. The program correctly resolved all but four of the anaphors, yielding a success rate of 87%. However, it must be borne in mind that this is a very small sample and further evaluation is needed for more definitive results.

4.10 Other work

There is much more work which regrettably cannot be discussed in detail owing to limitations of space. Among the early research not covered explicitly, **Charniak's thesis** (Charniak 1972) deserves special attention. Even though this work does not offer a solution or implementation, it does show how complex pronoun resolution can be. Charniak's work points to a wealth of difficult cases from the domain of children's stories which demonstrate that arbitrarily detailed world knowledge may be required to decide upon an antecedent. **Wilks's preference semantics** (1973, 1975a) approach[51] uses, among other (more sophisticated) devices, knowledge of individual lexeme meanings in order to successfully solve cases such as 'Give the bananas to the monkeys although they are not ripe, because they are very hungry.' In this example each *they* is interpreted correctly using the knowledge that since the monkeys are animate, they are likely to be hungry, whereas the bananas, being fruit, are likely to be ripe or unripe.

Kantor (1977) investigates the problem of why some pronouns in discourse are more comprehensible than others, even when there is no ambiguity or anomaly. He defines the notion of *activatedness* of a concept: the more activated a concept is, the easier it is to understand an anaphoric reference to it. The notion of activatedness is very close to that of *focus* proposed by **Grosz** (1977a, b; see also Chapter 3, section 3.3). **Webber** (1979) applies a set of rules to a logical-form representation of the text to derive a set of entities available for subsequent reference. Webber's formalism attacks problems caused by quantification[52] which were not previously considered.

Sidner's focusing approach to interpretation of definite anaphora (Sidner 1979, 1983) resolves full definite noun phrases and definite pronouns on the basis of the focus state, as defined by the six focus registers (see section 3.3). The rules assume the existence of hierarchical/associative knowledge representation which provides for generic nodes, representing classes of objects or events. Sidner describes partial implementations of her algorithm in PAL (Personal Assistant Language Understanding Program), which was part of the PA

(Personal Assistant) project at MIT, and in TDUS (Task Discourse Understanding System) at SRI. **The PUNDIT text understanding system** for limited domains (Dahl 1986) also uses a simplified version of Sidner's algorithm with no actor focus and a single ordered focus list rather than separate current, potential and stacked foci. The algorithm is applied to pronouns, elided noun phrases, associative anaphors and 'one' anaphors.

Günther and Lehmann's (1983) rules for pronoun resolution operate in the restricted context of relational database query dialogues.[53] Their system constructs a DRS[54] for a dialogue and applies morphological, configurational (syntactic), semantic and pragmatic factors, in that order, to accessible candidate antecedents until only one candidate remains. The morphological rules (referred to as 'criteria') test for agreement of gender and number. The configurational criteria 'specify the concrete syntactic configurations where disjoint reference holds'; the c-command criterion is rejected as 'too strict' in some cases. The semantic criteria check mainly that the proposed antecedent does not give rise to a query which is anomalous in terms of the database relations. The pragmatic criteria are worth mentioning and are expressed in the following preferential rules (in order of application): (i) noun phrases in more recent sentences are preferred to less recent; those in the sentence containing the pronoun are most preferred (principle of proximity), (ii) pronouns are preferred to lexical noun phrases,[55] (iii) noun phrases in a matrix clause or phrase are preferred to noun phrases in embedded clauses or phrases, (iv) subject noun phrases are preferred to non-subjects, (v) accusative object noun phrases are preferred over non-subject NPs and (vi) anaphora is preferred to cataphora. Some of these preferences are similar to or the same as those used by other authors (e.g. (iv) and (v) are used in centering).

Rolbert's approach to resolution of pronouns in French (Rolbert 1989), implemented within an NL database query system, is based on syntactic, semantic and pragmatic factors. The syntactic factors include c-command (Reinhart 1983b) and a particular version of this syntactic relationship called *direct c-command* (Rolbert 1989). Different 'semantic' types of anaphora such as identity-of-sense anaphora and identity-of-reference anaphora[56] are dealt with at the semantic level using a logical representation of the discourse. If the pronominal anaphors cannot be resolved by the syntactic and semantic factors only, pragmatic criteria (such as criterion (iv) used by Günther and Lehmann) are called on to select the antecedent.

Other contributions during the 1970s and the 1980s worth mentioning include, but are not limited to, **Lockman's contextual reference resolution algorithm** (Lockman 1978) and **Asher and Wada's model** (Asher and Wada 1988) which employs syntactic, semantic and discourse factors.

4.11 Summary

The very early approaches to anaphora resolution used heuristics such as simple matching (Bobrow 1964) but also more elaborate ones which produced excellent

results for the time (Winograd 1972). Later work evolved as more theoretically-oriented and ambitious in terms of the types of anaphora handled. It typically resorted to extensive use of linguistic and non-linguistic knowledge (Carter 1986; Carbonell and Brown 1988; Rich and LuperFoy 1988; Sidner 1979; Wilks 1973) and was therefore of less practical value with the implementations being either limited or, in some cases, non-existent. As a consequence, the evaluation (if any) was carried out on a very small scale from the point of view of today's evaluation requirements. Nevertheless, the work on anaphora resolution in the 1960s, 1970s and 1980s is remarkable in that it addressed a number of fundamental issues and produced sophisticated models. Many of the approaches developed (e.g. Hobbs 1976, 1978; Brennan et al. 1987) still serve as benchmarks and are extensively cited in the current literature.

Notes

1 In general, much of the NLP work in the 1970s and 1980s was inspired by various knowledge representation theories such as Minsky's frame theory (1975) and as a consequence, many systems were built on the assumption that the required (domain) knowledge can be encoded and subsequently accessed by the system.
2 Some of the outlines are based on Hirst (1981) and Carter (1987a).
3 It should be noted, however, that this rule does not always work. The following counter-example has been provided by Minsky (1968): *He put the box on the table. Because it wasn't level, it slid off*. Here the two 'adjacent' occurrences of *it* are not coreferential: the first *it* refers to *the table* and the second to *the box*.
4 A breadth-first search of a tree is one in which every node of depth N is visited before any node of depth $N + 1$.
5 Hobbs notes (personal communication) that numerals can function as NPs as in the example 'I shall be glad when *2000* is over. *It* has been the worst year of my life'.
6 Hobbs correctly points out that the utility of these constraints is limited. They cannot be discriminative for the pronoun *he* for instance, because what one male human can do, another can do too. Even with *it* the utility is limited. However, in the present example, such heuristics can help.
7 These were later improved and recast in the more precise formal terms of the binding theory (see Chapter 3, section 3.2.2).
8 As Hobbs points out, this constraint is not 100% precise and fails on a number of examples such as *John saw a picture of him*.
9 Compare with c-command, Chapter 3, section 3.2.
10 This constraint, too, is not perfect and will fail on examples such as *Girls who he has dated say that Sam is charming* (Ross 1967) where *he* and *Sam* are coreferential. For more precise (but still not 100% precise to date!) and modern treatment of these constraints see Chapter 3, section 3.2.
11 Hobbs's algorithm was able to handle possessives too.
12 Original wording: '*it* occurring in a time or weather construction' (Hobbs 1976, 1978).
13 By 'plausible antecedent' Hobbs means candidates encountered along the way as the algorithm traversed the parse trees.
14 In the early 1980s, however, Hobbs implemented the algorithm as part of the DIALOGIC parser at SRI but no evaluation was carried out at that time (personal communication, J. Hobbs).

15 See Chapter 7, section 7.4.1 for more discussion on this topic.
16 The algorithm has been implemented by Donna Byron; the program operates on fully annotated sentences (fully bracketed with labels for word-class and features) and therefore does not use a parser to generate full parse trees.
17 Known otherwise (and better) as Walker.
18 The (original) typology of transitions is based on two factors: whether or not the backward-looking center, C_B, is the same from U_{N-1} to U_N, and whether or not this entity coincides with the preferred center of U_N. See section 3.1 for a brief overview on centering.
19 Originally termed *shifting – 1*.
20 Originally termed *shifting*.
21 Note that centering is a local phenomenon and operates within a segment.
22 Here again, for reasons of brevity Cb and Cf denote Cb(U_N) and Cf(U_N). Note that Cb (the backward-looking center) is a single entity, whereas Cf (the forward-looking center) is a list of entities.
23 These combinations are referred to as *anchors* (Brennan et al. 1987).
24 As Tetreault (1999) points out, the number of these combinations may be very high which could make the filtering phase very time-consuming.
25 Note that $U_{4.27}$ features *retaining* which signals an impending center shift.
26 The success rates reported in this evaluation varied slightly from those published in Hobbs (1976, 1978). Walker conjectured that this was probably due to a discrepancy in exactly what the dataset consisted of.
27 Note, however, that Hobbs beats the BFP on the *Newsweek* text by a comfortable margin.
28 Error chaining refers to the phenomenon of the algorithm's performing wrongly due to errors in preceding steps. Walker's analysis showed that error chains caused 22 failures of Hobbs's algorithm and 19 failures of BFP.
29 Analysis of the cases in which the algorithm performs incorrectly.
30 In the sense of Sidner (see section 3.3).
31 As Carter states, world and domain knowledge are notoriously hard to process accurately.
32 SPAR can resolve other linguistic ambiguities too.
33 In Sidner's theory pronoun interpretation rules are applied to each pronoun in a sentence independently of the others. The rules suggest candidate antecedents, normally one at a time, according to the contents of a set of focus registers (see Chapter 3, section 3.3) which have been set during processing of earlier sentences. If a candidate agrees syntactically with the pronoun, it is temporarily bound to it, and inference is invoked using semantic and common-sense knowledge. Sometimes, however, the rules suggest two or more candidates at once and when this is the case, inference is invoked in a 'special mode' to decide which candidate is most plausible.
34 See also section 3.3. The basic rule says that the focus should be suggested as the most probable antecedent.
35 Carter notes that this happens very seldom in practice.
36 See Sidner's definition of *actor focus*, section 3.3.
37 Direct comparison of the success rates of Hobbs's and Carter's approaches to pronoun resolution (note that Carter's approach tackles the broader class of nominal anaphora) is not possible since the results have been obtained on different texts which also probably differ in complexity. For further discussion on that topic and on evaluation issues in general, see Chapter 8.
38 The semantic (case) role *agent* is defined as the 'instigator of the action', whereas *patient* describes who/what is 'acted upon' or 'undergoing change' (see Dillon 1977 for a concise introduction to semantic roles). By way of illustration in the example 'John broke the window', *John* is the *agent* and *the window* the *patient*.

39 Therefore, in order to avoid confusion, the 'modules' ('constraint sources') are simply referred to as 'factors' henceforth.
40 Rich and LuperFoy refer to them as 'constraint sources'.
41 The last 6 factors have been implemented shortly after the submission of Rich and LuperFoy's 1988 paper (personal communication from Susann LuperFoy), hence in the paper these factors are referred to as 'envisioned but not yet implemented'.
42 These constraints are similar to the *selectional restrictions* or *case-role constraints* referred to by other authors.
43 Rich and LuperFoy refer to the semantic role *patient* as *object*.
44 See also note 23 of Chapter 2 which says that the distinction between 'semantic' and 'real-world' knowledge is unclear.
45 In fact, if taken in isolation, this example does not feature any antecedent of *they*.
46 Case roles (also termed *semantic roles*) used are agent (originally termed actor), patient (originally termed object), recipient, etc. (See examples below and also note 38 of this chapter.)
47 Referred to as 'case-role persistence' by the authors.
48 In terms of semantic roles: note that the role of *the club* and *there* (4.45) is that of *goal*; *the park* fills the role *source*.
49 It should be noted that syntactic parallelism and semantic parallelism overlap in the case when surface roles coincide with deep roles (e.g. the NP representing the subject represents also the agent or when the NP which is direct object is also patient, etc.).
50 This appears to be of limited utility and will not work in cases where multiple doctors are mentioned (comment Linda C. Van Guilder).
51 Developed for an English-to-French translation system.
52 As in the example 'Ross gave each girl a crayon. They used them to draw pictures of Daryel in the bath'. For examples of the quantifier structure of similar examples (2.14) and (2.15), see Chapter 2, section 2.1.3.
53 The authors express the hope that the rules might be extendable to other types of dialogue.
54 Discourse Representation Structure.
55 An entity referred to by a pronoun is more likely to be in focus.
56 See section 1.7 for definition of identity-of-reference anaphora.

CHAPTER FIVE

The present: knowledge-poor and corpus-based approaches in the 1990s and beyond

5.1 Main trends in recent anaphora resolution research

Much of the early work in anaphora resolution heavily exploited domain and linguistic knowledge (Carbonell and Brown 1988; Carter 1986, 1987a; Rich and LuperFoy 1988; Sidner 1979) which was difficult both to represent and to process, and required considerable human input. However, the pressing need for the development of robust and inexpensive solutions to meet the demands of practical NLP systems encouraged many researchers to move away from extensive domain and linguistic knowledge and to embark instead upon knowledge-poor anaphora resolution strategies. A number of proposals in the 1990s deliberately limited the extent to which they relied on domain and/or linguistic knowledge (Baldwin 1997; Dagan and Itai 1990; Kameyama 1997; Kennedy and Boguraev 1996; Mitkov 1996, 1998b; Nasukawa 1994; Williams et al. 1996) and reported promising results in knowledge-poor operational environments.

The drive towards knowledge-poor and robust approaches was further motivated by the emergence of cheaper and more reliable corpus-based NLP tools such as POS taggers and shallow parsers, alongside the increasing availability of corpora and other NLP resources (e.g. ontologies). In fact the availability of corpora, both raw and annotated with coreference links, provided a strong impetus to anaphora resolution with regard to both training and evaluation. Corpora, especially when annotated, are an invaluable resource not only for empirical research but also for automated or machine learning methods, and they also provide an important resource for evaluation of the implemented approaches. From deriving simple co-occurrence rules (Dagan and Itai 1990) through training decision trees to identify anaphor–antecedent pairs (Aone and Bennett 1995) to inducing genetic algorithms to optimise the resolution factors (Orasan et al. 2000; Mitkov et al. 2002), the performance of more and more modern approaches depends on the availability of large suitable corpora.

While the shift towards knowledge-poorer strategies and the use of corpora represented the main trends of anaphora resolution in the 1990s, there are other significant highlights in recent anaphora resolution research. The inclusion of the coreference task in MUC-6 and MUC-7 gave a considerable momentum to the development of coreference resolution algorithms and systems (Baldwin

et al. 1995; Gaizauskas and Humphreys 1996; Kameyama 1997[1]). The last decade of the twentieth century saw a number of anaphora resolution projects for languages other than English including French (Popescu-Belis and Robba 1997), German (Dunker and Umbach 1993; Fischer et al. 1996; Leass and Schwall 1991; Stuckardt 1996, 1997), Japanese (Mori et al. 1997; Murata and Nagao 2000; Nakaiwa and Ikehara 1992, 1995; Nakaiwa et al. 1995, 1996; Wakao 1994), Portuguese (Abraços and Lopes 1994) and Turkish (Tin and Akman 1994). Against the background of a growing interest in multilingual NLP, multilingual anaphora/coreference resolution has gained considerable momentum in recent years (Aone and McKee 1993; Azzam et al. 1998a; Harabagiu and Maiorano 2000; Mitkov 1999c; Mitkov and Stys 1997; Mitkov et al. 1998). Other milestones of recent research include the employment of probabilistic and machine learning techniques (Aone and Bennett 1995; Ge et al. 1998; Kehler 1997b; Cardie and Wagstaff 1999), the continuing interest in centering, used either in original or in revised form (Abraços and Lopes 1994; Hahn and Strube 1997; Strube and Hahn 1996; Tetreault 1999) and proposals related to the evaluation methodology in anaphora resolution (Mitkov 1998a, 2000, 2001b; Byron 2001).

In the following sections the approaches that have emerged as the most influential in current anaphora resolution research will be summarised. Whereas for practical reasons some of the main features of each approach (such as the factors employed and its evaluation) have been presented, the reader is encouraged to consult the original work for more details. Section 5.2 outlines Dagan and Itai's approach based on extracting collocation (co-occurrence) patterns from corpora, heralding a decade where corpus resources and corpus-based techniques took over from the less practical knowledge-dependent solutions. Following this, Lappin and Leass's RAP algorithm, which benefits from a Slot Grammar Parser to apply a powerful intrasentential filter, is presented in greater detail. The subsequent sections describe Kennedy and Boguraev's parser-free modification of RAP and Baldwin's knowledge-poor CogNIAC. Vieira and Poesio's work on definite descriptions is then outlined. Sections 5.7, 5.8 and 5.9 focus on recent machine learning, statistical and clustering trends summarising the machine learning approaches of Aone and Bennett, McCarthy and Lehnert, and Soon et al. as well as Ge and Charniak's statistical model and Cardie and Wagstaff's clustering algorithm. Section 5.10 briefly outlines other successful approaches that, due to space constraints, cannot be presented in any greater detail. The last section of the chapter discusses the growing importance of anaphora resolution for different applications in Natural Language Processing.

5.2 Collocation patterns-based approach

Ido Dagan and Alon Itai (1990, 1991) describe an approach for resolving third person pronouns based on collocation (or co-occurrence) patterns as an alternative solution to the expensive implementation of full-scale selectional restrictions. These patterns are collected automatically from large corpora and are used to filter out unlikely candidates for antecedent.

Table 5.1 Co-occurrence patterns associated with the verb *collect* based on an excerpt from the *Hansard* corpus

subject–verb	collection	collect	0
subject–verb	money	collect	5
subject–verb	government	collect	198
verb–object	collect	collection	0
verb–object	collect	money	149
verb–object	collect	government	0

Selectional restrictions used in anaphora resolution require that the antecedent must satisfy the constraints imposed on the anaphor (see section 2.2.3.1). In particular, if the anaphor participates in a certain syntactic relation (such as being a subject or object of a verb), then the substitution of the anaphor with the antecedent should also be possible since the antecedent will satisfy the selectional restrictions stipulated by the verb. Dagan and Itai's model substitutes the anaphor with each of the candidates, and the candidate that produces the most frequent co-occurrence patterns is preferred.

The authors illustrate their approach on a sentence taken from the *Hansard* corpus of proceedings of the Canadian Parliament:

(5.1) They knew full well that the companies held tax money aside for collection later on the basis that the government said it was going to collect it.

There are two occurrences of *it* in the above sentence. The first is the subject of *collect* and the second is its object. Statistics are gathered for the three candidates for antecedents in this sentence: *money*, *collection* and *government*. Table 5.1 lists the patterns produced by substituting each candidate with the anaphor, and the number of times each of these patterns occurred in the corpus. According to these statistics, *government* is preferred as the antecedent of the first *it* (which is in subject position), and *money* of the second (which is in object position).

Example (5.1) shows how 'selectional restrictions' based on collocation patterns practically eliminate all but the correct alternatives (*money* is not a real candidate for the first *it*). Other examples demonstrate that when there is more than one alternative satisfying the collocation patterns,[2] one solution[3] would be to pick the more frequently[4] occurring one as the antecedent:

(5.2) When the hog producers were in trouble a year ago and asked for some help, they got it immediately.

In this case both *help* and *trouble* may serve as the object of *get*. According to the statistics gathered from the corpus, the pattern 'get help' (verb–object) was counted 94 times, whereas the pattern 'get trouble' occurred 42 times.

Dagan and Itai's model consists of two separate phases. The first phase is the so-called 'acquisition' phase in which the corpus is processed and the statistical

database is built. The second is the 'disambiguation' phase, in which the statistical database is used to resolve (disambiguate) third person anaphors. The statistical database contains co-occurrence patterns for the following pairs of syntactic relations: 'subject–verb', 'verb–object' and 'adjective–noun'. To identify these relations, each sentence is parsed by the PEG parser (Jensen 1986).[5]

An experiment was performed to resolve anaphoric *it* in the *Hansard* corpus. The test data was manually selected from the corpus in the following way. Firstly, sentences containing *it* were extracted randomly from the corpus. Only candidates within the same sentence as the anaphor were considered (the *Hansard* corpus used for the experiment did not contain consecutive sentences).[6] Then, instances of non-anaphoric (pleonastic) occurrences of *it*, instances of anaphoric *it* whose antecedent was not an NP and instances where the anaphor was not involved in one of the three syntactic relations above were manually filtered. In addition, all trivial cases in which the anaphor had only one possible antecedent were also removed. As a result, about two-thirds of the original sentences were removed and the experiment was conducted on 59 examples.

The statistics were collected from a part of the corpus consisting of 28 million words. In 21 out of the 59 examples the algorithm could not approve any of the candidates because the threshold of 5 occurrences per alternative could not be reached. In the remaining 38 examples, Dagan and Itai's method proposed the correct antecedent 33 times (87% of the cases).

Dagan and Itai also explored the usefulness of their statistics by combining their algorithm with other methods that did not exploit co-occurrence patterns. First they examined the possibility of improving Hobbs's algorithm which, in its original version, proposed only one candidate. Hobbs's algorithm was modified to continue the search after proposing the antecedent and to produce additional candidates in the order encountered during the search (Dagan and Itai 1991).

The two methods were combined in the following way. The co-occurrence statistics overrode Hobbs's first preference whenever the patterns in which one of Hobbs's next candidates[7] occurred were observed in the corpora much more frequently than the patterns involving the first candidate.[8]

Statistics for the co-occurrence patterns were collected from the following three corpora: *The Washington Post* articles (about 40 million words), the Associated Press news wire (24 million words) and the *Hansard* corpus (85 million words). Sentences of no more than 25 words containing the pronoun *it* were extracted. For each sentence the previous sentence was also extracted.[9] These sentences were parsed by the ESG parser and Hobbs's algorithm was run on the resulting tree to produce the list of candidates. In addition, the syntactic relations involving the pronoun *it* were obtained. Each candidate was substituted in these relations to generate alternative patterns which were matched against the statistical database.

As in the first experiment, 'examples that were not appropriate for the use of selectional constraints' (Dagan and Itai 1991: 131) were removed. In addition to the cases described above for the first experiment, the authors removed sentences for which the parser failed to produce a reasonable tree. Also, they did not

consider the cases 'where the pronoun was not involved in any semantically meaningful relation (such as being the subject of the verb *to be*' (ibid.). Instances where the antecedent was a proper noun were discounted as well.[10] Finally, cases when one of the antecedents was an anaphor were also ignored, since the lexical NP antecedent may have been in a preceding sentence, not available for the test.

The filtering process yielded 74 cases of 'ambiguous' third person pronominal anaphors. Out of these examples, 38 did not qualify because the patterns observed did not exceed the threshold. On the remaining only 36 examples, Hobbs's algorithm alone scored a success rate of 64% which was boosted to 86% when combined with the statistical filter. When all examples were taken into account, including those that were not amenable to the statistical filter, the overall success rate was 74%, still marking a 10% increase owing to the application of the co-occurrence statistics.

Dagan and Itai's co-occurrence-based method was also tested in enhancing Lappin and Leass's syntax-based RAP algorithm (Dagan et al. 1995) within the genre of technical manuals. The results show that while increasing the success rate by 3%, within this genre, lexical preference patterns alone are not as efficient in pronoun resolution as an algorithm based on syntactic and attentional measures of salience (for more details see section 5.3.4).

Dagan and Itai note that their model deals with patterns involving specific words (e.g. *government*) and not semantic classes (e.g. *institution*) as in some semantic models. They argue that the use of word level patterns directly collected from the corpus has the advantage of getting more accurate 'constraints'. At the same time, they agree that the use of semantic classes has the advantage of generality: when there is insufficient data about a specific pattern, data about patterns containing words of the same semantic classes may be helpful.[11]

Finally it should be pointed out that domain of the corpus influences the frequency of patterns: corpora pertaining to different domains may feature different collocation patterns.

Although tested on a very small set of data, Dagan and Itai's model appears to be a very useful technique for resolving anaphors, especially when very large corpora are available to collect collocation statistics. One problem is that due to the possible sparsity of data, the approach may not be applicable in all cases. An alternative proposed by Mitkov (1996, 1998b) is to use collocations as complementary but not as the sole preference.[12]

5.3 Lappin and Leass's algorithm

5.3.1 Overview

Shalom Lappin and Herbert Leass (1994) describe an algorithm for resolving third person pronouns (including reflexives and reciprocals) whose antecedents are NPs. The algorithm, termed Resolution of Anaphora Procedure (henceforth RAP) operates on syntactic representations generated by McCord's Slot

Grammar parser (McCord 1990, 1993). It relies on salience measures derived from the syntactic structure as well as on a simple dynamic model of attentional state to select the antecedent of a pronoun from a list of NP candidates. It does not employ semantic information or real-world knowledge in choosing from the candidates. RAP contains the following main components:

- An intrasentential syntactic filter for ruling out coreference between a pronoun and an NP on syntactic grounds.
- A morphological filter for ruling out coreference between a pronoun and an NP due to non-agreement of person, number, or gender features.
- A procedure for identifying pleonastic pronouns.
- An anaphor binding algorithm for identifying the possible antecedent of a reflexive or reciprocal pronoun within the same sentence.
- A procedure for assigning values to several salience parameters for an NP, including syntactic role, parallelism of syntactic roles, frequency of mention, proximity, and sentence recency. Higher salience weights are assigned to (i) subject over non-subject NPs, (ii) direct objects over other complements, (iii) arguments of a verb over adjuncts and objects of prepositional phrase adjuncts of the verb, and (iv) head nouns over complements of head nouns.
- A procedure for identifying anaphorically linked NPs as an equivalence class for which a global salience value is computed as the sum of the salience values of its elements.
- A decision procedure for selecting the preferred element from a list of antecedent candidates for a pronoun.

The *syntactic filter on pronoun–NP coreference* consists of six conditions for NP–pronoun non-coreference within a sentence. These conditions are presented below and are illustrated by examples (NPs and pronouns carrying different indexes cannot be coreferential). In particular, a pronoun P is non-coreferential with a (non-reflexive or non-reciprocal) noun phrase NP if any of the following conditions hold:

1. P and NP have incompatible agreement features.[13]
   The woman$_i$ said that he$_j$ is funny.
2. P is in the argument domain of NP.[14]
   She$_i$ likes her$_j$.
   John$_i$ seems to want to see him$_j$.
3. P is in the adjunct domain of NP.[15]
   She$_i$ sat near her$_j$.
4. P is an argument of a head H, NP is not a pronoun, and NP is contained in H.
   He$_i$ believes that the man$_j$ is amusing.
   This is the man$_i$, he$_j$ said, John$_k$ wrote about.
5. P is in the noun phrase domain of NP.[16]
   John$_i$'s portrait of him$_j$ is interesting.
6. P is a determiner of a noun Q, and NP is contained in Q.[17]
   His$_i$ portrait of John$_j$ is interesting.
   His$_i$ description of the portrait by John$_j$ is interesting.

The *procedure for identifying pleonastic pronouns* includes lexical and syntactic tests by looking up a list of modal adjectives (such as *necessary*, *possible*, *certain*, *likely*, *difficult*, *legal*, etc.) and cognitive verbs (such as *recommend*, *think*, *believe*, etc.) to identify the constructions specified in the examples below in which *it* is considered pleonastic. Syntactic variants of these constructions are recognised as well:

It is *ModalAdj* that *S*
It is *ModalAdj* (for *NP*) to *VP*
It is *CogV-ed* that *S*
It seems/appears/means/follows (that) *S*
NP makes/finds it *ModalAdj* (for *NP*) to *VP*
It is time to *VP*
It is thanks to *NP* that *S*

The *anaphor binding algorithm* uses the following hierarchy of 'argument slots': *subj* > *agent* > *obj* > *iobj* / *pobj*. Here *subj* is the surface subject slot as identified by the slot grammar parser, agent is the deep subject slot of a verb heading a passive VP, *obj* is the direct object slot, *iobj* is the indirect object slot, and *pobj* is the object of a PP complement of a verb, as in *put NP on NP*. A noun phrase NP is the antecedent[18] for a reflexive or reciprocal pronoun R iff R and NP do not have incompatible agreement features, and any of the following conditions hold:

1. R is in the argument domain of NP, and NP fills a higher argument slot than R.
   They$_i$ wanted to see themselves$_i$.
   Mary knows the people$_i$ who John introduced to each other$_i$.
2. R is in the adjunct domain of NP.
   He$_i$ worked by himself$_i$.
   Which friends$_i$ plan to travel with each other$_i$?
3. R is in the noun phrase domain of NP.
   John likes Bill$_i$'s portrait of himself$_i$.
4. NP is an argument of a verb V, there is a noun phrase Q in the argument domain or the adjunct domain of NP such that R has no noun determiner, and R is (i) an argument of Q, or (ii) an argument of a preposition PREP and PREP is an adjunct of Q.
   They$_i$ told stories about themselves$_i$.
5. R is a determiner of a noun Q, and (i) Q is in the argument domain of NP and NP fills a higher argument slot than Q, or (ii) Q is in the adjunct domain of NP.
   [John and Mary]$_i$ like each other$_i$'s portraits.

Salience weighting applies to discourse referents and is computed on the basis of *salience factors*. In addition to *sentence recency* (where recent sentences are given higher weight), the algorithm gives additional weight to subjects (*subject emphasis*), predicate nominals in existential constructions (*existential emphasis*), direct objects (*accusative emphasis*), noun phrases that are not contained in other noun phrases (*head noun emphasis*) and noun phrases that are not contained in

Table 5.2 Salience factor types with initial weights

Factor type	Initial weight
Sentence recency	100
Subject emphasis	80
Existential emphasis	70
Accusative emphasis	50
Indirect object and oblique complement emphasis	40
Head noun emphasis	80
Non-adverbial emphasis	50

adverbial prepositional phrases (*non-adverbial emphasis*). The salience factors and their weights are given in Table 5.2. The following three examples illustrate the factors existential emphasis (the italicised NP is a predicate nominal in an existential construction), head noun emphasis (the NP in italics does not receive head noun emphasis) and non-adverbial emphasis (the NP in italics does not receive non-adverbial emphasis) respectively.

1. There are only *a few restrictions* on LQL query construction for Wordsmith.
2. the assembly in *bay C*
3. In the *Panel definition panel*, select the 'Specify' option from the action bar.

5.3.2 The resolution algorithm

The RAP's procedure for identifying antecedents of pronouns works as follows[19]:

1. First a list of all NPs in the current sentence is created and the NPs are classified according to their type (definite NP, pleonastic pronoun, other pronoun, indefinite NP).
2. All NPs occurring in the current sentence are examined.
   (a) NPs that evoke new discourse referents are distinguished from NPs that are presumably coreferential with already listed discourse referents as well as from those used non-referentially (e.g. pleonastic pronouns).
   (b) Salience factors are applied to the discourse referents evoked in the previous steps as appropriate.
   (c) The syntactic filter and reflexive binding algorithm are applied.
      (i) If the current sentence contains any personal or possessive pronouns, a list of pronoun–NP pairs from the sentence is generated. The pairs for which coreference is ruled out on syntactic grounds are identified.
      (ii) If the current sentence contains any reciprocal or reflexive pronouns, a list of pronoun-NP pairs is generated so that each pronoun is paired with all its possible antecedent binders.
   (d) If any non-pleonastic pronouns are present in the current sentence, their resolution is attempted in the linear order of pronoun occurrence in the sentence.

In the case of reflexive or reciprocal pronouns, the possible antecedent binders are identified by the anaphor binding algorithm. If more than one candidate is found, the one with the highest salience weight is chosen.

In the case of *third person pronouns*, a list of possible antecedent candidates is created. It contains the most recent referent of each equivalence class. The salience weight of each candidate is calculated (as the sum of the values of all salience factors that apply to it), and included in the list. The salience weight of these candidates can be additionally modified. For example, cataphora is strongly penalised, whereas parallelism of grammatical roles is rewarded. Also, the salience weights of candidates from previous sentences are degraded by a factor of 2 when each new sentence is processed. Unlike the salience factors shown in Table 5.2, these modifications of the salience weights are local to the resolution of a particular pronoun. Next, a salience threshold is applied: only those candidates whose salience weight is above the threshold are considered further.

In the final step agreement of number and gender is checked. This procedure seems to be much simpler for English than for other languages, which may exhibit ambiguity of the pronominal forms as to gender and number.[20] First the morphological filter is applied, followed by the syntactic filter. If more than one candidate remains, the candidate with the highest salience weight is chosen. In the event of more than one candidate's remaining, the candidate closest to the anaphor is selected as the antecedent.

For more details of the stages of the algorithm, see Lappin and Leass (1994).

5.3.3 Evaluation

RAP was tuned on a corpus of five computer manuals containing a total of approximately 82 000 words. From this corpus 560 occurrences of third person pronouns (including reflexives and reciprocals) and their antecedents were extracted. In the training phase the authors experimented extensively with salience weighting in order to optimise RAP's success rate.[21] The parallelism reward was introduced at this stage, as it seemed to substantially improve the results. A salience factor that was originally present, viz. *matrix emphasis*, was revised, modified and termed *non-adverbial emphasis*. In its original form this factor contributed to the salience of NPs not contained in a subordinate clause or in an adverbial prepositional phrase demarcated by a separator. However, this factor was found to be too general because it did not take into account the positions of the pronouns and their candidates for antecedents. Lappin and Leass also experimented with the initial weights for the various factors, with the size of the parallelism award and cataphora penalty, attempting to optimise RAP's overall success rate.[22]

The blind test was performed on 360 pronoun occurrences, which were randomly selected from a corpus of computer manuals containing 1.25 million words. RAP performed successful resolution in 86% of the cases, with 72% success for the intersentential cases (altogether 70) and 89% for intrasentential cases (altogether 290).[23]

Lappin and Leass also investigated the relative contribution of each of the salience factors by switching some of them off and running a blind test. The following evaluation variants were tested:

I. 'standard' RAP (as used in the blind test);
II. parallelism reward de-activated;
III. non-adverbial and head emphasis de-activated;
IV. matrix emphasis used instead of non-adverbial emphasis;
V. cataphora penalty de-activated;
VI. subject, existential, accusative and indirect object/oblique complement emphasis (i.e. hierarchy of grammatical roles) de-activated;
VII. equivalence classes de-activated;
VIII. sentence recency and salience degradation de-activated;
IX. all 'structural' salience weighting de-activated (II + III + V + VI);
X. all salience weighting and degradation de-activated.

The results of these tests suggest that the recency factor has the highest relative impact on the overall score, bringing down the overall success rate by 22%.

5.3.4 RAP enhanced by lexical preference

Dagan et al. (1995) constructed a procedure (referred to as RAPSTAT) for using statistically measured lexical preference patterns to re-evaluate RAP's salience rankings of antecedent candidates, in an attempt to enhance RAP's performance. RAPSTAT assigns a statistical score to each element of a candidate list that RAP generates. This score is calculated on the basis of a corpus-based collocation preference, in a similar way to that described in Dagan and Itai (1990).[24] If the scores proposed by RAPSTAT significantly differ from the salience preferences prescribed by RAP and if the difference in the salience weightings is still under the admissible threshold, then RAP is overruled by RAPSTAT in deciding on the antecedent.

The following is an example of a case where RAPSTAT overrules RAP:

(5.3) The Send Message display is shown, allowing you to enter your message and specify where it will be sent.

RAP assigns salience values of 345 and 315 to the candidates *display* and *message* respectively (see Lappin and Leass 1994). In the corpus used for testing RAPSTAT, the verb–object pair *send-display* appeared only once, while *send-message* occurred 289 times. As a result, *message* received a considerably higher statistical score than *display* and was selected correctly by RAPSTAT as the antecedent for *it* (the difference in the salience weightings of the two candidates was under the difference threshold which was set to 100 for this experiment).

The blind test for RAPSTAT was carried out on the corpus used for evaluation of RAP. RAPSTAT scored a success rate of 89% which represented a 3% improvement on RAP's performance. RAPSTAT disagreed with RAP in 41 cases, 25 times (61%) correctly and 16 times (39%) incorrectly. The results show that within the restricted genre of technical manuals, incorporating statistical

information on lexical preference patterns into a salience-based anaphora resolution procedure provides a modest improvement in performance.

5.3.5 Comparison with other approaches to anaphora resolution

RAP was compared on the same data with Hobbs's (1976, 1978) algorithm (see section 4.5). The test excluded pleonastic, reflexive and reciprocal pronouns since Hobbs's algorithm did not deal with these. Moreover, the Slot Grammar implementation of the algorithm[25] gave it the full advantage of RAP's syntactic-morphological filter, which is more powerful than the configurational filter built into the original specification of the algorithm. Therefore, the test results provided a direct comparison of RAP's salience metric and Hobbs's search procedure (Lappin and Leass, 1994: 555).

The results of the blind test (360 pronoun occurrences of which 70 were intersentential anaphors and 290 intrasentential anaphors) showed that, overall, RAP performed better with an 86% success rate as opposed to 82% obtained by Hobbs's algorithm. RAP scored better on intrasentential anaphora (89% vs. 81%) which was much more frequent in the corpus (see above). However, Hobbs's algorithm was more successful than RAP in resolving intersentential anaphora (87% vs. 74%).

Lappin and Leass conclude that because of the high rate of agreement between RAP and Hobbs's algorithm, there is a significant degree of convergence between salience as measured by RAP and the configurational prominence defined by Hobbs's search procedure. This is to be expected in English, where grammatical roles are identified by means of phrase order. The authors also conjecture that in languages where grammatical roles are case marked and word order is relatively free, there will be greater divergence in the predictions of the two algorithms.

Lappin and Leass's work is one of the most influential contributions to anaphora resolution in the 1990s: it has served as a basis for the development of other approaches (see next section) and has been extensively cited in the literature.

5.4 Kennedy and Boguraev's parse-free approach

Kennedy and Boguraev (1996) report on a modified version of RAP which does not require in-depth, full syntactic parsing but works instead from the output of a part-of-speech tagger enriched with annotations of grammatical function. The system uses a phrasal grammar for identifying NP constituents and, similarly to Lappin and Leass (1994), employs salience preference to rank candidates for antecedents. It should be pointed out that Kennedy and Boguraev's approach is not a simple knowledge-poor adaptation of RAP: it is rather an extension, given that some of the factors used in the new system are unique (Table 5.3).

The main motivation for developing a parser-free version of RAP is the fact that while one of the strong points of Lappin and Leass's algorithm is that it operates primarily on syntactic information alone, this seems to be a limiting factor for its wider use: the state of the art of parsing technology still falls short of broad-coverage, robust and reliable output.[26] Additionally, the authors were

interested in developing a more general text-processing framework which, due to the lack of full syntactic parsing capability, would normally have been unable to use a high-precision anaphora resolution tool.

Kennedy and Boguraev use the ENGCG part-of-speech tagger (Voutilainen et al. 1992; Karlsson et al. 1995) which in addition to delivering high recall (99.77%) and precision (95.54%) over a variety of text genres, supplies the grammatical function such as subject, object, etc., for each input token.

For each lexical item in a sentence the tagger provides a set of values which indicate its morphological, lexical and grammatical features. In addition, the tagger output is enriched by a simple position-identification function which associates an integer with each token in a text sequentially (referred to as *offset*). As an example, consider (5.4).

(5.4) For 1995 the company set up its headquarters in Hall 11, the newest and most prestigious of CeBIT's 23 halls.

This text would be presented in the following way to the anaphora resolution algorithm (note the information on the grammatical function such as *@SUBJ* – subject, *@FMAINV* – main verb, etc.; *off* denotes offset):

'For/off139' 'for' PREP @ADVL
'1995/off140' '1995' NUM CARD @<P
'the/off141' 'the' DET CENTRAL ART SG/PL @DN>
'company/off142' 'company' N NOM SG/PL @SUBJ
'set/off143' 'set' V PAST VFIN @+FMAINV
'up/off144' 'up' ADV ADVL @ADVL
'its/off145' 'it' PRON GEN SG3 @GN>
. . .
'$./off160' '.' PUNCT

A simple NP grammar (reduced to modifier-head groups) identifies all noun phrases on the basis of the tagger's output. The NP boundaries are returned in offset values; the offset also provides important information about precedence relations. In addition, a set of patterns is used to detect nominal sequences in two subordinate syntactic environments (containment in an adverbial adjunct or containment in an NP). This is accomplished by running patterns that identify NPs that occur locally to adverbs and relative pronouns as well as to noun–preposition and noun–complementiser sequences. Pattern matching also identifies occurrences of pleonastic *it*.

Once the extraction procedure is completed, a set of discourse referents is generated on the basis of the detected NPs. A discourse referent has the form:

TEXT:	text form
TYPE:	referential type (e.g. REF, PRO, RFLX)
AGR:	person, number, gender
GFUN:	grammatical function
ADJUNCT:	T or NIL
EMBED:	T or NIL
POS:	text position

Each discourse referent contains information about itself and the context in which it appears, the only information about its relation to other discourse referents being in the form of precedence relations (as indicated by the text position). The absence of explicit information about configurational relations marks the crucial difference between Kennedy and Boguraev's algorithm and that of Lappin and Leass.[27]

Once the representation of the text has been recast as a set of discourse referents (ordered by offset value), it is sent to the anaphora resolution algorithm. The basic logic of the algorithm parallels that of Lappin and Leass. The text is examined sentence by sentence and the discourse referents are interpreted from left to right.[28] Coreference is determined by first eliminating from consideration those discourse referents to which an anaphor cannot possibly refer, and then selecting the antecedent from the candidates that remain by means of salience measure. The salience factors used by Kennedy and Boguraev are a superset of those used in Lappin and Leass (1994);[29] in addition, they introduce the salience factors *possessive*, which rewards discourse referents whose grammatical function is possessive, and *context*, which boosts the score of candidates that appear in the same discourse segment as the anaphor. The discourse segment is determined by a text-segmentation algorithm which follows Hearst (1994).

The salience factors employed and their values are presented in Table 5.3. As with RAP, the values of the two new factors have been determined experimentally on the basis of the relative importance of each factor as a function of the overall success rate of the algorithm.

Following Lappin and Leass (1994), Kennedy and Boguraev calculate the salience of each coreference class by adding up all the values of the salience factors which are satisfied by some member of the class. When a pronoun is resolved to a previously introduced discourse referent, the pronoun is added to the equivalence class associated with the discourse referent and the salience of the coreference class is re-calculated on the basis of its newly added member.

Table 5.3 Salience factor types with initial weights and their abbreviations as used by Kennedy and Boguraev (1996)

Factor type	Initial weight
Sentence recency (SENT-S; iff in the current sentence)	100
Context emphasis (CNTX-S; iff in the current context)	50
Subject emphasis (SUBJ-S; iff GFUN = subject)	80
Existential emphasis (EXST-S; iff in an existential construction)	70
Possessive emphasis (POSS-S; iff GFUN = possessive)	65
Accusative emphasis (ACC-S; iff GFUN = direct object)	50
Indirect object emphasis (DAT-S; iff GFUN = indirect object)	40
Oblique complement emphasis (OBLQ-S; iff the complement of a preposition)	30
Head noun emphasis (HEAD-S; iff EMBED = NIL)	80
Non-adverbial emphasis (ARG-S; iff ADJUNCT = NIL)	50

The salience decreases or increases according to the frequency of reference to a specific coreference class: for instance, it decreases gradually if no recent mentions to discourse referents from this class have been made.

The resolution strategy, by and large, follows that of Lappin and Leass. The first step in interpreting the discourse referents in a new sentence is to decrease by a factor of 2 the salience weights of the coreference classes that have been already established. Next, all non-anaphoric discourse referents in the current sentence are identified and a new coreference class for each one is generated,[30] calculating its salience weight based on how the discourse referent satisfies the set of salience factors.

The second step involves reflexives and reciprocals. A list of candidate antecedent–anaphor pairs is generated for each one of them, based on the hypothesis that a reflexive or reciprocal must refer to a co-argument. In the absence of syntactic configurational information, co-arguments are located with the help of grammatical function (as determined by ENGCG) and precedence relations. A reflexive can have three possible grammatical function values: direct object, indirect object and oblique. In the first case, the closest preceding discourse referent marked with the grammatical function subject is identified as antecedent. In the latter cases, both the preceding subject and the closest preceding direct object that is not separated from the anaphor by a subject are proposed as possible antecedents. If more than one candidate is returned as a possible antecedent, the one with the highest salience weight is declared as the actual antecedent. Once the antecedent has been identified, the anaphor is added to the coreference class associated with the antecedent and the salience weight of this class is re-calculated accordingly.

The third and final step addresses the interpretation of personal pronouns. The resolution strategy is as follows. First a set of possible candidate antecedents is generated. This is accomplished by running the morphological agreement and disjoint reference filters over candidates whose salience weights exceed a certain threshold. The morphological agreement filter tests for person, number and gender agreement between the pronoun and the candidate.

The determination of disjoint reference represents 'a significant point of divergence' between Kennedy and Boguraev's algorithm and that of Lappin and Leass. In the absence of full syntactic analysis which makes possible the incorporation of an intrasentential syntactic filter in RAP, Kennedy and Boguraev's parser-free approach relies on inferences from grammatical function and precedence to approximate the following three configurational constraints that play an important role in ruling out coreference.

- *Constraint 1:* A pronoun cannot refer with a co-argument.
- *Constraint 2:* A pronoun cannot co-refer with a non-pronominal constituent that it both commands and precedes.
- *Constraint 3:* A pronoun cannot co-refer with a constituent that contains it.

Constraint 1 is implemented by tracking down all discourse referents in direct object, indirect object or oblique positions which follow a pronoun identified

as subject or object, as long as no subject intervenes: it is hypothesised that a subject marks the beginning of the next clause. Discourse referents that satisfy these conditions are singled out as disjoint.

Constraint 2 is implemented for every non-adjunct and non-embedded pronoun by removing from further consideration all non-pronominal discourse referents following the pronoun in the same sentence. The command relation is indicated by the precedence relation and by the syntactic environment: an argument that is not contained in an adjunct or embedded in another noun phrase commands those expressions which it precedes.

Constraint 3 makes use of the observation that a discourse referent contains every object to its right with a non-nil EMBED value (see Table 5.3). This constraint identifies as disjoint a discourse referent and every pronoun that follows it and has a non-nil EMBED value, until a discourse referent with EMBED value of NIL is located (marking the end of the containment domain). Constraint 3 rules out coreference between a possessive pronoun and the NP that it modifies.

The candidates that pass the agreement and disjoint reference filters are evaluated further. Cataphoric pronouns are penalised, whereas intrasentential candidates that satisfy either the locality heuristics or the parallelism heuristics have their salience weight increased. The locality heuristic was proposed by Kennedy and Boguraev to negate the effect of subordination when both the candidate and the anaphor appear in the same subordinate context (determined as a function of precedence relations and EMBED and ADJUNCT values). The salience of the candidate in the same subordinate context as the pronoun is temporarily increased to the level it would have if this candidate were not in the subordinate context; the level is returned to normal after the anaphor is resolved.

The parallelism heuristic (different from the one used by Lappin and Leass) rewards candidates where the syntactic functions (GFUN values) of candidate and anaphor are the same as the syntactic functions of a previously identified anaphor–antecedent pair.

Finally, the candidates under consideration are ranked according to their salience weight and the one with the highest value is proposed as the antecedent. If two or more candidates have the same salience weight, the one immediately preceding the anaphor is chosen to be the antecedent. For various examples illustrating how the algorithm works see Kennedy and Boguraev (1996).

Kennedy and Boguraev's experiment shows that with little compromise of accuracy (as compared to RAP) their approach delivers wide coverage. The dataset used for evaluation featured 27 texts taken from a random selection of genres, including press releases, product announcements, news stories, magazine articles, and other documents from the World Wide Web. These texts contained 306 third person anaphoric pronouns of which 231 were correctly resolved.[31] This gives an accuracy of 75%, which is not much below Lappin and Leass's 86% accuracy obtained on the basis of data from one genre only (technical manuals).[32]

The authors conducted an error analysis which showed that 35% of the errors were due to gender mismatch problems and 14% of the errors came from quoted speech. The persistence of gender mismatches reflects the lack of a consistent gender slot in the ENGCG output. Kennedy and Boguraev believe that augmenting the algorithm with a lexical database that includes more detailed gender information would result in improved accuracy. They also conjecture that to ensure better results, quoted speech has to be handled separately from the rest of the surrounding text.

Interestingly, Kennedy and Boguraev find that only a small number of errors can be directly attributed to the absence of configurational information. Of the 75 misinterpreted pronouns, only 2 involved failure to establish configurationally determined disjoint reference (both of these involved Constraint 3). This finding is different from that outlined in Lappin and Leass (1994) and Dagan et al. (1995), which suggests that syntactic filters have a prominent role in anaphora resolution.

5.5 Baldwin's high-precision CogNIAC

The pronoun resolution program CogNIAC (Baldwin 1997) was used as the pronoun component of the University of Pennsylvania's coreference entry in the MUC-6 evaluation. The main theoretical assumption underlying CogNIAC's strategy is that there is a subclass of anaphora which does not require general-purpose reasoning and can be resolved with the help of limited knowledge and resources. What distinguishes CogNIAC from a number of other algorithms is that it does not resolve a pronoun in cases of ambiguity, i.e when it is not sufficiently confident about a proposed antecedent. This results in a system that produces very high precision, but unsatisfactory recall.[33]

CogNIAC makes use of limited knowledge and resources and its pre-processing includes sentence detection, part-of-speech tagging and recognition of basal noun phrases (i.e. consisting of head nouns and modifiers, but without any embedded constituents), as well as basic semantic category information such as gender and number (and in one configuration, partial parse trees).

CogNIAC employs six core rules and two additional rules, which are given below, together with their performance on training data consisting of 200 pronouns in a narrative text.

1. *Unique in discourse*: If there is a single possible antecedent *i* in the read-in portion of the entire discourse, then pick *i* as the antecedent (this rule worked 8 times correctly and 0 times incorrectly on the training data).
2. *Reflexive*: Pick the nearest possible antecedent in the read-in portion of current sentence if the anaphor is a reflexive pronoun (16 correct, 1 incorrect).
3. *Unique in current and prior*: If there is a single possible antecedent *i* in the prior sentence and the read-in portion of the current sentence, then pick *i* as the antecedent (114 correct, 2 incorrect).
4. *Possessive pronoun*: If the anaphor is a possessive pronoun and there is a single exact string match *i* of the possessive in the prior sentence, then pick *i* as the antecedent (4 correct, 1 incorrect).

5. *Unique current sentence*: If there is a single possible antecedent *i* in the read-in portion of the current sentence, then pick *i* as the antecedent (21 correct, 1 incorrect).
6. *Unique subject/subject pronoun*: If the subject of the prior sentence contains a single possible antecedent *i*, and the anaphor is the subject of the current sentence, then pick *i* as the antecedent (11 correct, 0 incorrect).

CogNIAC works as follows: pronouns are resolved from left to right in the text and, for each pronoun, the above rules are applied in the order presented above. If for a specific rule an antecedent is found, then no further rules are applied. If no rules resolve the pronoun, then it is left unresolved.

CogNIAC's evaluation was conducted in two separate experiments, one of which was a comparison with Hobbs's naïve algorithm and another which was carried out on MUC-6 data. In the first experiment third person pronouns only were considered. The pre-processing consisted of part-of-speech tagging, delimitation of base noun phrases and identification of finite clauses. The results of the pre-processing were subjected to hand correction in order to make comparison with Hobbs's algorithm fair.[34] Errors were not chained, i.e. while processing the text from left to right, earlier mistakes were corrected before proceeding to the next noun phrase.

Since Hobbs's algorithm resolves all pronouns (unlike CogNIAC, which does not propose an antecedent in circumstances of ambiguity), two lower-precision rules were added to Rules 1–6 so that both algorithms could operate in robust[35] mode.

7. *Cb-picking*: If there is a backward-looking center Cb in the current finite clause that is also a candidate antecedent, then pick *i* as the antecedent.
8. *Pick most recent*: Pick the most recent potential antecedent in the text.

Baldwin notes that even though these two rules are of lower precision than the first six, they perform well enough to be included in the 'resolve all pronouns' configuration. Rule 7 was correct 10 times out of 13 based on training data with 201 pronouns, whereas Rule 8 succeeded 44 times out of 63.

The results of the first experiment indicate that both Hobbs's algorithm and CogNIAC did almost equally well on the evaluation texts: the naïve algorithm was correct in 78.8% of the cases, whereas the robust version of CogNIAC was successful in 77.9% of the cases (based on 298 pronouns from a text about 'two same gender people'). On the other hand, the high-precision version of CogNIAC scored a precision of 92% (190/206) and a recall of 64% (190/298).

The second experiment was performed on data from the *Wall Street Journal*. For this experiment, a few changes were made to the original version of CogNIAC by incorporating the following rules/modules:

- Rule(s) for processing quoted speech in 'a limited fashion'.
- Rule that searched back for a unique antecedent through the text at first 3 sentences, 8 sentences back, 12 sentences back and so on.
- Partial parser (Collins 1996) to identify finite clauses.
- Pattern for selecting the subject of the immediately surrounding clause.
- Detector of pleonastic *it*.

Also, Rules 4, 7 and 8 were disabled because they did not appear to be appropriate for the particular genre.

The performance of CogNIAC was less successful on this data with 75% precision and 73% recall. 'Software problems' accounted for 20% of the incorrect cases and another 30% were due to misclassification of a noun phrase as person or company or incorrect identification of number. The remaining errors were due to incorrect noun phrase identification, inability to recognise pleonastic *it* or cases without antecedent.

5.6 Resolution of definite descriptions

Research on anaphora resolution has focused almost exclusively on the interpretation of pronouns with a few notable exceptions of earlier work covering definite descriptions (Alshawi 1992; Carter 1986, 1987a; Sidner 1979) and of more recent projects (Cardie and Wagstaff 1999; Kameyama 1997; Poesio et al. 1997; Vieira and Poesio 2000a, 2000b; Muñoz and Palomar 2000; Muñoz et al. 2000, Muñoz 2001)[36] including indirect anaphora (Gelbukh and Sidorov 1999; Murata and Nagao 2000). A significant recent work on interpretation of anaphoric definite descriptions[37] is that of Vieira and Poesio (2000b). Their work led to the development of a shallow processing system relying on structural information, on the information provided by existing lexical resources such as WordNet, on minimal amounts of general hand-coded information or on information that could be acquired automatically from a corpus. As a result of the relatively knowledge-poor approach adopted, the system is not really equipped to handle definite descriptions which require complex reasoning; nevertheless, a few heuristics have been developed for processing this class of anaphoric NPs. On the other hand, the system is domain independent and its development was based on an empirical study of definite description use involving a number of annotators.

Vieira and Poesio classify the types of definite descriptions in the following way: **direct anaphora** for subsequent-mention definite descriptions referring to an antecedent with the same head noun as the description,[38] **bridging descriptions** which have an antecedent denoting the same discourse entity but represented by a different head noun[39] and thus often requiring extralinguistic knowledge for their interpretation and **discourse new** for first-mention definite descriptions denoting objects not related by shared associative knowledge to entities already introduced in the discourse.[40] The paper does not discuss indirect anaphora.[41]

Vieira and Poesio's system does not only attempt to find the antecedent of definite description anaphors. It is also capable of recognising discourse-new descriptions which appear to represent a large portion of the corpus investigated. The system does not carry out any pre-processing of its own and benefits from an annotated subset of the Penn Treebank I corpus (Marcus et al. 1993) containing newspaper articles from the *Wall Street Journal*. The corpus was divided into two parts: one containing approximately 1000 definite descriptions

used for the development of the system and another part of approximately 400 definite descriptions kept aside for testing. The algorithm used a manually developed decision tree created on the basis of extensive evaluation; the authors also experimented with automatic decision-tree learning algorithms (Quinlan 1993).

The system achieved 62% recall and 83% precision for direct anaphora resolution, whereas the identification of discourse-new descriptions was performed with a recall of 69% and a precision of 72%. Overall, the version of the system that only attempts to recognise first-mention and subsequent-mention definite descriptions obtained a recall of 53% and a precision of 76%. The resolution of bridging descriptions was a much more difficult task because lexical or world knowledge was often necessary for their resolution. For instance, the success rate in the interpretation of semantic relations between bridging descriptions (e.g. synonymy, hyponymy, meronymy) using WordNet was reported to be in the region of 28%.

5.7 Machine learning approaches

Natural language understanding requires a huge amount of knowledge about morphology, syntax, semantics, discourse and pragmatics and general knowledge about the real world but the encoding of all this knowledge represents an insurmountable impediment for the development of robust NLP systems. As an alternative to knowledge-based systems, machine learning methods offer the promise of automating the acquisition of this knowledge from annotated or unannotated corpora by learning from a set of examples (patterns). The term machine learning is frequently used to refer specifically to methods that represent learned knowledge in a declarative, symbolic form as opposed to more numerically-oriented statistical or neural-network training methods. In particular, it concerns methods that represent learned knowledge in the form of interpretable decision trees, logical rules and stored instances (Mooney 2002). The following section will describe a few anaphora resolution systems based on decision trees. The decision trees are classification functions represented as trees in which the nodes are attribute tests, the branches are attribute values and the leaves are class labels. Among the most extensively used and cited decision-tree algorithms are ID3 (Quinlan 1986) and C4.5 (Quinlan 1993). For a brief introduction to machine learning see Mooney (2002).

5.7.1 Aone and Bennett's approach

Aone and Bennett (1995, 1996) describe an anaphora resolution system for Japanese which is trained on a corpus of newspaper articles tagged with discourse information.[42] The work is a continuation of their multilingual anaphora resolution project (Aone and McKee 1993) where they report a 'robust, extendible and manually trainable' system.

The machine learning resolver (MLR) employs the C4.5 decision-tree algorithm (Quinlan 1993). The decision tree is trained on the basis of feature vectors for an anaphor and its possible antecedents. The training features can be unary and related either to the anaphor or to the candidate for antecedent (e.g. number or gender), or they can be binary and represent relations between the anaphor and the antecedent (e.g. distance). Altogether 66 features are used including lexical (e.g. lexical class), syntactic (e.g. grammatical function), semantic (e.g. semantic class) and positional (e.g. distance between the anaphor and the antecedent).

The training method operates on three parameters: anaphoric chains, anaphoric type identification and confidence factors. The *anaphoric chains* parameter is used for selecting both a set of positive training examples and a set of negative training examples. When this parameter is *on*, the positive training examples for each anaphor are all pairs consisting of the anaphor and any of the preceding NPs in the same coreferential chain as the anaphor. The negative training examples are the pairs including the anaphor and an NP which is not in the same coreferential chain. When the anaphoric chain parameter is *off*, only the pairs consisting of anaphors and their antecedents[43] are considered as positive examples.[44]

The *anaphoric type identification* parameter is used for training decision trees. When this parameter is *on*, a decision tree is trained to return 'no' when an anaphor and a candidate are not coreferential, or return the anaphoric type when they are coreferential. When the parameter is *off*, a binary decision tree is trained to answer just 'yes' or 'no' and does not have to return the type of the anaphor.

The *confidence factor* parameter (0–100) is used to prune decision trees. A higher confidence factor does less pruning of the tree and tends to overfit the training examples. With a lower confidence factor, more pruning is performed and this results in a smaller, more generalised tree. The confidence factors used are 25, 50, 75 and 100%.

The training corpus used to train decision trees contained 1971 anaphors which were spread over 259 different texts: 929 of the anaphors were proper names (of organisations), 546 were 'quasi-zero pronouns',[45] 282 were zero pronouns and 82 were definite descriptions. All the antecedents of these anaphors were organisations. The evaluation corpus featured 1359 anaphors of which 1271 were of the four anaphoric types mentioned above. Both the training and the evaluation texts were joint ventures and each article mentioned one or more organisations.

The evaluation was carried out on six different modes of the system; each mode was defined on the basis of the different values of the anaphoric chain, anaphoric type identification and confidence factors. With a view to moving the attention away from the inaccuracy in pre-processing and focusing instead on the resolution, the evaluation was done on the basis of only those anaphors which were identified by the program and not on the basis of all the anaphors in the text.[46] The measures used in the evaluation were *recall* and *precision* which were defined by Aone and Bennett as shown in Table 5.4:[47]

Table 5.4 Recall and precision as defined by Aone and Bennett (1995)

Recall = N_c/N_a, Precision = N_c/N_t	
N_a	Number of anaphors identified by the program
N_c	Number of correct resolutions
N_t	Number of resolutions attempted

as well as the combined *F-measure* expressed as

$$F = \frac{(\beta^2 + 1.0) \times P \times R}{\beta^2 \times P \times R}$$

where P is precision, R is recall and β is the relative importance given to recall over precision (in this case β = 1).

Using the F-measure as an indicative metric for overall performance, the modes with chain parameters turned on and type identification turned off[48] performed best with recall ranging from 67.53% to 70.20%, precision from 83.49% to 88.55% and F-measure from 76.27% to 77.27%. For more on the performance of each mode see Aone and Bennett (1995, 1996).

5.7.2 *McCarthy and Lehnert's approach*

McCarthy and Lehnert's RESOLVE system (1995) uses the C4.5 decision-tree system to learn how to classify coreferent noun phrases in the domain of business joint ventures. The feature vectors used by RESOLVE were created on the basis of all pairings of references and coreference links among them from a text manually annotated for coreferential noun phrases. The pairings that contained coreferent phrases formed positive instances, whereas those that contained non-coreferent phrases formed negative instances. From the 1230 feature vectors (or instances) that were created from the entity references marked in 50 texts, 322 (26%) were positive and 908 (74%) were negative. The following features and values were used (the first two features were applied to each NP individually; the other four features were applied to each pair of NPs):

- *Name*: Does a reference contain a name? Possible values {yes, no}.
- *Joint venture child*: Does a reference refer to a joint-venture child, e.g. a company formed as a result of a tie-up among two or more entities? Possible values {yes, no, unknown}.
- *Alias*: Does one reference contain an alias of the other, i.e. does each of the two references contain a name and is one of the names a substring of the other name? Possible values {yes, no}.
- *Both joint venture child*: Do both references refer to a joint-venture child? Possible values {yes, no, unknown}.
- *Common NP*: Do both references share a common NP? Possible values {yes, no}.
- *Same sentence*: Do the references come from the same sentence? Possible values {yes, no}.

The evaluation of RESOLVE, which was carried out on the MUC-5 English Joint Venture corpus[49] and reported in McCarthy and Lehnert (1995), focused on the coreference resolution algorithm since all pre-processing errors were manually post-edited. In this restricted genre the unpruned version of the algorithm scored 85.4% recall, 87.6% precision and 86.5% F-measure, whereas the pruned version obtained 80.1% recall, 92.4% precision and 85.8% F-measure.

5.7.3 Soon, Ng and Lim's approach

Soon, Ng and Lim (1999) describe a C4.5-based learning approach to coreference resolution of noun phrases in unrestricted text. The coreference resolution module is part of a larger coreference resolution system also featuring sentence segmentation, tokenisation, morphological analysis, part-of-speech tagging, noun phrase identification, named entity recognition and semantic class determination (via WordNet). The feature vectors used for training and evaluation consist of ten features. The following features apply to pairs of noun phrases.

- *Distance*: Possible values are {0, 1, 2, . . . }. If two noun phrases are in the same sentence, the distance feature is assigned a value of 0, if they are located in two consecutive sentences 1, and so on.
- *String match*: Possible values {yes, no}. If one string matches another, the value is yes, otherwise no.
- *Number agreement*: Possible values {yes, no}. If two NPs agree in number, the value of this feature is positive, otherwise negative.
- *Semantic class agreement*: Possible values {yes, no, unknown}. Two NPs are in agreement with regard to their semantic class either if they are of the same semantic class (e.g. *he* and *Mr. Dow* are both from the semantic class *male*) or if one is a parent of the other (e.g. as in the case of the semantic classes *male* and *person*). Two NPs are in disagreement with regard to their semantic class if their semantic classes are not the same and none of them is parent of the other (e.g. as in the case of the semantic classes *male* and *organisation*).
- *Gender agreement*: Possible values {yes, no, unknown}. The gender is marked as unknown for NPs such as *the president*, *chief executive officer* etc.
- *Proper name*: Possible values {yes, no}. The value of this feature is positive if both NPs are proper names.
- *Alias*: Possible values {yes, no}. The value of this feature is positive if both NPs are proper names that refer to the same entity.

The following features apply to individual NPs.

- *Pronoun*: Possible values {yes, no}. If an NP is a pronoun, then the value of this feature is yes, otherwise no.
- *Definite noun phrase*: Possible values {yes, no}.
- *Demonstrative noun phrase*: Possible values {yes, no}.

The size of the training data amounted to about 13 000 words, whereas the evaluation documents consisted of about 14 000 words. The coreference resolution system achieved a recall of 52%, precision 68%, yielding an F-measure of 58.9%.

It should be noted, however, that these results cannot be directly compared with those obtained by Aone and Bennett (1995) and by McCarthy and Lehnert (1995) since these researchers evaluated their systems on noun phrases that have been correctly identified. In contrast, Soon, Ng and Lim's approach was evaluated in a fully automatic mode against the background of pre-processing errors. Also, whereas the evaluation of McCarthy and Lehnert's system was carried out on specific types of NPs (organisations and business entities) and Aone and Bennett covered Japanese texts only, Soon et al.'s method processed all types of English NPs.

An updated version of Soon et al.'s system is reported in Soon et al. (2001). The new system deploys 12 features (as opposed to 10 in the original experiment), uses the C5 decision-tree algorithm and is trained and tested against both MUC-6 and MUC-7 data. The authors also use cross-validation to obtain the learning parameters.

Another approach that uses machine learning techniques is that of **Connoly et al. (1994).**

5.8 Probabilistic approach

Ge, Hale and Charniak (1998) propose a statistical framework for resolution of third person anaphoric pronouns. They combine various anaphora resolution factors into a single probability which is used to track down the antecedent. The program does not rely on hand-crafted rules but instead uses the Penn *Wall Street Journal* Treebank to train the probabilistic model.

The first factor the authors make use of is the *distance* between the pronoun and the candidate for an antecedent. The greater this distance, the lower the probability for a candidate NP to be the antecedent. The so-called 'Hobbs's distance' measure is used in the following way. Hobbs's algorithm is run for each pronoun until it has proposed N (in this case $N = 15$) candidates. The Kth candidate is regarded as occurring at 'Hobbs's distance' = K. Ge and co-workers rely on features such as *gender*, *number* and *animacy* of the proposed antecedent. Given the words contained in an NP, they compute the probability that this NP is the antecedent of the pronoun under consideration based on probabilities computed over the training data, which are marked with coreferential links. The authors also make use of *co-occurrence patterns*[50] by computing the probability that a specific candidate occurs in the same syntactic function (e.g. object) as the anaphor. The last factor employed is the *mention count* of the candidate. Noun phrases that are mentioned more frequently have a higher probability of being the antecedent; the training corpus is marked with the number of times an NP is mentioned up to each specific point.

The four probabilities discussed above are multiplied together for each candidate NP. The procedure is repeated for each NP and the one with the highest probability is selected as the antecedent. For more on the probabilistic model and the formulae used, see Ge et al. (1998).

The authors investigated the relative importance of each of the above four probabilities (factors employed) in pronoun resolution. To this end, they ran the

program 'incrementally', each time incorporating one more probability. Using only Hobbs's distance yielded an accuracy of 65.3%, whereas the lexical information about gender and animacy brought the accuracy up to 75.7%, highlighting the latter factor as quite significant. The reason the accuracy using Hobbs's algorithm was lower than expected was that the Penn Treebank did not feature perfect representations of Hobbs's trees.[51] Contrary to initial expectations, knowledge about the governing constituent (co-occurrence patterns) did not make a significant contribution, only raising the accuracy to 77.9%. One possible explanation could be that selectional restrictions are not clear-cut in many cases; in addition, some of the verbs in the corpus such as *is* and *has* were not 'selective' enough. Finally, counting each candidate proved to be very helpful, increasing the accuracy to 82.9%.

The annotated corpus consisted of 93 931 words and contained 2477 pronouns, 1371 of which were singular *he*, *she* and *it*. The corpus was manually tagged with reference indices and repetitions of each NP. In addition, cases of pleonastic *it* were excluded when computing the accuracy of the algorithm.[52] Ten per cent of the corpus was reserved for testing, whereas 90% was used for training.

In their paper Ge, Hale and Charniak also propose a method for unsupervised learning of gender information which they incorporate in the pronoun resolution system. The evaluation of the enhanced approach on 21 million words of *Wall Street Journal* text indicates improved performance, bringing the accuracy up to 84.2%.

5.9 Coreference resolution as a clustering task

Cardie and Wagstaff (1999) describe an unsupervised algorithm which views NP coreference resolution as a clustering task. Each noun phrase is represented as a vector of 11 features[53] and their computed values; the clustering algorithm coordinates these to partition the set of noun phrases into equivalence classes of coreferential chains.[54]

First, all noun phrases are located using the Empire NP finder (Cardie and Pierce 1998). Empire identifies only base noun phrases, i.e. simple noun phrases which contain no other smaller noun phrases. For example, *Chief Financial Officer of Prime Corp.* is too complex to be a base noun phrase. It contains two base noun phrases *Chief Financial Officer* and *Prime Corp*. (Cardie and Wagstaff 1999). Next, each NP in the input text is represented as a set of the features shown in Figure 5.1. Their values are automatically determined (and therefore not always accurate).

The degree of 'coreference closeness' between each two noun phrases in Figure 5.1 is computed on the basis of the 'distance metric'. The closer the distance, the higher the probability that two noun phrases are coreferential. Consequently, two noun phrases are considered as coreferential if the distance between them is smaller than a specific threshold (what they term the *clustering radius threshold*).

Individual words. The words contained in each NP are stored as a feature.

Head noun. The last word in the NP is considered the head noun.

Position. NPs are numbered sequentially, starting at the beginning of the document.

Pronoun type. Pronouns are marked as NOMinative (e.g. *he*, *she*), ACCusative (e.g. *him*, *her*), POSSessive (e.g. *his*, *her*) or AMBiguous (e.g. *you* and *it*).

Article. Each NP is marked INDEFinite if it contains *a* or *an* or DEFinite, if it contains *the*, or NONE.[55]

Appositive. A simple (and, admittedly, restrictive) heuristic is used to determine whether or not a noun phrase is in an appositive construction: if the noun phrase is surrounded by commas, contains an article, and is immediately preceded by another noun phrase, then it is marked as an appositive.

Number. If the head noun ends in an 's' the noun phrase is considered PLURAL, otherwise it is taken to be SINGULAR.

Proper name. A simple heuristic used is to look at two adjacent capitalised words, optionally containing a middle initial.

Semantic class. WordNet is made use of to obtain coarse semantic information about the head noun. The head noun is classified as one of TIME, CITY, ANIMAL, HUMAN or OBJECT. If none of these classes pertains to the head noun, its immediate parent in the class is returned as the semantic class, e.g. PAYMENT for the head noun pay. A separate algorithm identifies NUMBERs, MONEY, and COMPANYs.

Gender. Gender (MASCuline, FEMinine, EITHER or NEUTER) is determined via WordNet. A list of common first names is used to recover the gender of proper names.

Animacy. Noun phrases returned as HUMAN or ANIMAL are marked as ANIM; all others are considered to be INANIMate.

Figure 5.1 Features used in Cardie and Wagstaff's unsupervised algorithm.

The distance metric is defined as

$$\text{dist}(\text{NP}_i, \text{NP}_j) = \Sigma_{f \in A} w_a * \text{incompatibility}_a(\text{NP}_i, \text{NP}_j)$$

Here A corresponds to the NP feature set described above, while $incompatibility_a$ is a function which returns a value between 0 and 1 inclusive, and indicates the degree of incompatibility of the feature a for NP_i and NP_j. Finally, w_a denotes the relative importance of compatibility with regard to the feature a. The incompatibility functions and the corresponding weights are listed in Table 5.5.

The weights are chosen to correspond to the degree of restriction or preference imposed by each feature. Constraints with a weight ∞ represent filters that rule out coreference: two noun phrases can never corefer if they have incompatible values with regard to a certain feature. In the implemented version of the system, *number*, *proper name*, *semantic class*, *gender* and *animacy* operate as coreference filters. On the other hand, features with weight $-\infty$ force coreference

Table 5.5 Incompatibility functions and weights for each term in the distance metric

Feature *a*	Weight	Incompatibility function
Words	10.0	(number of mismatching words) / (number of words in the longer NP)
Head	1.0	1 if the head nouns differ; else 0
Position	5.0	(difference in position) / (maximum difference in document)
Pronoun	*r*	1 if NP_i is a pronoun and NP_j is not; else 0
Article	*r*	1 if NP_j is indefinite and not appositive; else 0
Word-substring	$-\infty$	1 if NP_i subsumes (entirely includes as a substring) NP_j
Appositive	$-\infty$	1 if NP_j is appositive and NP_i is its immediate predecessor; else 0
Number	∞	1 if they do not match in number; else 0
Semantic class	∞	1 if they do not match the class; else 0
Gender	∞	1 if they do not match in gender (allows EITHER to match MASC or FEM); else 0
Animacy	∞	1 if they do not match in animacy; else 0

between two noun phrases with compatible values for this feature. The *appositive* and *word-substring* features operate in such a capacity.

When computing a sum that involves both ∞ and $-\infty$, the approach chooses to be on the safe side, ∞ is given priority and the two noun phrases are not considered coreferential. Cardie and Wagstaff illustrate this by the following example:

(5.5) [$_{NP_1}$Reardon Steel Co.] manufactures several thousand tons of [$_{NP_2}$steel] each week.

In this example NP_1 subsumes NP_2 which results in a distance $-\infty$ for the *word-substring* term of the distance metric. On the other hand, NP_1's semantic class is COMPANY, whereas NP_2's class is OBJECT, thus generating a distance of ∞ for the *semantic class* feature. Therefore, dist (NP_1, NP_2) = ∞ and the two noun phrases are not considered coreferential.

The clustering algorithm starts at the end of the document and works backwards, comparing each noun phrase to all preceding noun phrases. If the distance between two noun phrases is less than the clustering radius *r*, then their classes are considered for possible merging (initially, each NP represents a coreference class on its own). Two coreference equivalence classes can be merged unless there exist any incompatible NPs in the classes to be merged.

The clustering approach was evaluated using the 'dry run' and 'formal evaluation' modes (MUC-6). For the dry run data set, the clustering algorithm obtained 48.8% recall and 57.4% precision, which came to an F-measure of 52.8%. The formal evaluation scores were 52.7% recall and 54.6% precision, coming to an F-measure of 53.6%.[56] Both runs used $r = 4$ which was obtained by testing different values on the dry run corpus. Different values of *r* ranging from 1.0 to 10.0 were tested and, as expected, the increase of *r* raised recall, but lowered precision.

The clustering approach was also compared with three baseline algorithms. The first baseline marked each pair of noun phrases as coreferential (i.e. all NPs in a document form one class), scoring 44.8% F-measure for the dry run data test and 41.5% for the formal run dataset. The second baseline considered each two NPs that had a word in common as coreferential; it produced scores of 44.1% and 41.3% respectively. Finally, the third baseline marked as coreferential only NPs whose heads matched; this baseline obtained *F*-measures of 46.5% and 45.7% respectively.

Cardie and Wagstaff's approach is knowledge-poor since it does not require full syntactic parsing, domain knowledge, etc. The approach is unsupervised[57] in that it requires neither annotation of training data nor a large corpus for computing statistical occurrences. In addition, this approach not only handles pronoun resolution, but tackles NP coreference as well. Its limitations lie in the 'greedy nature' of the clustering algorithm (an NP_j is linked to *every* preceding NP_i) and in the low accuracy of pre-processing (NPs are identified at base level only; most of the heuristics for computing the 11 features are very crude). Also, the clustering algorithm does not handle pleonastic *it* and reflexive pronouns.

5.10 Other recent work

Kameyama's algorithm (1997) for resolution of nominal anaphora[58] uses syntactically incomplete inputs (sets of finite-state approximations of sentence parts) which are even more impoverished than the inputs to Kennedy and Boguraev's system. The three main factors in Kameyama's algorithm are (i) accessible text regions, (ii) semantic consistency and (iii) dynamic preference. The accessible text region for proper names is the entire preceding text, for definite noun phrases it is set to 10 sentences, and for pronouns 3 sentences (ignoring paragraph boundaries). The semantic consistency filters are number consistency, type consistency[59] (anaphors must be either of the same type as their antecedents or subsume their type; e.g. *company* subsumes *automaker* and *the company* can take a *Chicago-based automaker* as an antecedent) and modifier consistency (e.g. *French* and *British* are inconsistent but *French* and *multinational* are consistent). The basic underlying hypothesis of the dynamic preference is that intrasentential candidates are more salient than intersentential ones and that syntax-based salience fades with time. Since information on the grammatical functions is unavailable, the syntactic prominence of grammatical functions such as subjects is approximated by left–right linear ordering. The algorithm was first implemented for the MUC-6 FASTUS information extraction system (Kameyama 1997) and produced one of the top scores (recall 59%, precision 72%).

Tetreault (1999) proposes a centering-based pronoun resolution algorithm called the *Left–Right Centering Algorithm* (LRC).[60] The LRC is an alternative of the original BFP algorithm (Brennan et al. 1987; see also section 46) in that it processes the utterances incrementally. It works by first searching for an antecedent in the current sentence and, if not successful, continues the search on the Cf-list of the

previous and the other preceding utterances[61] in a left-to-right fashion. The LRC was compared with Hobbs's naïve algorithm, BFP and Strube's S-list approach[62] on an annotated subset of the Penn Treebank containing 1696 pronouns.[63] Quoted text was removed from the corpus, being outside the 'remit' of the BFP and the S-list. The evaluation compared algorithms searching on all previous Cf-lists (Hobbs's algorithm, LRC-N, Strube-N) and those considering Cf(U_{N-1}) only (LRC-1, Strube-1, BFP).[64] Among the algorithms that searched all sentences, Hobbs's algorithm scored best (72.8%), followed closely by LRC-N (72.4%) and Strube-N (68.8%). The algorithms that searched the previous sentence only performed more modestly: LRC-1 (71.2%), Strube-1 (66.0%), BFP (56.7%).[65] Tetreault's evaluation, similar to that of Ge et al. (1998), was concerned with the evaluation of the pronoun resolution only, since the availability of an annotated corpus did not require any pre-processing. For discussion on the distinction between evaluating algorithms and systems, see Chapter 8, section 8.1.

Ferrández, Palomar and Moreno's algorithm[66] (1997, 1998, 1999) employs a Slot Unification Parser and works in two modes, the first benefiting from ontology and dictionary, and the second working from the output of a part-of-speech tagger in a knowledge-poorer environment. Various extensions and improvements of this algorithm, incorporated later in the PHORA system, have been described in Palomar et al. (2001a). The evaluation reports a success rate of 76.8% in resolving anaphors in Spanish. In a recent publication Palomar et al. (2001b) present the latest version of the algorithm which handles third person personal, demonstrative, reflexive and zero pronouns. This version features, among other improvements, syntactic conditions on Spanish NP–pronoun non-coreference and an enhanced set of resolution preferences. The authors also implement several known methods and compare their performance with that of their own algorithm. An indirect conclusion from this work is that an algorithm needs semantic knowledge in order to hope for a success rate higher than 75%.

The developments in anaphora resolution take place in the wider context of NLP, where the search for multilingual applications is a live issue. Against the background of growing interest in *multilingual work*, it is natural that anaphora resolution projects have started looking at the multilingual aspects of the approaches and in particular at how a specific approach can be used or adapted for other languages. An increasing number of projects have focused on languages other than English, which means that the initial monolingual (English) orientation of the field is no longer dominant. Recent works such as Mitkov and Stys (1997), Mitkov et al. (1998), Azzam et al. (1998a), Harabagiu and Maiorano (2000) and Mitkov and Barbu (2000) have established a new trend towards multilinguality in the field.

As an illustration, Harabagiu and Maiorano (2000) use an annotated bilingual English and Romanian corpus to improve the coreference resolution in both languages. The knowledge-poor system COCKTAIL (Harabagiu and Maiorano 1999) and its Romanian version are trained on the bilingual corpus and the results obtained outperform the coreference resolution in each of the individual languages. Mitkov and Barbu (2000) propose a 'mutual enhancement' approach

which benefits from a bilingual English–French corpus in order to improve the performance in both languages.[67]

The methodology of *evaluation in anaphora resolution* has been the focus of several recent papers (Bagga 1998; Byron 2001; Mitkov 1998a, 2000, 2001b). It is proposed in Mitkov (2001b) that evaluation should be carried out separately for anaphora resolution algorithms and for anaphora resolution systems. This paper argues that it would not be fair to compare the performance of an algorithm operating on 100% correct, manually checked input with an algorithm which is part of a larger system and works from the prone-to-error output of the pre-processing modules of the system. Even though extensive evaluation has become a must in anaphora resolution, one of the problems has been that most of the evaluations do not say much as to where a specific approach stands with respect to others, since there has been no common ground for comparison. A possible way forward is the evaluation workbench which has recently come into existence (Mitkov 2000, 2001b; Barbu and Mitkov 2000, 2001; see also section 8.6) and which offers comparison of different algorithms not only on the basis of the same evaluation corpus but also on the basis of the same pre-processing tools. The evaluation for anaphora resolution is not the same as that for coreference resolution since the output is different in both cases. In anaphora resolution the system has to determine the antecedent of the anaphor; for nominal anaphora any preceding NP which is coreferential with the anaphor is considered as the correct antecedent. On the other hand, the objective of coreference resolution is to identify all coreferential chains. In contrast to anaphora resolution, the MUC-6 and MUC-7 have encouraged fully automatic coreference resolution; also the coreferentially annotated data produced for MUC, however small they are, have provided good grounds for comparative evaluation. Chapter 8 offers detailed discussion of various evaluation issues in anaphora resolution. Outstanding evaluation issues are also discussed in Chapter 9.

Finally, recent work also includes: Morton's (2000) system for resolution of pronouns, definite descriptions, appositives and proper names; the latest version of Harabagiu's COCKTAIL (Harabagiu et al. 2001), which employs, among other things, bootstrapping[68] to check semantic consistency between noun phrases; Stuckardt's (2001) work focusing on the application of the binding constraints on partially parsed texts; Hartrumpf's (2001) hybrid method, which combines syntactic and semantic rules with statistics derived from an annotated corpus; Barbu's (2001) hybrid approach based on the integration of high-confidence filtering rules[69] with automatic learning; and work on anaphora resolution in spoken *dialogues* (Rocha 1999; Martínez-Barco et al. 1999).

5.11 Importance of anaphora resolution for different NLP applications

Recent projects have increasingly demonstrated the importance of anaphora or coreference resolution in various NLP applications. In fact, the successful identification of anaphoric or coreferential links is vital for a number of applications

in the field of natural language understanding including Machine Translation, Automatic Abstracting, Question Answering and Information Extraction.

The interpretation of anaphora is crucial for the successful operation of a **Machine Translation** system. In particular, it is essential to resolve the anaphoric relation when translating into languages that mark the gender of pronouns from languages that do not, or between language pairs that contain gender discrepancies. Unfortunately, the majority of MT systems developed in the 1970s and 1980s did not adequately address the problems of identifying the antecedents of anaphors in the source language and producing the anaphoric 'equivalents' in the target language. As a consequence, only a limited number of MT systems have been successful in translating discourse, rather than isolated sentences. One reason for this situation is that, in addition to anaphora resolution itself being a very complicated task, translation adds a further dimension to the problem. The reference to a discourse entity encoded in a source language anaphor by the speaker (or writer) has not only to be identified by the hearer (translator or translation system) but also re-encoded in a different language. This complexity is variously due to gender discrepancies across languages, to number discrepancies of words denoting the same concept, to discrepancies in gender inheritance of possessive pronouns and to discrepancies in target language anaphor selection (Mitkov and Schmidt 1998). Building on Mitkov and Schmidt's work, Peral et al. (1999) reported specifically upon discrepancies related to the lexical transfer of anaphors between Spanish and English, whereas Geldbach (1999) discussed these discrepancies within a context of Russian to German Machine Translation.

The 1990s have seen an intensification of research efforts in anaphora resolution for Machine Translation. This can be seen in the growing number of related projects which have reported promising results (e.g. Wada 1990; Leass and Schwall 1991; Nakaiwa and Ikehara 1992; Chen 1992; Saggion and Carvalho 1994; Preuβ et al. 1994; Nakaiwa et al. 1994; Nakaiwa and Ikehara 1995; Nakaiwa et al. 1995; Mitkov et al. 1995; Mitkov et al. 1997; Geldbach 1997).[70]

The importance of coreference resolution in **Information Extraction**[71] led to the inclusion of the coreference resolution task in the Message Understanding Conferences, which in turn simulated the development of a number of coreference resolution systems (e.g. Baldwin et al. 1995; Gaizauskas and Humphreys 1996; Kameyama 1997). The **coreference resolution task**[72] takes the form of merging partial data objects about the same entities, entity relationships, and events described at different discourse positions. A recent application of anaphora resolution in information extraction has been reported in a system that identifies and analyses statements in court opinions (Al-Kofani et al. 1999).

Researchers in **Text Summarisation** are increasingly interested in anaphora resolution since techniques for extracting important sentences are more accurate if anaphoric references of indicative concepts are taken into account as well. More generally, coreference and coreferential chains have been extensively exploited for abstracting purposes. Baldwin and Morton (1998) describe a query-sensitive document summarisation technique which extracts sentences containing phrases that corefer with expressions in the query. Azzam, Humphreys and

Gaizauskas (1999) use coreferential chains to produce abstracts by selecting a 'best' chain to represent the main topic of a text. The output is simply the concatenation of sentences from the original document that contain one or more expressions occurring in the selected coreferential chain. Finally, Boguraev and Kennedy (1997) employ their anaphora resolution algorithm (Kennedy and Boguraev 1996) in what they call 'content characterisation' of technical documents.

It should be noted that *cross-document coreference resolution* has emerged as an important trend due to its role in **Cross-Document Summarisation**.[73] Bagga and Baldwin (1998b) describe an approach to cross-document coreference resolution which extracts all sentences containing expressions coreferential with a specific entity (e.g. *John Smith*) from each of several documents. In order to decide whether the documents discuss the same entity (i.e the same *John Smith*), the authors employ a threshold vector space similarity measure between the extracts.

Coreference resolution has proved to be helpful in **Question Answering**. Morton (1999) retrieves answers to queries by establishing coreference links between entities or events in the query and those in the documents.[74] The sentences in the searched documents are ranked according to the coreference relationships, and the highest ranked sentences are displayed to the user. Breck et al. (1999) successfully employ coreference resolution along with shallow parsing and named entity recognition for this application as well. Finally, Vicedo and Ferrández (2000) report improved performance of their question-answering system after applying pronoun resolution in the retrieved documents.

Other applications include the use of coreference constraints to improve the performance (from 92.6% to 97.0%) in the learning of person name structures from unlabelled data (Charniak 2001), and the employment of anaphora resolution to check the correct translation of terminology in a machine-aided translation (Angelova et al. 1998). An interesting recent application (Canning et al. 2000) focuses on readers with acquired dyslexia, helping them to replace pronouns with their antecedents given their difficulty in processing anaphora.[75]

5.12 Summary

Most of the recent and current research in anaphora resolution is geared towards robust and knowledge-poor solutions which often support practical applications such as information extraction and text summarisation. In addition, recent developments benefit extensively from the availability of corpora and demonstrate rising awareness of the necessity of evaluation to show where a specific approach stands. Whereas research in the 1970s and 1980s hardly addressed evaluation issues, no project today would be taken seriously if sufficient evaluation results were not reported. However, it remains difficult to compare approaches in a fair and consistent way since the evaluation is usually done on different sample data and because of the different degrees of pre-processing. For more on this topic see Chapter 8.

Notes

1 See section 5.10.
2 In the experiment conducted (see below in the text) a threshold of 5 occurrences per alternative was used.
3 The authors (Dagan and Itai 1990) offer no preferred solution as to how the candidate should be selected in such cases. Other means mentioned are syntactic heuristics or leaving the case ambiguous for the user to decide on the antecedent.
4 Understandably, if the difference between the selected frequency and the next best ones exceeds a certain threshold.
5 In another experiment, Dagan and Itai (1991) use the ESG (English Slot Grammar) Parser (McCord 1989). See below in the text for an outline of an experiment which combines the authors' approach with that of Hobbs.
6 To provide enough candidates, the authors examined occurrences of *it* after the 15th word of the sentence. These examples provided between 2 and 5 candidates with an average of 2.8 candidates per anaphor.
7 Only the first three candidates of Hobbs's preference list were considered (in all 74 examples used, the correct antecedent was one of the first three candidates). An example was considered amenable to the statistical filter only if at least one of the three candidates had patterns that were more frequent than a specific threshold.
8 A factor of 2 was used in this experiment.
9 This restriction of the search scope to 2 sentences only is apparently based on Hobbs's (1978) finding that about 90% of the anaphoric pronouns *he*, *she*, *it* and *they* have their antecedents either in the same sentence as the pronoun or in the previous one. It should be pointed out, however, that Hobbs's statistics were produced on the basis of 300 pronouns taken from 3 different genres (see Chapter 4, section 4.5.2).
10 The reason for this is that proper nouns are more vulnerable to the statistical approach, due to their higher frequency.
11 Dagan and Itai express the view that the use of semantic classes is not feasible in manually constructed semantic models. It should be pointed out however that whereas this may have been a valid point at the time when their project was undertaken, the emergence of WordNet has made the use of patterns involving semantic classes rather than just words perfectly feasible (Saiz-Noeda et al. 2000). In fact, Saiz-Noeda et al. (2000) report an increase of 19.4% in the success rate when such semantic patterns are added to an anaphora resolution algorithm which does not make use of any semantic information (Ferrández et al. 1998).
12 Co-occurrence patterns are fine-tuned by defining four types of collocations: collocations within the paragraph, collocation within the document, genre-specific collocations and cross-genre collocations (see Chapter 7).
13 The agreement features of an NP here are its number, person, and gender.
14 A phrase F is said to be in the *argument domain* of a phrase G iff F and G are both arguments of the same head.
15 A phrase F is in the *adjunct domain* of G iff G is an argument of a head H, F is the object of a preposition PREP, and PREP is an adjunct of H.
16 A phrase F is in the *noun phrase domain* of G iff G is the determiner of a noun Q and (i) F is an argument of Q, or (ii) F is the object of a preposition PREP and PREP is an adjunct of Q.
17 A phrase F is *contained in* a phrase Q iff (i) F is either an argument or an adjunct of Q, i.e. F is *immediately contained in* Q, or (ii) F is immediately contained in some phrase R, and R is contained in Q.
18 Or as more accurately referred to in Lappin and Leass (1994) 'antecedent binder'.

19 The description is slightly simplified by omitting reference to ID (identifier) for easier understanding.
20 On the other hand, the automatic identification of gender is harder for English than for many other languages. It should be noted that incorrect gender information can lead to a drop in the performance of an anaphora resolution algorithm (see section 5.4 of this chapter for an outline of Kennedy and Boguraev's algorithm; see also Chapter 2, section 2.2.3.5).
21 These experiments were carried out manually (e.g. analysing errors and trying out alternative values with a view to achieving better results). For automatic optimisation procedures, see Chapter 7.
22 See previous note.
23 The reader is referred to Lappin and Leass's paper for further details on the evaluation.
24 See also section 5.2.
25 Recall that Hobbs's algorithm was not implemented in its original version.
26 At the time when Kennedy and Boguraev undertook this research; this statement, however, is still valid today!
27 Recall that configurational information is used in Lappin and Leass's algorithm both in the determination of the salience of a discourse referent (as in the case of head noun emphasis or non-adverbial emphasis) and in the disjoint reference filters (as in syntactic filter on pronoun–NP coreference).
28 There are two possible interpretations of a discourse referent: it could either introduce a new participant in the discourse, or could refer to a previously interpreted discourse referent.
29 Table 5.3 shows separately *Indirect object emphasis* and *Oblique complement emphasis* (complement of a preposition).
30 Note that the coreference class so generated may merge at a later stage with some of the already established ones.
31 The authors note that the set of 306 pronouns excluded 30 occurrences of pleonastic pronouns which could not be recognised by the pleonastic patterns and were manually removed; also manually removed were 6 occurrences of *it* which referred to a VP or prepositional constituent.
32 Kennedy and Boguraev rightly argue that the comparison is not trivial. They maintain that computer manuals are well-behaved texts and it is not clear how RAP's figure would have 'normalised' over a wide range of text types which feature frequent examples of quoted speech and which are not always completely 'clean'.
33 In his paper, Baldwin argues that high precision is vital for tasks such as information retrieval and information extraction.
34 Recall that in its original form Hobbs's algorithm was executed manually.
35 By 'robust' is meant that an antecedent is proposed for every pronoun.
36 See also the machine learning approaches to NP coreference in the following section.
37 The authors use the term definite description (Russell 1905) to indicate definite noun phrases with the definite article *the*, such as *the book*. They are not concerned with other types of definite noun phrases such as pronouns, demonstratives or possessive constructions. See Chapter 1, sections 1.4.2 and 1.6 for further discussion on definite descriptions.
38 As in the example 'They have two daughters and *a son*. I met *the son* last week'.
39 As in 'They have two daughters and *a son*. I met *the boy* last week'.
40 As in 'They have two daughters and a son. I met them all at *the station* last week'.
41 As in 'I left Bill *a valuable book*, but when he returned it, *the cover* was filthy and *the pages* were torn.' See also Chapter 1, section 1.6 for more on indirect anaphora.
42 The tagging here was carried out with the so-called 'Discourse tagging tool' (Aone and Bennett 1994). Apart from marking anaphors and antecedents in an SGML form, the types

of the anaphors (e.g. definite NPs, proper names, quasi-zero pronouns, zero pronouns etc., which are further subdivided as organisations, people, locations, etc.) were also marked.

43 The authors consider the most recent NP in an anaphoric chain as an antecedent.

44 The anaphoric chain parameter has been employed because an anaphor may have more than one 'correct' antecedent (see section 1.2).

45 Aone and Bennett (1995) distinguish between 'quasi-zero pronouns', where zero pronouns refer back to the subject of the initial clause in complex sentences with more than one clause, and simple zero pronouns.

46 It is natural to expect that the results would have been lower if the evaluation had been done on all anaphors as marked by humans. For related discussion see Chapter 8, section 8.2.

47 This definition is somewhat different from that proposed in Baldwin 1997 and Gaizauskas and Humphreys 1996. For further discussion on that topic see Chapter 8, section 8.2.

48 There were four such modes.

49 This corpus consisted of news articles describing business joint ventures.

50 Originally called *governing head information*.

51 Hobbs's algorithm operates on $\bar{N}$ parse-tree nodes that are absent from the Penn Treebank trees.

52 Therefore, Ge, Hale and Charniak's algorithm did not need any automatic pre-processing.

53 Features are constraints or preferences in the terminology of this book.

54 Recall that coreferential chains constitute equivalence classes (see section 1.2).

55 It should be noted that NPs which do not contain definite articles can also be of definite status (see example (1.10), section 1.2).

56 These results place the clustering algorithm between the best performing and worst performing coreference resolution programs at MUC-6, outperforming the only other corpus-based learning approach.

57 Cardie and Wagstaff admit in their paper though (section 5, note 4) that it is not clear whether clustering can be regarded as a 'learning' approach.

58 As introduced in Chapter 1, nominal anaphora is exhibited by pronouns, definite noun phrases and proper names referring to an NP antecedent.

59 Originally termed 'sort consistency'.

60 An extended version of this paper was recently published (Tetreault 2001).

61 In his project, Tetreault simplifies the notion of utterance to a sentence.

62 While the original version of the S-list approach incorporates both semantics and syntax, a shallow modification was implemented for Tetreault's study.

63 The corpus was the one used by Ge et al. (1998) – see also section 5.8 of this chapter and sections 6.2 and 6.3 of Chapter 6. Sentences were fully bracketed and had labels that indicated word-classes and features (gender, number).

64 For this experiment, Tetreault implemented two separate versions of his own algorithm (LRC-1 and LRC-N) and of Strube's approach (Strube-1 and Strube-N).

65 In its original version the BFP did not process intrasentential anaphors but for this experiment the LRC intrasentential technique was used to resolve pronouns that could not be resolved by the BFP.

66 This work served as a basis for the development of algorithms for resolution of definite descriptions (Muñoz and Palomar 2000, 2001) and zero pronouns (Ferrández and Peral 2000).

67 For more details on this approach see Chapter 7, section 7.3.

68 Bootstrapping is a new machine-learning technique presented by Riloff and Jones (1999).

69 The high-precision rules were those proposed in CogNIAC by Baldwin (1997).

70 A brief survey of anaphora resolution in Machine Translation can be found in Mitkov (1999a).

71 **Information Extraction** is the automatic identification of selected types of entities, relations or events in free text. It covers a wide range of tasks, from finding all the company names in a text, to finding all the murders, including who killed whom, when and where. Such capabilities are increasingly important for sifting through the enormous volumes of on-line text for the specific information required (Grishman 2002).

72 Recall, however, that coreference and anaphora are not the same phenomenon (see Chapter 1, section 1.2).

73 **Cross-Document Summarisation** is the task of summarising a collection of thematically related documents.

74 The coreference relationships that Morton's system supports are identity, part–whole and synonymy.

75 Acquired dyslexia is a form of aphasia which results in reading impairment. Some readers suffering from this disability are unable to process pronominal anaphora, especially if there is more than one candidate for antecedent.

CHAPTER SIX

The role of corpora in anaphora resolution

6.1 The need for anaphorically or coreferentially annotated corpora

Since the early 1990s research and development in anaphora resolution have benefited from the availability of corpora, both raw and annotated. While raw corpora, successfully exploited for extracting collocation patterns (Dagan and Itai 1990, 1991), are widely available, this is not the case for corpora annotated with coreferential links. The annotation of corpora is an indispensable, albeit time-consuming, preliminary to anaphora resolution (and to most NLP tasks or applications), since the data they provide are critical to the development, optimisation and evaluation of new approaches.[1] The automatic training and evaluation of anaphora resolution algorithms require that the annotation cover not only single anaphor–antecedent pairs, but also anaphoric chains, since the resolution of a specific anaphor would be considered successful if any preceding element of the anaphoric chain associated with that anaphor were identified. Unfortunately, anaphorically or coreferentially **annotated corpora** are not widely available and those that exist are not of a large size.

The act of annotating corpora follows a specific **annotation scheme**, an adopted methodology prescribing how to encode linguistic features in text. Annotation schemes usually comprise a set of ASCII strings such as labelled syntactic brackets to delineate grammatical constituents or word class tags (Botley 1999). Once an annotation scheme has been proposed to encode linguistic information, user-based tools (referred to as **annotation tools**) can be developed to facilitate the application of this scheme, making the annotation process faster and more user-friendly. Finally, an **annotation strategy** is essential for accurate and consistent mark-up.

This chapter will briefly introduce the few existing corpora annotated with anaphoric or coreferential links and will then present the major annotation schemes that have been proposed. Next, several tools that have been developed for the annotation of anaphoric or coreferential relationships will be outlined. Finally, the chapter will discuss the issue of annotation strategy and inter-annotator agreement.

6.2 Corpora annotated with anaphoric or coreferential links

One of the few anaphorically annotated resources, the **Lancaster Anaphoric Treebank** is a 100 000-word sample of the Associated Press (AP) corpus (Leech and Garside 1991), marked up with the UCREL[2] anaphora annotation scheme (see section 6.3). The original motivation for constructing this corpus was to investigate the potential for developing a probabilistic anaphora resolution program. In late 1989, an agreement was made between the UCREL and IBM Yorktown Heights teams, with funding from the latter, to construct a corpus marked to show a variety of anaphoric or, more generally, cohesive relationships in texts.

Before the anaphoric relationships were analysed and encoded, each text already included the following annotations:

(i) A reference code for each sentence (e.g. A001 69, A001 70, A009 90, A009 91).
(ii) A part-of-speech tag for each word.
(iii) Parsing labels indicating the main constituent structure for each sentence.

The original AP corpus was divided into units of approximately 100 sentences,[3] and the syntactic and anaphoric markings were carried out on each of these units, so that the anaphoric reference numbering began afresh with each unit.

The **MUC coreference task** (MUC-6 and MUC-7) gave rise to the production of texts annotated for coreferential links for training and evaluation purposes. The annotated data which complied with the MUC annotation scheme (see section 6.3) was mostly from the genre of newswire reports on subjects such as corporate buyouts, management takeovers, airline business and plane crashes.[4] All the annotated texts amounted to approximately 65 000 words.[5]

A **part of the Penn Treebank**[6] was annotated to support a statistical pronoun resolution project at **Brown University** (Ge 1998). The resulting corpus contains 93 931 words and 2463 pronouns. In addition to providing information on coreference between pronouns and noun phrases, or generally between any two noun phrases, pleonastic pronouns were also marked.

A corpus containing around 60 000 words, annotated in a way similar to the MUC annotation scheme with the help of the annotation tool ClinKA (see section 6.4) has been produced at the **University of Wolverhampton** (Mitkov et al. 2000). The corpus features fully annotated coreferential chains and covers texts from different user manuals (printers, videorecorders, etc.).

An ongoing project conducted by members of the **University of Stendahl, Grenoble**, and **Xerox Research Centre Europe** (Tutin et al. 2000) is to deliver a million-word corpus annotated for anaphoric and cataphoric links. The annotation is limited to anaphor–closest antecedent pairs rather than full anaphoric chains[7] and involves the following types of anaphors: third person personal pronouns, possessive pronouns, demonstrative pronouns, indefinite pronouns, adverbial anaphors and zero noun anaphors.

Texts annotated for **coreferential links in French** are also reported by Popescu-Belis (1998). The first one, marked up in both MUC's and Bruneseaux

and Romary's schemes (see section 6.3), is part of a short story by Stendahl (*Victoria Accoramboni*). The second one, produced at LORIA,[8] is part of a novel by Balzac (*Le Père Goriot*) and follows Bruneseaux and Romary's scheme. In the first sample all referential expressions (altogether 638) were marked, whereas in the second sample only entities representing the main characters in the novel were annotated (a total of 3812).

Finally, as a consequence of the increasing number of projects in multilingual anaphora resolution, the need for parallel bilingual and multilingual corpora annotated for coreferential or anaphoric links has become obvious. To the best of this writer's knowledge there are no such corpora yet apart from a small-size **English–Romanian corpus** developed for testing a bilingual coreference resolution system (Harabagiu and Maiorano 2000). Another **parallel English–French** corpus covering texts from technical manuals was annotated for coreferential links at the University of Wolverhampton and exploited by an English and French bilingual anaphora resolution algorithm (see Chapter 7, section 7.3.2). The English part of the corpus contains 25 499 words and the French part 28 037 words.

It should be noted that annotated corpora are an invaluable resource not only to computational linguistics projects but also to different types of linguistic analysis. A corpus of identifiable surface markers of anaphoric items and relationships that can be used to examine current theories will undoubtedly prove to be very useful in any linguistic studies focusing on anaphora.

6.3 Annotation schemes

In recent years, a number of corpus annotation schemes for marking up anaphora have come into existence. Notable amongst these are the UCREL anaphora annotation scheme applied to newswire texts (Fligelstone 1992; Garside et al. 1997) and the SGML-based (MUC) annotation scheme used in the MUC coreference task (Hirschman 1997). Other well-known schemes include Rocha's (1997) scheme for annotating spoken Portuguese, Botley's (1999) scheme for demonstrative pronouns, Bruneseaux and Romary's scheme (1997), the DRAMA scheme (Passonneau and Litman 1997), the annotation scheme for marking up definite noun phrases proposed by Poesio and Vieira (1998) and the MATE scheme for annotating coreference in dialogues proposed by Davies et al. (1998).

The **UCREL scheme** was initially developed by Geoffrey Leech (Lancaster University) and Ezra Black (IBM). The coding method was then elaborated and tested by its application to corpus texts by the UCREL team, whose feedback triggered further elaboration and testing for the scheme. This development cycle was iterated several times.

The scheme allows the marking of a wide variety of cohesive features ranging from pronominal and lexical NP anaphora through ellipsis to the generic use of pronouns. Special symbols added to anaphors and antecedents can encode the direction of reference (i.e. anaphoric or cataphoric), the type of cohesive relationship involved and the antecedent of an anaphor, as well as various

semantic features of anaphors and antecedents. For example, the following text fragment (Tanaka 2000) has been encoded using some of the features of this scheme:

(6.1) Anything (108 Kurt Thomas 108) does, <REF = 108 he does to win. Finishing second, <REF = 108 he says, is like finishing last.

As example (6.1) shows, the UCREL scheme brackets antecedent noun phrases and indicates the direction of pronominal references with arrows ('<' for anaphoric references and '>' for cataphoric references) whilst a string denotes the type of relationship involved (in this case reference rather than substitution). An index number is uniquely assigned to each antecedent and any subsequent references to it. Since elements that are related anaphorically share the same index number, anaphoric chains can be readily identified, either manually or automatically (Botley and McEnery 1991).

The 'indexing mode' may differ depending on whether the antecedent is a single discourse entity, consists of multiple discourse entities, is either one or the other type of discourse entity, or it is not quite certain. For instance, in the following example (Tanaka 2000), the antecedents of the anaphoric pronouns *they* and *their* are felt to be probably *Phil Esposito* (indexed '28') and *Ron Greschner* (indexed '29') but this is not quite certain, hence an uncertainty indicator '?' is marked before each index number:

(6.2) Right wing (27 Anders Hedberg 27) withdrew because of shoulder problems and was replaced by center (28 Phil Esposito 28), while defenseman (29 Ron Greschner 29) took over for teammate (30 Barry Beck 30) (elbow).[9] Monday, <REF = ?28,?29 they devoted much of <REF = ?28,?29 their time to the tenure of Alan Eagleson, executive director . . .[10]

Since the process of identifying anaphoric relationships is a complex one and may lead to disagreement between the annotators (for further discussion on this topic see section 6.5), it was decided that the UCREL scheme should avoid too detailed a level of analysis. This decision was also prompted by the need to produce a substantial volume of annotated texts for development of an anaphora resolver and, as a consequence, by the need to speed up the annotation. Finally, although the coding scheme was influenced by Halliday and Hasan (1976), the resulting corpus was hoped to be as theoretically neutral as possible, so that it could be used by researchers from a wide range of divergent theoretical positions. This became a further reason for opting for a less detailed type of analysis.

The resulting scheme, therefore, reflects the resolution of the 'tension' between the practical requirement to avoid too detailed a level of analysis (caused by the inter-annotator consistency, the speed of marking up, and the demand for theoretical neutrality) and the requirement to meet potential users' theoretical interests as much as possible (Tanaka 2000).

In addition to applying the scheme to the data produced for the MUCs, the SGML-based **MUC annotation scheme** (Hirschman 1997) has been used by a number of researchers to annotate coreferential links (Gaizauskas and

Humphreys 1996; Mitkov et al. 1999). Given an antecedent A and an anaphor B, where both A and B are strings in the text, the basic MUC coreference annotation has the form

<COREF ID ="100"> A </COREF> . . .
<COREF ID ="101" TYPE = IDENT REF ="100"> B </COREF>

So, for example, in *The Kenya Wildlife Service estimates it loses $1.2 million a year in park entry fees because of fraud*[11] the anaphor *it* and its antecedent *The Kenya Wildlife Service* would be marked up as

<COREF ID ="100"> The Kenya Wildlife Service </COREF> estimates
<COREF ID ="101" TYPE = IDENT REF ="100"> it </COREF> loses $1.2 million a year in park entry fees because of fraud.

In the MUC scheme, and in the above example, the attribute ID uniquely denotes each string in a coreference relation, REF identifies which string is coreferential with the one which it tags, TYPE indicates the type of relationship between anaphor and antecedent and IDENT indicates the identity relationship between anaphor and antecedent.[12] The MUC scheme only covers the identity (IDENT) relation for noun phrases and does not include other kinds of coreference relations such as set/subset, part/whole,[13] etc. In addition to these attributes, the annotator can add two more, the first of which is MIN, which is used in the automatic evaluation of coreference resolution systems. The value of MIN represents the smallest continuous substring of the element that must be identified by a system in order to consider a resolution correct. Secondly, the attribute STATUS can be used and set to the value OPT. This information is used to express the fact that markup of the tagged element is optional. Dates, currency expressions and percentages are considered noun phrases.

The MUC scheme stipulates which noun phrases should be marked up as coreferential and when. For example, bound anaphors and their antecedents are regarded as coreferential and in the example *Most lotions don't give percentages of their ingredients*,[14] a coreference link between *Most lotions* and *their* is recorded. Appositional phrases are considered coreferential to the noun phrase to which they apply, even if they are indefinite (*Reza Khatami, the brother of the President of the Islamic Republic of Iran*,[15] but also *John Smith, a 10-year MUC veteran*). Similarly, all predicate nominals, including indefinite ones, are regarded as coreferential with the subject, allowing a coreference link to be marked not only between *Tony Blair* and *the Prime Minister* in *Tony Blair is the Prime Minister of Britain* but also between *Tony Blair* and *a Prime Minister* in *Tony Blair is a Prime Minister*. Coreference is not recorded if the text only asserts the possibility of identity between two noun phrases such as *Marcelo Rodriguez may be the only person in Los Angeles complaining of too much exposure*;[16] nor can appositives be coreferential if they are negative (*Oliver James, never one for late-night socialising, arrived home at 10.00* . . .). In particular, two NPs[17] should be recorded as coreferential if the text asserts them to be coreferential at any time (Hirschman and Chinchor 1997). The MUC annotation scheme also recommends how much of the NP to annotate. The head of a noun phrase is considered as the minimum string to be annotated

(such as *task* in the NP *coreference task* or *contract* in the NP *the last contract you will get*). The maximum noun phrase includes all text which may be considered a modifier of the noun phrase (such as *The Love Bug*, or *The Kenya Wildlife Service, which runs the country's national parks*[18] or *A computer virus that may have originated in the Philippines*[19] . . .).

The view adopted in this book as to the coreferential status of indefinite predicate nominal and indefinite, unspecific appositives is different and they are not considered coreferential with their subjects or with the NP to which they apply (see section 1.2). Van Deemter and Kibble (1999) have criticised the MUC coreference annotation scheme in that 'it goes beyond annotation of coreference as it is commonly understood' since it marks non-referring NPs (and therefore which cannot corefer) such as quantifying NPs (e.g. *every man*, *most computational linguists*) as part of the coreferential chain.[20] Van Deemter and Kibble also express their reservation regarding the marking of indefinite NPs and predicate NPs as possibly coreferential, arguing that if in the example

(6.3) Henry Higgins, who was formerly sales director of Sudsy Soaps, became president of Dreamy Detergents.

Henry Higgins, *sales director of Sudsy Soaps* and *president of Dreamy Detergents* are all marked as standing in the IDENT relation, and if two NPs should be recorded as coreferential if the text asserts them to be coreferential at any time, then one could conclude that Henry Higgins is presently sales director of Sudsy Soaps as well as president of Dreamy Detergents, which is not what the text asserts.[21]

However, despite its imperfections, the MUC scheme has the strength of offering a standard format. Also, although it has been designed to mark only a small subset of anaphoric and coreferential relations, the SGML framework does provide a useful starting point for the standardisation of different anaphoric annotation schemes.

The **DRAMA**[22] **scheme** (Passonneau 1996; Passonneau and Litman 1997) identifies anaphors and antecedents in a text, and marks coreference relationships between them. Although similar to the MUC scheme, the DRAMA scheme classifies and marks different kinds of bridging relationships. DRAMA includes instructions for dealing with some difficult problems of identifying the 'markable' entities in dialogues, and allows a wider set of these than does the MUC scheme, such as clauses, verb phrases or adjectival phrases which might be the antecedents of certain anaphors such as *it* and *that*.

Bruneseaux and Romary's scheme (1997) identifies anaphors and antecedents in the text and marks the relationships between them, as is the case with other schemes such as MUC, DRAMA and UCREL. An innovation of this scheme is that it allows references to the visual context to be encoded, due to the fact that the corpora annotated include conversational data in human–computer interaction systems where speakers are using a geological simulation program. This scheme also allows the marking of deixis in the form of pointing and mouse-click gestures.

Poesio and Vieira's (1998) first scheme classified definite noun phrases and their textual relationships with other NPs rather than linking referential

expressions as in the MUC and DRAMA schemes. As a result of this, the number of markable entities in this scheme is much more limited. In addition to classifying definite NPs, **Poesio and Vieira's (1998) second scheme** also marked the referential link between referring definite NPs and their antecedents. This latter scheme allowed a wider range of markables than the first scheme.

The **MATE scheme for annotating coreference in dialogues** (Davies et al. 1998) draws on the MUC coreference scheme, adding mechanisms for marking up further types of information about anaphoric relations as done in the UCREL, DRAMA and Bruneseaux and Romary schemes. In particular, this scheme allows for the markup of anaphoric constructs typical in Romance languages such as clitics and of some typical dialogue phenomena. The scheme also provides for the markup of ambiguities and misunderstandings in dialogue. The MATE scheme consists of a *core scheme* and an *extended scheme*. The core scheme has been developed with a two-fold objective in mind: to produce a scheme which (i) is likely to be reliable in terms of the inter-annotator agreement and which (ii) offers coverage roughly analogous to that offered by the MUC scheme. The extended scheme enables more detailed annotation of various relationships which can occur between discourse entities such as bound anaphora (*Nobody* likes to lose *his job*), set relationship (see example (6.5)), possessive relationship, event relationship, etc.; examples of all relationships involved can be found in Davies et al. (1998). As expected, the inter-annotator agreement for marking up these more complex relations is considerably lower: once one moves beyond the IDENT relation, it can be difficult to decide how to classify the link between two elements (Poesio and Vieira 1998).

Each discourse entity (de) in the text is given an ID number and <de> tags corresponding to the <coref> tags in the MUC scheme. The <link> pointer specifies the type of link between two discourse entities and lists their IDs as values of the ARGS attribute. For instance, the IDENT relation between the two mentions of *orange juice* in the dialogue

(6.4) When do we have *orange juice* at Elmira?
We have *orange juice* at Elmira at 6 a.m.

would be tagged as follows:

> When do we have <de ID = "01"> orange juice </de> at Elmira?
> We have <de ID = "02"> orange juice </de> at Elmira at 6 a.m.
> <link type = "ident" args = "01 02">

Also, the example

(6.5) *The kids* went to a party last weekend. *Paul* wanted to wear his new suit, but *Jane* insisted on wearing her jeans.

is annotated as[23]

> <de ID = "85"> The kids </de> went to a party last weekend. <de ID = "86"> Paul </de> wanted to wear his new suit, but <de ID = "87"> Jane </de> insisted

on wearing her jeans
<link type = "element" args = "86 85">
<link type = "element" args = "87 85">

where the 'element' link represents a set relation holding when one discourse entity is an element of the set denoted by the other discourse entity.

The strength of the MATE scheme is that while based on the widespread MUC scheme, and adopting the popular SGML standard, similarly to the UCREL scheme it covers a rich variety of anaphoric relations, which makes it a promising general-purpose framework.

Tutin et al.'s (2000) XML-based scheme supports the annotation of a variety of anaphoric relations such as coreference, set membership, substitution, sentential anaphora[24] and indefinite relation which includes all cases not covered by the first four types such as bound anaphora.[25] The annotation scheme encodes the boundaries of each expression, the link between two expressions and the type of relationship between them. The boundaries are marked by means of the <exp> and </exp> tags, with an ID number inserted in <exp>. The link between an anaphor and antecedent is encoded by the <ptr> tag which is added to the anaphor and which is represented as a string containing an *src* antecedent label and a pointer to the ID number of the antecedent. Finally, the type of relation is marked by the 'type' attribute in the <ptr> string. As an illustration, (6.6)

(6.6) *Des quatre locomotives de Savoie*, *l'une* est à redresseurs [. . .]. Les trois autres montrent une sorte de coexistence . . . .

Of the four locomotives of Savoie, one is of the erector type [. . .]. The three others show a kind of coexistence . . .

is annotated as follows ('mde' stands for 'membre de' which is the French expression for *set membership*).

<exp id = "f50"> Des quatre locomotives de Savoie </exp>, <exp id = "f51"> <ptr type = "mde" src = "f50"/> l'une </exp> est à redresseurs [. . .].> <exp id = "f52"> <ptr type = "mde" src = "f50" /> Les trois autres</exp> montrent une sorte de coexistence.

Tutin et al.'s scheme can also encode special cases such as identity-of-sense anaphora, ambiguity of anaphors and coordinated (split) antecedents.

Ge's annotation (1998) covers five kinds of relationships involving pronouns. The author marks pronouns which have explicit nominal antecedents, pronouns with split antecedents, pronouns pointing to an action or event not represented by a single noun phrase, and two types of pleonastic pronouns: those that are not specific enough and those that appear in cleft constructions.

Rocha (1997) described a detailed annotation scheme for marking anaphoric references in a corpus of spoken Portuguese dialogues, and extracts from the London–Lund corpus. Rocha's scheme explores the relationship between anaphora and the topic structure of discourse, by signalling the discourse, segment and subsegment topics. In addition to being able to mark discourse

structure features, Rocha's scheme can also mark different aspects of anaphors, such as the type of anaphor (e.g. 'subject pronoun' or 'full noun phrase'), the type of antecedent (implicit, non-surface or explicit, surface antecedent), the topicality status of the antecedent (whether the antecedent is the discourse topic, segment topic or subsegment topic) and the type of knowledge required for the processing of the anaphor (such as syntactic, collocational or discourse knowledge). Rocha's scheme allows for anaphora in spoken (and presumably written) texts to be analysed according to a rich variety of interrelated factors, in a way which extends beyond the descriptive analysis of Halliday and Hasan (which is largely implemented in the UCREL annotation scheme outlined above); however, it is very labour-intensive to apply.

Botley's scheme (1999) describes the different functions of anaphoric demonstratives in written and spoken texts. Essentially, this scheme classifies demonstrative anaphors according to five distinctive features, each of which can have one of a series of values. The features employed are recoverability of the antecedent (e.g. directly recoverable, indirectly recoverable, non-recoverable, not applicable e.g. exophoric), direction of reference (anaphoric, cataphoric, not applicable), phoric type (referential, substitution, not applicable), syntactic function (non-modifier, non-head, not applicable) and antecedent type (nominal antecedent, prepositional/factual antecedent, clausal antecedent, adjectival antecedent, no antecedent).

Botley (1999), Davies et al. (1998) and Tutin et al. (2000) provide further discussion on issues relating to the annotation of anaphors and the annotation schemes.

6.4 Annotation tools

In order to help the human annotator it is necessary to provide him/her with a tool which makes it possible to quickly identify the entities in the discourse and the relations between them. A good graphical interface offers the human annotator trouble-free and efficient interaction with the annotated text. It should also display the resulting annotation in a way that is easy for a user to interpret, hiding unnecessary or hard-to-read aspects of the annotation, such as raw SGML encoding.

The first tool for annotation of anaphoric links, **XANADU**, written by Roger Garside at Lancaster University, is an X-windows interactive editor which offers the user an easy-to-navigate environment for manually marking pairs of anaphors–antecedents within the UCREL scheme (Fligelstone 1992). In particular, XANADU allows the user to move around a block of text, displaying circa 20 lines at a time. Users can use a mouse to mark any segment of text to which they wish to add some labelling. Apart from the text window, there are two primary windows which are always displayed. The first of these contains a set of 'command buttons', which for the most part refer specifically to categories of anaphora that are recognised as being within the scope of the scheme. The second window contains a list of already identified antecedents.

The **DTTool** (Discourse Tagging Tool) enables the annotation of anaphoric relations in Japanese, Spanish and English (Aone and Bennett 1994). This is done in a graphical manner – by the colour-coding of different types of anaphors (e.g. third person pronoun, definite NP, proper name, zero pronoun, etc.) and antecedents which are displayed on the screen with arrows linking them. For instance, third person pronouns referring to organisation nouns are highlighted in orange, while definite NPs referring to a person are highlighted in azure blue. The annotated data can be viewed in five different modes: all tags, each anaphor–antecedent pair, all anaphor–antecedent pairs of the same type, all anaphoric chains and the text without any tags.

The **Alembic Workbench** was developed at MITRE, and has been used among other things to mark up coreference relations. The Alembic Workbench is a component of the trainable multilingual information extraction system Alembic (Day et al. 1997) and is designed to enable the rapid manual or semi-automatic production of data for training and testing. The data include annotated parts of speech, named entities, coreference chains, etc. For the coreference annotation task, the workbench features a right window which produces a sorted list of all tagged elements to facilitate finding the coreferring expressions. The semi-automatic mode extends to simple annotation tasks such as tagging named entities. For instance, if a certain string has been marked as a proper name, the tool proposes that the same string be marked as proper name if it appears again in the text. The Alembic Workbench offers a choice of tag sets, including all those necessary for the MUC, and provides a graphical interface which allows the modification of existing tags and the addition of new ones. In addition, the users of the system are able to construct their own task-specific annotation schemes.

Referee is a discourse annotation and visualisation tool which operates in three modes – reference mode, segment mode and dialogue mode (DeCristofaro et al. 1999). In *reference mode*, the user can mark words or expressions by associating features (e.g. syntactic function, pronominalisation, distance, definiteness, etc.) with each of them and assigning coreference. The annotated information can then be easily retrieved. For example, clicking on a specific word will not only display the values of its features (e.g. syntactic function = subject, pronominalisation = no) but will also highlight all other expressions which corefer with it. At this point the user can update the coreference links or feature values, or store additional information. In *segment mode* the user can partition the text into arbitrarily nested and overlapping segments, whereas the *dialogue mode* enables him/her to code a dialogue by breaking it into turns.

CLinkA is a tool for annotating coreferential links which operates by default on the MUC scheme, but also gives the user the option to define his/her own annotation scheme (Orasan 2000). The program uses two types of tags: one for marking the initial mention of an element in a coreference chain and one for marking the remaining element of that chain. Similarly to the Alembic Workbench, the following attributes can be added to each tag: (i) counters which identify uniquely every element in the coreference chain and are generated automatically by the program, (ii) index numbers which are uniquely assigned to

each antecedent and any subsequent references to it and (iii) values specified by each annotator such as MIN, which stands for a 'minimal noun phrase' (see the section above on the MUC annotation scheme). CLinkA also displays a right window listing all identified NPs in the text, which helps the annotator in linking the coreferential items. The process of annotation is kept as simple as possible. As an illustration, boundaries of entities are identified by mouse clicks and the addition of an entity to an existing chain can be done by clicking on an element already in the chain. To speed up the annotation, the tool offers several features for semi-automatic marking. For instance, identical strings are likely to be in the same coreferential chain and each time the program establishes identity between a new string and an already marked one, users are asked if they would like to add the new string to the chain of the preceding identical string. CLinkA also features a graphical interface for comparing the annotations carried out by different annotators, displaying them in different colours. The tool is language independent and has been used for annotating coreferential links in English, Spanish and Bulgarian. This language independence has been facilitated by the fact that CLinkA is implemented in Java which supports Unicode. In addition, the tool operates on any platform that has a Java virtual machine.

Day et al. (2000) describe a **cross-document annotation toolset** that supports among other things the annotation of cross-document coreference. Individual documents are annotated with pointers to a single entity repository. In turn, the repository maintains references to all the documents (and locations within documents) where information has been individually annotated. Other recent developments relevant to the annotation of coreferential links are the general-purpose annotation tool **FAST** (Friendly Annotator for SGML Texts) which operates in three different modes: manual, semi-automatic and fully automatic (Barbu 2000) and the ongoing work on **ATLAS**, 'flexible and extensible architecture for linguistic annotation' which will include support for multidomain, multilayered and multilinked annotations (Bird et al. 2000).

In spite of their attractive features, the tools for annotating anaphoric and coreferential relations are still largely based on manual antecedent identification and labelling, which is not always easy and straightforward. The manual annotation process imposes a considerable demand on human time and labour. The main reason why annotation of anaphoric or coreferential data has not yet been able to benefit from the level of automation enjoyed by its lexical, syntactic and semantic counterparts (part-of-speech tagging, parsing or word-sense disambiguation) is the complexity of the linguistic phenomena of anaphora and coreference. However, with a view to accelerating the marking-up process, the idea for semi-automatic annotation of anaphoric and coreferential links has already been put forward (Mitkov 1997b). It has been suggested that an annotation tool could employ a high-precision anaphora resolution system to propose antecedent(s) which are then post-corrected by a human annotator by either choosing from a list of returned antecedents or by manually marking up omitted anaphoric relationships. On a less ambitious but more practical scale CLinkA already provides some level of semi-automatic marking such as the inclusion of two matching NPs in the same coreferential class after consultation with the user.

6.5 Annotation strategy and inter-annotator agreement

The annotation of anaphoric or coreferential relations is a notoriously difficult, time-consuming and labour-intensive task even when focusing on one single variety of the phenomenon.[26] Compared with syntactic analysis, the discourse-level analysis of anaphoric relations involves a much more interpretative process, and the possibility of disagreement in interpretation between annotators is much greater than in syntactic analysis. Fligelstone (1992) notes that 'The nature of the task, with its heavy reliance on interpretation, suggests that it may prove impossible to achieve such a high degree of inter-analyst consistency as with the parsing scheme . . .'.[27] The complexity of the task imposes a restriction that the annotation process should not follow a detailed level of analysis (as in the case of the UCREL and MUC schemes) but focus instead on identity relation only (MUC scheme). The experience with the MUC annotation scheme shows that even within the narrow domain of NP coreference it is not always easy to decide which NPs should be marked as coreferential. This is indicative of how complex anaphora and coreference are. As a consequence, the annotation process is often considered to be far from reliable in that inter-annotator agreement may be disappointingly low. For related discussion on the complexity of anaphora and coreference see van Deemter and Kibble (1999).

Given the complexity of the anaphora and coreference annotation task, a recent project carried out at the University of Wolverhampton[28] (Mitkov et al. 2000) adopted a less ambitious but clearer approach regarding the variety of anaphora annotated. This move was motivated by the fact that (i) annotating anaphora and coreference is a very difficult task and (ii) the aim was to produce annotated data for the types of anaphora most widely used in NLP: that of identity-of-reference direct nominal anaphora featuring a relation of coreference between the anaphors (pronouns, definite descriptions or proper names) and any of their antecedents (non-pronominal NPs).[29] The annotation covered identity-of-reference direct nominal anaphora, which included relationships such as specialisation, generalisation and synonymy, but excluded part-of and set membership relations that are considered instances of indirect anaphora. Whilst it was obvious that such a corpus would be of less interest in linguistic studies, it was believed that the vast majority of NLP work on anaphora and coreference resolution (and all those tasks which rely on it) would be able to benefit from this corpus by using it for evaluation and training purposes. The view was that the trade-off of a wide-coverage, but complicated and potentially error-prone, annotation task with low consistency across annotations for a simpler but more reliable annotation task with a NLP-oriented end product was a worthwhile endeavour.

An annotation strategy in the form of guidelines outlining what to annotate and when to annotate it, and recommending the best annotation practice, can be very helpful to the annotators, and could enhance the annotation consistency and the inter-annotator agreement which are often disappointingly low. The guidelines produced for the objectives of the Wolverhampton project cited above discuss which classes of anaphora should be annotated (identity-of-reference

direct nominal anaphora in this particular project) and what are the markables (all kinds of NPs including base, complex and coordinated), and advise in which cases two NPs should be marked as coreferential (e.g. definite descriptions in copular relation). These guidelines also explain which types of anaphora or coreference should not be annotated (e.g. identity-of-sense anaphora, bound anaphora, cross-document coreference), what are not markables and in which cases NPs should not be marked as coreferential (e.g. copular relation when one of the NPs is indefinite). Useful annotation practices used to improve the inter-annotator agreement included printing out the whole text prior to annotation so that the annotators familiarise themselves with the text, identifying all the noun phrases to be marked as either initial or subsequent mentions, making a note of all troublesome or ambiguous cases and discussing them with other annotators, and ensuring that the annotation is done in one intensive period, as sporadically annotating a file can lead to the annotator's having to re-read the document for familiarisation several times. For more details see Mitkov et al. (2000).

On the other hand, however difficult the annotation task is, in order to ensure that a specific text or corpus is marked up as correctly and objectively as possible, it is necessary that in addition to adhering to annotation guidelines, each sample be marked by at least two annotators independently. Since there is no guarantee that the annotators will agree on how each instance should be annotated, and with a view to performing quality checks, Mitkov et al. (2000) describe a program which matches all annotations and flags up instances marked up differently by the annotators. The program works by extracting the full coreference chains from two annotated files and then producing the chains that are present in one file but are not identical to any chains in the file being compared. Similarly, differing elements are written and the number of elements shared between the files is returned. This allows a qualitative assessment of the differences between the annotations as well as subsequent discussion and adjudication.

Orasan (2000) and Mitkov et al. (2000) used the following measure to compute the similarity/closeness of the annotations produced by two different annotators:

$$\mu = \frac{2C}{A + B}$$

where A is the number of items marked by the first annotator, B is the number of items marked by the second and C is the number of items which were marked by both annotators.[30] If both annotators marked the same items then the agreement is equal to 1, otherwise it is a value greater or equal to 0 and less than 1 ($0 \leq \mu \leq 1$). Mitkov et al. (2000) found that the average proportion of shared elements on the corpora annotated varied from 0.66 to 0.72.

Hirschman et al. (1998) conducted a small-scale analysis on the inter-annotator agreement in the coreference task as defined by the Message Understanding Conferences (MUC-6 and MUC-7). The study, which was based on the annotation produced by two annotators, suggested that only 16% of the disagreement cases represented genuine disagreement about coreference since the remainder of the cases were typographical errors or errors of omission. Initially the agreement

was in the low 80s but in order to improve it, the authors ran several experiments. In one of the experiments they separated the tagging of the noun phrases from the linking of the actual coreferring expressions, and as a result the inter-annotator agreement climbed to the low 90s.[31]

6.6 Summary

This chapter has highlighted the importance of corpora for anaphora and coreference resolution. Corpora annotated with anaphoric or coreferential links are particularly important for the research in anaphora resolution. They are invaluable resources for obtaining empirical data and rules in the building of new anaphora resolution approaches and for training, optimisation and evaluation of the existing approaches. The production of annotated corpora is a challenging and time-consuming task, which follows a specific annotation scheme and strategy, and uses task-specific annotation tools. The chapter has also outlined the existing corpora annotated for coreference, the annotation schemes proposed and the tools developed, and has discussed the related issue of annotation strategy and inter-annotator agreement.

Notes

1 For further details, see Chapters 5 and 7.
2 UCREL (Unit for Computer Research on the English Language) comprises members of the Departments of Computing and of Linguistics and Modern English Language in Lancaster University. Since 1980, one of the main research goals of UCREL has been the creation of annotated corpora.
3 Occasionally a text sample did not include the beginning of the original text.
4 Some of the articles are also about reports on scientific subjects. Management of defence contracts is covered and there are also reports on music concerts, legal matters (lawsuits, etc.) and broadcasting business.
5 This figure is based on data and information kindly provided to us by Nancy Chinchor.
6 The Penn Treebank is a corpus of manually parsed texts from the *Wall Street Journal*.
7 This limitation makes the corpus more suitable for theoretical linguistic research than for evaluation and testing anaphora resolution systems where full anaphoric or coreferential chains are needed (see section 1.2).
8 LORIA stands for Laboratoire Lorrain de Recherche en Informatique et ses Applications.
9 The reference code of this sentence in the AP is A007 17.
10 The reference code of this sentence is A007 18.
11 Adapted from *Time*, 31 July 2000, p. 4.
12 See also identity-of-reference anaphora in section 1.7.
13 See also indirect anaphora as defined in section 1.6.
14 *Time Europe*, 28 August 2000.
15 *Time Europe*, 28 August 2000.
16 *Time Europe*, 21 February 2000, p. 22.
17 The original statement uses the term 'markables' which in the coreference task are nouns, noun phrases and pronouns.

18 *Time*, 31 July 2000, p. 4.
19 *Time*, 15 May 2000, p. 3.
20 The authors argue that MUC mixes up coreferential and anaphoric relations. For more on the difference between anaphora and coreference see Chapter 1, section 1.2.
21 Van Deemter and Kibble propose alternative solutions in their paper.
22 DRAMA stands for Discourse Reference Annotation for Multiple Applications.
23 As in the previous example, the ID numbers are chosen as an illustration and may not correspond to the actual enumeration in a real text.
24 Exhibited by anaphors whose antecedents are clauses or sentences; sentential anaphora itself may involve coreference or substitution.
25 The scheme does not cover lexical noun phrase anaphora.
26 Consider the case of demonstrative anaphora – it is well known that when the antecedent is a text segment longer than a sentence, it is often difficult to decide exactly which text portion represents the antecedent.
27 See McEnery (2002) for a detailed general discussion on various issues related to corpus annotation including consistency and accuracy.
28 By the Research Group in Computational Linguistics (http://www.wlv.ac.uk/sles/compling/).
29 Since the task of anaphora resolution is considered successful if any element of the anaphoric (coreferential) chain preceding the anaphor is identified, the project addressed the annotation of whole anaphoric (coreferential) chains and not only pairs of anaphors and their closest antecedents.
30 Another measure for computing the agreement between annotators used in the literature is the kappa statistic (Carletta, 1996). This measure only considers those items marked by both annotators and indicates how many times the annotators used the same tags and the same values for their attributes. The kappa statistic (κ) is computed as $\kappa = (P(A) - P(E))/(1 - P(E))$ where $P(A)$ is the proportion of times the annotators agree and $P(E)$ is the proportion of times that we would expect the annotators to agree by chance. It has been successfully employed for computing the feature-value agreement between annotators in several annotation projects (Davies et al. 1998; Vieira and Poesio 2000b). However useful this measure seems, it is not straightforward to compute it with respect to coreference annotation. This is because when the kappa statistic is computed, it is assumed that the possible values of features are known a priori. In the case of the annotation adopted, this would mean that the initial mentions of the chains are known. Different models were tried in order to find a solution to this problem, but none was found useful.
31 Given the limited scope of the study, the authors suggest that these results need more extensive evaluation.

CHAPTER SEVEN

An approach in focus: Mitkov's robust, knowledge-poor algorithm

The development of my robust, knowledge-poor approach[1] was motivated by the pressing need for anaphora resolution algorithms operating robustly in real-world, knowledge-poorer environments in order to meet the demands of practical NLP systems. I reported the first version of the algorithm in Mitkov (1996) as an inexpensive, fast and yet reliable alternative to the labour-intensive and time-consuming construction of a knowledge-based system.[2] This project was also an example of how anaphora can be resolved quite successfully (at least in a specific genre) without any sophisticated linguistic knowledge or even without parsing. In addition, the evaluation showed that the basic set of factors[3] employed can work well not only for English, but also for other languages.

In this chapter I shall describe my robust, knowledge-poor algorithm for pronoun resolution. In section 7.1 I shall introduce the original algorithm, discuss the antecedent indicators which form the basis of its resolution strategy and report evaluation results. Section 7.2 will focus on the multilingual character of the approach and will outline extensions to other languages. In section 7.3 I shall present a combined English–French version which mutually enhances the performance in both languages, benefiting from bilingual corpora and differences in gender or number agreement. Section 7.4 will introduce the recent fully automatic implementation MARS,[4] discuss the differences and improvements to the original model and report on its evaluation. Finally, in section 7.5 I shall outline the fully automatic version of my approach for Bulgarian.

7.1 The original approach

Mitkov's approach avoids complex syntactic, semantic and discourse analysis, relying instead on a list of preferences known as *antecedent indicators*. The approach operates as follows: it works from the output of a text processed by a part-of-speech tagger and an NP extractor, locates noun phrases which precede the anaphor within a distance of two sentences, checks them for gender and number agreement with the anaphor and then applies the indicators to the remaining candidates by assigning a positive or negative score (2, 1, 0 or –1). The noun phrase[5] with the highest composite score is proposed as antecedent.

7.1.1 Pre-processing strategy

The pre-processing strategy of the approach is simple: it uses a sentence splitter, a POS tagger and NP grammar rules to extract the preceding noun phrases in the current and two preceding sentences. Subsequent versions of the approach have used search scopes of different lengths (2, 3 or 4 sentences), but the original algorithm only considered two sentences prior to the sentence containing the anaphor. The NP patterns are limited to the identification of base NPs and do not include complex or embedded phrases.

7.1.2 Resolution strategy: the antecedent indicators

The detected noun phrases are passed on to a gender and number agreement test. The approach takes into consideration the fact that in English there are certain collective nouns which do not agree in number with their antecedents (e.g. *government*, *team*, *parliament*, etc., can be referred to by *they*; equally some plural nouns such as *data* can be referred to by *it*) and are thus exempted from the agreement test.

Next, the antecedent indicators are applied to all NPs which have passed the gender and number filter. The antecedent indicators can act in either a *boosting* or an *impeding* capacity. The boosting indicators apply a positive score to an NP, reflecting a positive likelihood that it is the antecedent of the current pronoun. In contrast, the impeding ones apply a negative score to an NP, reflecting a lack of confidence that it is the antecedent of the current pronoun. Most of the indicators are genre-independent and related to coherence phenomena (such as salience and distance) or to structural matches, whereas others are genre-specific.[6] In the following, the indicators employed by the pronoun resolution algorithm are outlined and illustrated by examples. The *boosting indicators* are:

- *First noun phrases*: A score of +1 is assigned to the first NP in a sentence.

For sentences containing subjects this preference is theoretically supported by the fact that, in the absence of a parse tree, it acts as a linear approximation of the subject preference as used in centering. From the viewpoint of another theory, in a coherent text the given or known information, or theme, usually appears first, and thus forms a coreferential link with the preceding text (Firbas 1992). The new information, or rheme, provides some information about the theme.

- *Indicating verbs*: A score of +1 is assigned to those NPs immediately following a verb which is a member of a predefined set (including verbs such as *analyse, assess, check, consider, cover, define, describe, develop, discuss, examine, explore, highlight, identify, illustrate, investigate, outline, present, report, review, show, study, summarise, survey, synthesise*, etc.). Empirical evidence suggests that noun phrases following the above verbs usually carry more salience.

- *Lexical reiteration*: A score of +2 is assigned to those NPs repeated twice or more in the paragraph in which the pronoun appears and a score of +1 is assigned to those NPs repeated once in that paragraph.

Lexically reiterated items are identified on the basis of simple string matching but lexical reiteration also includes sequences of noun phrases with the same head (e.g. *a bottle*, *the bottle* or *toner bottle*, *bottle of toner*, *the bottle*). In addition, a list of mutually excluding modifiers such as *first* and *second* is used to track down noun phrases which have the same head but are not coreferential (e.g. *the first channel* and *the second channel* do not count as lexical reiteration). Due to the fact that the approach does not use any ontology such as WordNet, synonyms or superordinates cannot be captured and thus such occurrences are not counted as lexical reiterations.

- *Section heading preference*: A score of +1 is assigned to those NPs that also occur in the heading of the section in which the pronoun appears. This score is awarded in addition to the score of +1 obtained through *lexical reiteration* due to the repetition of a specific NP in a following passage.
- *Collocation match*: A score of +2 is assigned to those NPs that have an identical collocation pattern to the pronoun.

Collocation match is restricted to the patterns <noun phrase/pronoun, verb> and <verb, noun phrase/pronoun> or if the verb is *to be*, to <noun phrase/pronoun, verb, adjective/past participle>. Example:

(7.1) Press *the key* down and turn the volume up . . . Press *it* again.

Owing to lack of syntactic information, this preference is somewhat weaker than the collocation preference described in Dagan and Itai (1990).

The collocation match preference has been extended to patterns <(un)V, NP/pronoun> and <NP/pronoun, (un)V>, i.e. verbs with an 'undoing action' meaning are considered to fall into collocation patterns along with their 'doing action' counterparts. This extended new rule helps in cases such as 'Loading *a cassette* or unloading *it*'. Also, certain patterns are still considered collocation matches if the verb takes the form of a gerund as in the case of 'When you plug in the power adapter, *the print head* moves to its protected position (you'll hear *it* moving)'.[7]

While this approach relies on collocation information from the active document and the current paragraph, an 'extension' of this indicator is envisaged to consider three types of collocation match: (i) a corpus-based collocation match – this involves extracting a collocation dictionary from an available corpus and updating it each time a new document/subcorpus is added, (ii) a document-based collocation match – this collocation information is based on the active document only and (iii) a paragraph-based collocation match – this is based on the patterns in the current paragraph. I also plan to fine-tune this preference by assigning appropriate scores to each of the three types of collocation. Preliminary observations suggest that the paragraph-based collocation is a stronger preference than document-based collocation which, in turn, is a stronger preference than corpus-based collocation.

- *Immediate reference*: A score of +2 is assigned to those NPs appearing in constructions of the form '. . . (You) V_1 NP . . . *con* (you) V_2 it (con (you) V_3 it)', where *con* ∈ {and/or/before/after/until . . . }.

This preference can be viewed as a modification of the collocation preference. It is highly genre-specific and occurs frequently in imperative constructions:

(7.2) To print the paper, you can stand *the printer* up or lay *it* flat.
(7.3) Gently push the diskette into the drive until it clicks into place.
(7.4) Unwrap *the paper*, form *it* and align *it*, then load *it* into the drawer.

This indicator, *prepositional noun phrases* and *collocation pattern preference* have proved to be the most 'confident' indicators in pointing to the correct antecedent. In fact the noun phrase awarded a score by *immediate reference* always emerges as the correct antecedent. Chapter 8 discusses the related measure of *decision power* and its values obtained for each indicator.

- *Sequential instructions*: A score of +2 is applied to NPs in the NP_1 position of constructions of the form: 'To V_1 NP_1, V_2 NP_2. (Sentence). To V_3 it, V_4 NP_4' where the noun phrase NP_1 is the likely antecedent of the anaphor *it* (NP_1 is assigned a score of 2).

  (7.5) To turn on *the video recorder*, press the red button. To programme *it*, press the 'Programme' key.

- *Term preference*: A score of +1 is applied to those NPs identified as representing terms in the genre of the text.

As with immediate reference, the last two indicators are genre-specific.

The *impeding indicators* are:

- *Indefiniteness*: Indefinite NPs are assigned a score of –1.

Indefinite noun phrases in previous sentences are very often less likely antecedents of pronominal anaphors than definite ones and are penalised by –1. The program regards a noun phrase as definite if the head noun is modified by a definite article, or by demonstrative or possessive pronouns.

- *Prepositional noun phrases*: NPs appearing in prepositional phrases are assigned a score of –1.

  (7.6) Insert *the cassette* into the VCR making sure *it* is suitable for the length of recording.

In (7.6) the noun phrase *the VCR* is penalised for being part of the prepositional phrase *into the VCR*. This preference can be explained in terms of salience from the point of view of the centering theory. The latter proposes the ranking 'subject, direct object, indirect object' (Brennan et al. 1987) and noun phrases which are parts of prepositional phrases are often indirect objects.

One indicator, *referential distance*, may impede or boost a candidate's chances of being selected as the antecedent of a pronoun depending on that NP's distance in terms of clause and sentence boundaries from the pronoun. NPs in the previous clause to (but in the same sentence as) the pronoun are assigned a score of +2, those in the previous sentence to the pronoun are assigned a score of +1, those in the sentence prior to that are assigned a score of 0 and more distant pronouns are assigned a score of –1.[8]

It should be pointed out that the antecedent indicators are preferences and not absolute factors. There might be cases where one or more of the antecedent indicators do not 'point' to the correct antecedent. For instance, in the sentence 'Insert the cassette into *the VCR* making sure *it* is turned on', the indicator *prepositional noun phrases* would penalise the correct antecedent. When all preferences (antecedent indicators) are taken into account, however, the right antecedent is still likely to be tracked down – in the above example, the *prepositional noun phrases* heuristic stands a good chance of being overturned by the *collocation match* heuristics since the collocation *The VCR is turned on* is likely to appear previously in the text, being typical of video technical manuals.

The antecedent indicators have proved to be reasonably efficient in identifying the right antecedent and the results show that for the genre of technical manuals they may be no less accurate than syntax- and centering-based methods (see Mitkov 1998b). The approach described is not dependent on any theories or assumptions; in particular, it does not operate on the assumption that the subject of the previous utterance is the highest-ranking candidate for the backward-looking center – an approach which can sometimes lead to incorrect results. For instance, subject-favouring methods or methods heavily relying on syntactic parallelism would incorrectly propose *the utility* as the antecedent of *it* in the sentence 'The utility shows you *the LIST file* on your terminal for a format similar to that in which *it* will be printed'[9] as it would prefer the subject as the most salient candidate. The *indicating verbs* preference of Mitkov's approach, however, would prefer the correct antecedent *the LIST file*.

7.1.3 Informal description of the algorithm

The algorithm for pronoun resolution can be described informally as follows:

1. Examine the current sentence and the two preceding sentences (if available). Look for noun phrases[10] only to the left of the anaphor.[11]
2. Select from the identified noun phrases only those which agree in gender and number[12] with the pronominal anaphor and group them as a set of potential candidates.
3. Apply the antecedent indicators to each potential candidate and assign scores; propose the candidate with the highest aggregate. If two candidates have an equal score, propose the candidate with the higher score for immediate reference. If immediate reference does not hold, propose the candidate with higher score for collocational pattern. If collocational pattern suggests a tie or does not hold, select the candidate with higher score for indicating verbs. If this indicator does not hold again, go for the most recent candidate.

7.1.4 Illustration

Consider the following example featuring the anaphor *it*.

(7.7) Positioning the original: Standard Sheet Original

> Raise the original cover. Place the original face down on the original glass so that *it* is centrally aligned against the original width scale. The center of the original must be aligned with the arrow marking on the original width scale.

Steps 1 and 2 of the described algorithm generate the set of potential candidates as {*original cover*, *original*, *original glass*}.

Step 3 assigns the following scores to the candidates:

original cover

1 (first noun phrases) + 0 (indicating verbs) + 0 (lexical reiteration) + 0 (section heading) + 0 (collocation) + 0 (immediate reference) + 0 (sequential instructions) + 1 (term preference) + 0 (indefiniteness) + 0 (prepositional noun phrases) + 1 (referential distance) = 3

original

1 (first noun phrases) + 0 (indicating verbs) + 1 (lexical reiteration) + 1 (section heading) + 0 (collocation) + 0 (immediate reference) + 0 (sequential instructions) + 1 (term preference) + 0 (indefiniteness) + 0 (prepositional noun phrases) + 2 (referential distance) = 6

original glass

0 (first noun phrases) + 0 (indicating verbs) + 0 (lexical reiteration) + 0 (section heading) + 0 (collocation) + 0 (immediate reference) + 0 (sequential instructions) + 1 (term preference) + 0 (indefiniteness) + (–1) (prepositional noun phrases) + 2 (referential distance) = 2

The noun phrase *the original* (score 6) is selected as antecedent for *it*.

7.1.5 Evaluation of Mitkov's original approach

For practical reasons, the approach presented does not incorporate syntactic and semantic knowledge (other than a list of domain terms) and this would suggest that the results are not likely to be as successful as those achieved through an approach making use of syntactic knowledge in the form of constraints and/or preferences. The lack of syntactic information, for instance, means giving up c-command constraints and subject preference (or on other occasions object preference: see Mitkov 1995b) which could be used in center tracking. Syntactic parallelism, useful in selecting the antecedent on the basis of syntactic function, also has to be forgone. Lack of semantic knowledge rules out the use of verb semantics and semantic parallelism. However, despite these limitations, the evaluation suggests that results are comparable to syntax-based methods (e.g. Lappin and Leass 1994). The evaluation results also show superiority over other knowledge-poor methods in the specific genre of user manuals. I believe that the good success rate is due to the fact that a number of efficient antecedent indicators are taken into account and no factor is given absolute preference. In

Table 7.1 Success rates of the knowledge-poor approach on different manuals

Manual	Number of anaphoric pronouns	Success rate in %
Minolta Photocopier	48	95.8
Portable StyleWriter (PSW)	54	83.8
Alba Twin Speed Recorder	13	100.0
Seagate Medallist Hard Drive	18	77.8
Haynes Car Manual	50	80.0
Sony Video Recorder	40	90.6
All manuals	223	89.7

particular, this strategy can often override incorrect decisions caused by strong centering preference (see the end of section 7.1.2) or syntactic and semantic parallelism preferences (Mitkov 1998b).

The knowledge-poor approach was evaluated by obtaining the *success rate*[13] on the basis of texts which were automatically pre-processed (POS tagging, NP identification) but were then manually post-edited to ensure that the input to the algorithm was correct. The evaluation in English[14] included texts from different technical manuals (Minolta Photocopier, Portable StyleWriter (PSW), Alba Twin Speed Video Recorder, Seagate Medalist Hard Drive, Haynes Car Manual, Sony Video Recorder) which contained a total of 223 anaphoric pronouns. The knowledge-poor approach resolved 200 anaphors correctly which gave a success rate of 89.7%. The success rates were different for each of the technical manuals (Table 7.1) which showed that even for texts belonging to the same genre, results may vary. Therefore, this means that for more definitive figures test data containing thousands of anaphors were needed.[15]

The *critical success rate* of the approach was measured as 82% on part of the evaluation data (Portable StyleWriter manual – see Table 7.3). This measure covers only the anaphors which still have more than one candidate for antecedent after gender and number filters and which, therefore, are more difficult to resolve (see Chapter 8). The high critical success rate[16] (almost as high as the success rate on these texts) and the significantly lower figures for the baseline models (see below) speak in favour of the efficiency of the antecedent indicators.

In order to evaluate the effectiveness of the approach and to explore if and to what extent it is superior to the *baseline models* for anaphora resolution, I tested the sample texts on (i) a baseline model which checks agreement in number and gender and, where more than one candidate remains, picks out as antecedent the most recent subject matching the gender and number of the anaphor and (ii) a baseline model which selects as antecedent the most recent noun phrase that matches the gender and number of the anaphor.[17] The evaluation results suggest success rates of 48.6% for the first baseline model and 65.9% for the second (Table 7.2).

Table 7.2 Comparison of the success rates of Mitkov's knowledge-poor approach with two baseline models

Approach	Number of anaphoric pronouns	Success rate in %
Knowledge-poor approach	223	89.7
Baseline Most Recent	223	65.9
Baseline Subject	223	48.6

Typically, the knowledge-poor approach proved superior to both baseline models when the antecedent was neither the most recent subject nor the most recent noun phrase matching the anaphor in gender and number. Consider the following example:

(7.8) Identify *the drawer* by the lit paper port LED and add paper to *it*.

The composite score for *the drawer*[18] is 5, whereas the composite score for the most recent matching noun phrase *the lit paper port LED*[19] is 2. Therefore, while the most recent NP model fails in this case, Mitkov's approach suggests the correct antecedent with a comfortable margin. From example (7.8) it can also be seen that the knowledge-poor approach successfully tackles cases in which the anaphor and the antecedent have not only different syntactic functions but also different semantic roles. Usually knowledge-based approaches encounter difficulties in such situations because they use preferences such as 'syntactic parallelism' or 'semantic parallelism'. Mitkov's approach does not use these because it has no information about the syntactic structure of the sentence or about the syntactic function/semantic role of each individual word.

As far as the typical cases of failure are concerned, the knowledge-poor approach has difficulty dealing with more complex syntactic structures. This should not be surprising, given that the approach does not rely on any syntactic knowledge and, in particular, it does not produce any parse tree.

The evaluation also included *comparison to similar approaches* such as Breck Baldwin's CogNIAC approach (Baldwin 1997) which was run on part of the evaluation texts. CogNIAC successfully resolved the pronouns in 75% of the cases. This performance is compatible with the results described by Baldwin (1997). The reason for choosing CogNIAC is that both Mitkov's and Baldwin's approaches share common principles – both are regarded as knowledge-poor and use POS taggers rather than parsers.[20]

In addition, a *comparison with well-established approaches* included a small-scale evaluation of Jerry Hobbs's naïve algorithm (Hobbs 1976) on the basis of the same texts used for the comparative evaluation of Baldwin's approach (StyleWriter 1994). The results obtained point to a success rate of approximately 71%.

The results in Table 7.3[21] show that on this small set of data from the genre of technical manuals, the knowledge-poor approach performs better than Baldwin's or Hobbs's approaches. These results, however, cannot be generalised

Table 7.3 Comparative evaluation and critical success rate based on the PSW corpus

Approach	Number of anaphoric pronouns	Success rate in %	Critical success rate
Knowledge-poor approach PSW	54	83.8	82
Baldwin's CogNIAC	54	75	–
Hobbs's naïve algorithm	54	71	–

for other genres or unrestricted texts and further extensive tests are necessary in order to obtain an accurate picture.[22]

Finally, it is worth pointing out that even though the knowledge-poor approach can be regarded as genre-specific, it appears that it can do well in other genres as well. An evaluation on a small dataset in the genre of research papers achieved a success rate of 77.9%.

7.2 The multilingual nature of Mitkov's approach: extensions to other languages

Mitkov's approach was initially developed and tested for English and was later adapted and tested for Polish and Arabic as well.[23] For both languages, it was found that adaptation required minimum modification and that even if used unmodified, the approach delivered very good success rates.

7.2.1 Agreement and antecedent indicators for Polish and Arabic

Agreement rules played a more prominent role in both Polish and Arabic and, as expected, they were able to filter out more candidates for antecedent. The lower number of remaining candidates in Polish and Arabic is a possible explanation for the higher success rates of the approach for these languages (Table 7.4). The gender and number agreement of an anaphor and its antecedent in Polish is compulsory. Polish gender distinctions are much more diverse than in English (e.g. feminine and masculine do not apply to a restricted number of nouns only). Moreover, one pronominal form can potentially refer to nouns of different gender. For instance, the singular genitive third person form *jego* can equally refer to either masculine or neuter nouns. In addition, certain pronouns such as the accusative form *je* can refer to either singular neuter or plural feminine nouns. Finally, unlike English, zero anaphors (in subject position) are typical of Polish declarative sentences.

Agreement rules in Arabic are different. For instance, a set of non-human items (animals, plants, objects) is referred to by a singular feminine pronoun. Since Arabic is an agglutinative language, the pronouns may appear as suffixes of verbs, nouns (e.g. in the case of possessive pronouns) and prepositions. In particular, in the genre of technical manuals there are five 'agglutinative' pronouns.

The pronoun *ho* is used to refer to singular masculine persons and objects, while *ha* refers to singular feminine ones. There are three plural anaphoric pronouns: *homa* which refers to a dual number (a set of two elements) of both masculine and feminine nouns, *hom* which refers to a plural number (a set of more than two elements) of masculine nouns and *honna* which refers to a plural number of feminine nouns.

The antecedent indicators employed for the Polish and Arabic versions of Mitkov's approach were the same as those used for English with the exception of one indicator. In most cases the indicators were applied with the same score as English which suggests that the indicators can be regarded as 'multilingual'. One additional indicator used for Arabic was the *relative pronoun indicator* based on the fact that the first anaphor following a relative pronoun refers exclusively to the most recent NP preceding it; this NP is considered as the most likely antecedent and is awarded a score of 1. The following example illustrates the importance of this indicator.

(7.9) Al-tahakkok min tahyiat al-moakkit
Checking the Timer Settings

Yomkino-ka a'rdh tahyiat moakkitoka li-at-tahakkok mina *al-baramij* al-lati targhabo fi tasjili-*ha*.
You can display your timer settings to confirm *the programmes* that you wish to recording *it*.

In this example the pronoun *ha* (it) is the first pronominal anaphor which follows the relative pronoun *al-lati* (that) and refers to the non-animate feminine plural *al-baramij* (the programmes; for agreement rules in Arabic see above) which is the most recent NP preceding *al-lati*.

With a view to applying the indicator *indefiniteness*, definiteness was identified in different ways for Polish and for Arabic. Since in Polish there are no definite articles, definiteness is signalled by word order, demonstrative pronouns or repetition. In Arabic, definiteness occurs in a richer variety of forms (Galaini 1992). In addition to the definiteness triggered by the definite article *al* (the), demonstrative and possessive pronouns, a noun phrase in Arabic is also regarded as definite if it is followed by a definite noun/noun phrase.[24] For example, the noun phrase *kitabu al-rajuli* (lit. 'book the man'), which means 'the book of the man', is considered definite since the non-definite noun *kitabu* (book) is followed by the definite noun *al-rajuli* (the man). In Arabic this form of definiteness is called 'Al-ta'rif bi-al-idhafa' (definiteness by addition).

The antecedent indicator *prepositional noun phrases* had to be adapted for both Polish and Arabic. The indicator was extended in Polish to frequently occurring genitive constructions such as *liczba komputerow* (number of computers). Nouns which are part of such genitive constructions and which are not in genitive form are penalised by –1. In Arabic the antecedent and the anaphor can belong to the same prepositional phrase (see section 7.2.3). Therefore, this indicator was modified for the 'Arabic version' accordingly: if an NP belongs to a prepositional

phrase which does not contain the anaphor, it is penalised by –1; otherwise it is not assigned any score.

Finally, the *referential distance* indicator was modified for Arabic. Since it was discovered that in Arabic the anaphor is more likely to refer to the most recent NP, the scoring system for Arabic gives a bonus to such candidates: the most recent NP is assigned a score of 2, the one that precedes it immediately 1, and the rest 0.

7.2.2 *Evaluation of the Polish version*

Given that manually parsed corpora were used in both cases, the evaluation of both the Polish and the Arabic versions focused on the performance of the algorithm. This was the only option because processing tools were not available for either of the languages.

The evaluation for Polish was based on technical manuals available on the Internet (Internet Manual 1994 and Java Manual 1998). The sample texts contained 180 pronouns among which were 120 instances of exophoric reference (most being zero pronouns). The robust approach adapted for Polish demonstrated a high success rate of 93.3% in resolving anaphors.

Similarly to the evaluation for English, the approach for Polish (Mitkov and Stys 1997) was compared with (i) a baseline model which discounts candidates on the basis of non-agreement in gender and number and, from the remaining candidates, selects as the antecedent the most recent subject matching the anaphor in gender and number and (ii) a baseline model which checks for agreement in gender and number and, from the remaining candidates, selects as the antecedent the most recent noun phrase that agrees with the anaphor.

The Polish version of the robust knowledge-poor approach showed clear superiority over both baseline models for Polish. The first baseline model (baseline subject) was successful in only 23.7% of the cases, whereas the second (baseline most recent) had a success rate of 68.4%. These results demonstrate the increase in performance due to the use of antecedent tracking indicators.

The most typical instances where the preference-based approach performed better than the baseline models were (as expected) cases in which neither the subject nor the most recent noun phrase were successful candidates for antecedent. This is illustrated by the following example:

(7.10) Opowiemy wam o typach dostepnych zasobow informacji
We-will-tell you about types accessible resources information

oraz o tym jak wyprobowac mozliwosci niektorych z nich.
and about how try-out possibilities some of them

The antecedent of *nich* (them) is *typach dostepnych zasobow informacji* (types of accessible resources of information) with a composite score of 5. The antecedent

in this case is neither the subject (which is an instance of exophoric zero anaphora not having its source in the text) nor the most recent noun *mozliwosci* (possibilities) which nevertheless agrees in number and gender with the pronoun *nich*.

The Polish version also showed a very high critical success rate of 86.2%. When used without any modification ('Polish direct'), the approach scored a 90% success rate (see Table 7.4).

7.2.3 Evaluation of the Arabic version

Mitkov's approach for Arabic (Mitkov et al. 1998) was evaluated operating in two modes: the first mode consisted of using the robust approach directly, without any adaptation/modification for Arabic, whereas the second mode used an adapted version which included modified antecedent indicators (see above) designed to capture some of the specific aspects of Arabic plus a new *relative pronoun indicator*.

The evaluation was based on 190 anaphors from a technical manual (Sony 1992). The success rate of the robust approach directly employed without any adaptation for Arabic (this version is referred to as 'Arabic direct' in Table 7.4) was found to be 77.9% (148 out of 190 anaphors were correctly resolved). Typical failures were examples in which the antecedent and the anaphor belonged to the same prepositional phrase:

(7.11) Tathhar al-surah fi awal *kanat* ta-stakbilo-*ha* fi mintakati-ka.
Appears the-picture on first *channel* you-receive-*it* in area-your.

Such failure cases did not occur in the adapted (improved) version for Arabic in which the 'non-prepositional phrase' rule was changed (see section 7.2.1).

Another typical problem rectified by changing the referential distance in Arabic was the case in which the anaphor appeared as part of a PP modifying the antecedent-NP:

(7.12) Kom bi-taghtiat thokb al-lisan bi-sharit plastic aw ista'mil kasit *akhar* bi-*hi* lisan al-aman.
Cover slot the-tab with-tape plastic or use *cassette another* in *it* tab the-safety.

The candidates for antecedent in this example are the noun phrases *safety tab slot*, *plastic tape* and *another cassette*. If the robust approach is used without any modification, each candidate scores 2 for referential distance; the composite score for *tab slot* is 3, for *plastic tape* it is 2 and for *another cassette* it is 2 (all of these candidates get an additional 1 score for term preference). Using the new referential distance scores, however, the correct candidate *another cassette* is assigned a total of 2 as opposed to the other two candidates which are each assigned a composite score of 1.

The evaluation of the adapted and improved version for Arabic (referred to as 'Arabic improved' in Table 7.4) reported a success rate of 95.8% (182 out of 190 anaphors were correctly resolved).

Table 7.4 Summary of the evaluation results

	Success rate	Critical success rate	Baseline most recent	Baseline subject
English	89.7%	82%	65.9%	48.6%
Polish direct	90%	–	–	–
Polish improved	93.3%	86.2%	68.4%	23.7%
Arabic direct	77.9%	70.4%	–	–
Arabic improved	95.8%	94.4%	–	–

Both evaluations for Arabic also showed a very high critical success rate. The robust approach used without any modification scored a critical success rate of 70.4%, whereas the improved Arabic version scored 94.4%.

Table 7.4 summarises the success rates and critical success rates obtained for the English, Polish and Arabic versions[25] of the knowledge-poor approach and provides a comparison with the baseline models of English and Polish.

7.2.4 Extension to French

Mitkov's approach was recently implemented for French as part of a bilingual project based on the so-called 'mutual enhancement' methodology (see next section).[26] The French version closely followed the original approach for English, the only modification being the re-formulation of the *immediate reference* indicator:

- *Immediate reference* (French): A score of +2 is assigned to those NPs appearing in constructions of the form 'V_1 NP (con) V_2 le/la/les' or 'V_1 NP (con) le/la/les/l' V_2', where *con* is any French conjunction.

This modification was necessary because in French the position of the third person pronoun (le/la/les/l') depends on the type of clause (imperative or not), so both constructions had to be captured. As an illustration, consider examples (7.13) and (7.14) with their translations in English:

(7.13) Pour forcer la fin du listage, éteindrez *l'imprimante* et puis allumez-*la* du nouveau.
To force cancelling a print job, turn the printer off and then turn it on again.

(7.14) Pour forcer la fin du listage, il faut éteindre *l'imprimante* et puis *l'*allumer du nouveau.
To force cancelling a print job, one should turn the printer off and then turn it on again.

The performance of the French version on the evaluation data used for the bilingual resolution project is outlined in section 7.3.5.

7.3 Mutually enhancing the performance for English and French: a bilingual English/French system

7.3.1 Rationale

The English and French versions of Mitkov's approach served as a basis for the development of a bilingual pronoun resolution system which seeks to exploit the output of the French module in order to improve the performance of the English one and vice-versa. The rationale for the development of the bilingual anaphora resolution system was the fact that while there is no gender discrimination in English, the gender distinction in French could be helpful in the resolution process of English pronouns. As an illustration, consider the following example:

(7.15a) John removes the cassette from the videoplayer and disconnects it.

Without information on co-occurrence patterns which may not be widely available or without subcategorisation knowledge (selectional restrictions) which is even more difficult to obtain, the majority of anaphora resolution approaches would select as antecedent the wrong candidate *the cassette* instead of the correct one *the videoplayer* because indirect objects and noun phrases which are contained in adverbial prepositional phrases are usually penalised (Lappin and Leass 1994; Mitkov 1998b). Similarly, centering theory regards direct objects as more salient than indirect objects (Walker et al. 1998).

On the other hand, an anaphora resolution system for French would not have problems processing the equivalent French example:

(7.15b) Jean éjecte la cassette du magnétoscope et le débranche.

and identifying *(le) magnétoscope* as the correct antecedent of the pronoun *le* since the other candidate *la cassette* does not match the pronoun in gender.

These and other similar examples where the gender distinction in French could be helpful motivated the development of a bilingual (English/French) pronoun resolution system which features a strategy of mutual enhancement of performance and operates on parallel English and French corpora aligned at word level. In addition to utilising gender discrimination in French, this strategy also benefits from a bilingual corpus (e.g. information on how a pronoun is translated in the target language) and from the performance of the English algorithm (e.g. the antecedent indicators for English usually perform more accurately). The English and French modules mutually enhance their performance in that their outputs are compared and, if they disagree, one of them is preferred depending on the case (see section 7.3.3). Both the English and the French modules are based on Mitkov's knowledge-poor approach.

7.3.2 Brief outline of the bilingual corpora

Parallel bilingual English–French corpora are produced in most cases either on the basis of translating an original English text into French or on the basis of

translating original French text into English. Normally translation is performed with a view to achieving maximal fluency and cohesion in the target language where the distribution of words may be different from the source language. The translation of technical texts is generally not as free as the translation of literary works but nevertheless is highly unlikely to be literal. In fact, it is not unusual to have the target text rewritten for reasons of clarity.

Three technical texts (Linux HOW TO documents) were used in this bilingual experiment: 'Beowulf HOW TO v.1.1.1' (referred to in the tables as BEO), 'Linux CD-Rom HOW TO v.1.14' (CDR) and 'Access HOW TO v.2.11' (ACC), containing about 30 000 words in each language. Table 7.5 shows the exact number of words in each language as well as the number of pronouns (third person pronouns, possessives and reflexives were considered). The original files were in English and translated into French. Some of the pronouns occurring in English were completely omitted in French, replaced by full noun phrases or replaced by other types of anaphors whose resolution was not tackled in the project (for example, demonstratives). Similarly, some English noun phrases were replaced by pronouns in the French translation, whereas a few additional French pronouns were introduced even though they did not have a corresponding pronoun in the English text. Table 7.6 presents a summary of the different ways in which English pronouns were translated into French and the cases giving rise to new French pronouns.

Table 7.5 Distribution of pronouns in the bilingual corpus

	Words		Pronouns	
File	English	French	English	French
ACC	9 617	10 168	130	156
CDR	9 490	11 028	83	136
BEO	6 392	6 841	68	98
Total	25 499	28 037	281	390

Table 7.6 Pronoun translation correspondences

File	Direct translations of pronouns	English pronoun to French NP	English NP to French pronoun	New French pronouns	English pronoun omitted
ACC	108	12	27	31	10
CDR	77	5	22	37	1
BEO	56	7	19	23	5
Total	241	24	68	91	16

The mutual enhancement strategy benefits from the differences in the translation of pronouns and in particular from cases where a pronoun has been translated as a noun phrase which is a translation equivalent of its antecedent.

7.3.3 The contributions of English and French

The strategy of mutual enhancement is based on the English and French modules' benefiting from each other, and therefore mutually enhancing their performance. In fact, there are certain cases where the French module is expected to perform more reliably, whereas in others the English module is likely to propose the correct antecedent with higher probability.

7.3.3.1 CASES WHERE FRENCH/THE FRENCH VERSION HELPS

The most obvious benefit of using a French anaphora resolver is to exploit the *gender discrimination in French*. Gender agreement between the pronominal anaphor and its antecedent holds in most of the cases in French. The exceptions refer to special cases like noun phrases representing professions or positions.[27] When a pronoun is used to refer to a person occupying a specific position, its gender does not match the grammatical gender of the noun phrase, but that of the person.

(7.16) *Le professeur* se mit en colère. *Elle* n'en pouvait plus.
The teacher got cross. She could not stand it any more.

On the other hand, when certain professions are used generically and are referred to by a pronoun, the latter will take the gender of the profession, not of the person involved

(7.17) Quand *un professeur* se met en colère, *il* perd son autorité sur ses élèves.[28]
When a teacher gets cross, he loses his authority over his students.

Since gender agreement works for most cases in French, whenever the antecedent in French is resolved directly after gender agreement, its equivalent[29] in English is adopted as the antecedent.

It has to be borne in mind, however, that not all the pronouns in French point to the gender of the noun phrase they refer to. The ones that convey gender information are third person plural and singular personal pronouns used in subject position (*il*, *elle*, *ils*, *elles*), reflexive pronouns in the singular (*elle-même*, *lui-même*) and singular personal pronouns in the accusative (*le*, *la*). Plural personal pronouns in the accusative and dative do not carry any kind of gender information (*les*, *eux*), whereas possessive pronouns only convey information about the noun phrase they modify, and therefore do not contribute to this methodology.

Another straightforward case where the French system will boost the performance of the English is when the *translations of the English pronouns are French noun phrases* which are identical to or coreferential with the antecedent. In that case, the equivalent of the French antecedent is taken to be the antecedent in

English. Since the system runs on aligned corpora which are not annotated for coreferential chains, this case is exploited by considering as antecedent an NP which has the same head as the translation of the English pronoun within the window of the search scope (two preceding sentences).

Finally, when *the highest-ranked French candidate is well ahead of its English 'competitor'* (with a numerical value of 4 adopted as the threshold), the French antecedent and its English equivalent are taken as antecedents. As an illustration, if the difference between the scores of the highest-ranked candidate and the second best in French is at least 5, and the difference between the two best English candidates is only 1, then the proposed antecedent of the French module will be preferred.

A small-scale study into the usability of the enhancement strategy, based on a small test corpus of 231 English and 255 French pronouns, showed that the resolution of up to 48% of English pronouns could be improved on the basis of the French gender discrimination and the translation of some of them as noun phrases. As for the French pronouns, gender agreement could contribute to the successful resolution of up to 65.4% of them.

7.3.3.2 CASES WHERE THE ENGLISH VERSION CAN HELP

Currently the algorithm for English is more developed than the one for French, and its success rate is normally higher. This is the reason why in one of the decision strategies described in section 7.3.4 below, a composite score is taken with weight assigned to the English score 0.6 as opposed to 0.4 for French. Also, if after applying all decision strategies the tie between two competing English–French candidates is still not broken (see section 7.3.4), the antecedent proposed by the English module is preferred. Another reason for favouring the algorithm for English is that in the French implementation the indicators were employed with the same scores in English. A thorough investigation of the optimal scores for French has yet to be conducted.

There are a number of other, more concrete cases where the English module can be of help. Mitkov's algorithm implemented for this project incorporates the following *syntax filters* adopted from Kennedy and Boguraev (1996):

(i) A pronoun cannot refer with a co-argument.
(ii) A pronoun cannot corefer with a non-pronominal constituent which it both commands and precedes.
(iii) A pronoun cannot corefer with a constituent which contains it.

These constraints are a modification of the syntax constraints reported in Lappin and Leass (1994) and work quite well for intrasentential anaphors, but similar constraints have not been implemented for French. Therefore, if the bilingual system tries to resolve an intrasentential anaphor and if the proposed antecedents for English and French are not equivalent, the decision of the English module is preferred.

One of the last tie-breaking heuristics is the use of the value of the *decision power* (Mitkov 2001b) which is indicative of the confidence of the proposed

antecedent (see also Chapter 8, section 8.3.3). The decision power is a measure well studied in English, as opposed to French. Therefore, the value of the decision power for English is preferred in cases where the other decision strategies are incapable of proposing the correct antecedent. Another case where the English module could contribute to enhancing the performance of the French module is when the *translation of the French pronoun is an English noun phrase*, identical or coreferential with its antecedent. In that case, the antecedent of the French pronoun is selected as the French equivalent of the English antecedent.

Collective nouns in English such as *parliament*, *government*, *army*, *team*, etc., can be referred to both by a plural pronoun (*they*) and by a singular one (*it*). On the other hand, in French, such nouns are only referred to by singular pronouns (*il*, *elle*). Therefore, if the pronoun is *they*, if there are no other plural candidates in English and if the English antecedent is a collective noun, the decision for English can help the resolution in French where the anaphor may have to compete with other candidates of the same gender.

Finally, the English module is helpful in cases where *the highest-ranked English candidate is well ahead of its French competitor* with 3 taken again as a threshold (see above and also the following section).

7.3.4 Selection strategy

The selection strategy is based on favouring cases where one of the systems is expected to perform better, as described in section 7.3.3 above, and addresses pronouns that cannot be resolved directly[30] in either of the languages. This strategy benefits from the outputs of Mitkov's algorithms (both the original version for English and its adaptation for French, specially developed for this project) and can be presented as a sequence of eight steps:

- *Step 1* If one of the English pronouns is translated as an NP in French, and if that French NP is preceded by an NP with the same head within a window of two sentences, the English equivalent of the preceding NP is taken as the antecedent for English. The same applies in reverse order for French.
- *Step 2* If a French pronoun is resolved after applying the gender agreement constraint, the corresponding English pronoun is resolved to the English equivalent of the identified French antecedent.
- *Step 3* If there is only one plural pronoun in English and if it refers to a collective noun such as *parliament*, *army*, *police*, etc., and if the corresponding French pronoun has not yet been resolved, the antecedent for French is set to the equivalent of the English collective noun.
- *Step 4* If an English pronoun is resolved as a result of applying the intrasentential constraints described in 7.3.3, the equivalent of the English antecedent is taken as antecedent for French.
- *Step 5* If the top candidates are such that they are different for each language and if the difference between the highest-ranked candidate and the second best in one language is much greater than that between the highest-ranked candidate and the second best in the other language (greater than or equal to

3 for English and 4 for French[31]), the highest-ranked candidate with greater score difference from its runner-up and its equivalent are taken as antecedents.

- *Step 6* If the top candidates for both languages are different and if the condition described in Step 5 does not apply, for each English candidate English_C_i ($i = 1, \ldots, N$; N is the number of all candidates) and its equivalent French candidate French_C_i ($i = 1, \ldots, N$), the weighted score $0.6 \times$ English_C_i + $0.4 \times$ French_C_i is computed. The pair of candidates English_C_k and French_C_k with the highest weighted score are declared as antecedents.
- *Step 7* In the event of a tie, the values of the decision power of the employed antecedent indicators are considered. If in one of the languages an indicator with a decision power >0.8 is employed and if the highest decision power of the indicators activated in the other language is <0.6, the proposed candidate in the first language and its equivalent in the second are declared as antecedents.
- *Step 8* If none of the Steps 1–7 can deliver an antecedent, the NP proposed by the English module and its French equivalent are chosen as antecedents.

7.3.5 *Evaluation*

The evaluation was based on parallel texts featuring 25 499 words (281 of which were pronouns) in English and 28 037 words (390 pronouns) in French (see Table 7.5). The evaluation files were annotated for morphological features and syntactic constituents and had tables, sequences of code, tables of contents, tables of references and translation notes removed.

The evaluation was performed in two passes. In the first pass the individual anaphora resolvers for English and French were run separately and their performance was computed in terms of success rate (number of correctly solved anaphors / number of all anaphors). In the second pass the mutual enhancing algorithm was activated, benefiting from the outputs of each individual resolver. The success rate of each resolver was computed again after enhancement and the improvement in performance recorded.

Tables 7.7 and 7.8 show that the improvement of the success rate on particular files could be up to 4.62% for English and up to 5.15% for French after

Table 7.7 Pronoun resolution for English before and after enhancement

File (English)	Number of pronouns	Before enhancement correctly resolved pronouns/success rate	After enhancement correctly resolved pronouns/success rate	Improvement (success rate)
ACC	130	106/81.54%	112/86.16%	4.62%
CDR	83	56/67.47%	59/71.09%	3.59%
BEO	68	41/60.30%	44/64.71%	4.41%
Total	281	203/72.25%	215/76.52%	4.27%

Table 7.8 Pronoun resolution for French before and after enhancement

File (French)	Number of pronouns	Before enhancement correctly resolved pronouns/success rate	After enhancement correctly resolved pronouns/success rate	Improvement (success rate)
ACC	156	107/68.59%	113/72.44%	3.85%
CDR	136	77/56.62%	84/61.77%	5.15%
BEO	98	43/43.88%	44/44.90%	1.02%
Total	390	227/58.21%	241/61.80%	3.59%

enhancement. During the analysis of the outputs of each of the resolvers the following cases were distinguished:

- The antecedent was initially wrongly identified for English, but correctly identified later due to the French gender filter (for 11 anaphors).
- The antecedent was correctly identified in English without the help of the French gender filter, and the antecedent was wrongly proposed for French (37).
- Both the English and the French pronoun resolvers proposed the wrong candidate (32).
- Both the English and the French pronoun resolvers identified the correct antecedent (26).

It should be noted that in all cases the gender filter in French helped the English module reduce its search space.

7.4 MARS: a re-implemented and improved fully automatic version

7.4.1 Fully automatic anaphora resolution

MARS[32] is a new implementation[33] of Mitkov's robust, knowledge-poor approach using the FDG[34]-parser as its main pre-processing tool. MARS operates in fully automatic mode, in contrast to the vast majority of approaches which rely on some kind of pre-editing of the input to the anaphora resolution algorithm[35] or which have only been manually simulated. As an illustration, Hobbs's naïve approach (1976, 1978) was not implemented in its original version. In Dagan and Itai (1990, 1991), Aone and Bennett (1995) and Kennedy and Boguraev (1996) pleonastic pronouns were removed manually,[36] whereas in Mitkov (1998b) and Ferrández et al. (1998) the outputs of the POS tagger and the NP extractor/partial parser were post-edited in a similar way to Lappin and Leass (1994) where the output of the Slot Unification Grammar parser was corrected manually. Finally, Ge et al.'s (1998) and Tetreault's (1999) approaches made use of annotated corpora and thus did not perform any pre-processing.

The development of MARS, and also the re-implementation of fully automatic versions of Baldwin's, as well as Kennedy and Boguraev's approaches for comparative purposes in a related project (Barbu and Mitkov 2000; see also Chapter 8), showed that fully automatic anaphora resolution is more difficult than previous work had suggested.[37] In the real world, fully automatic resolution must deal with a number of hard pre-processing problems such as morphological analysis/POS tagging, named entity recognition, unknown word recognition, NP extraction, parsing, identification of pleonastic pronouns, selectional constraints, etc. Each one of these tasks introduces error and thus contributes to a reduction in the success rate of the anaphora resolution system; the accuracy of tasks such as robust parsing and identification of pleonastic pronouns is way below 100%.[38] For instance, many errors will be caused by the failure of systems to recognise pleonastic pronouns – and their consequent attempt to resolve them as anaphors.

Given the limitations of pre-processing and the drop in performance that inevitably results, ways of improving the accuracy of anaphora resolution systems should be sought. One straightforward remedy is to use, if possible, high-quality pre-processing tools. Therefore, one of the best available 'super-taggers' in English – Conexor's FDG Parser (Tapanainen and Järvinen 1997) – was chosen for MARS. This super-tagger provides information on the dependency relations between words which allows the extraction of complex NPs. It also gives the lemmas and syntactic roles of words.

In the case of nominal anaphora resolution one of the obvious ways forward is to develop modules which recognise cataphora and non-anaphoric occurrences such as pleonastic pronouns or single out anaphors which have constituents other than NPs as their antecedents (e.g. clauses, sentences, discourse segments, etc.). This would prevent the algorithm from considering pronouns which either are not anaphoric or are outside the remit of the algorithm and thus save the drop in accuracy resulting from the assignment of antecedents to such pronouns. As a consequence, a special module for automatic identification of non-nominal anaphora was developed and included in MARS.[39]

7.4.2 Differences between MARS and the original approach

The initial implementation of MARS followed Mitkov's original approach closely, the main differences being (i) the addition of three new indicators and (ii) the change in the way some of the indicators were implemented or computed due to the available pre-processing tools. In its most recent version, MARS uses a program for automatically recognising instances of anaphoric or pleonastic pronouns and intrasentential syntax filters. As this book nears completion, experiments with a program for automatic gender identification are also under way.

The three new indicators included in MARS are:

- *Boost pronoun*: As NPs, pronouns are permitted to enter the list of candidates of other pronouns. The motivation for considering pronominal candidates is two-fold. Firstly, pronominalised entities tend to be salient. Secondly, the

NP corresponding to an antecedent may be beyond the range of the algorithm, explicitly appearing only prior to the two sentences preceding the one in which the pronoun appears. Pronoun candidates may thus serve as a stepping-stone between the current pronoun and its more distant nominal antecedent. On the other hand, it is not helpful if the system reports that the antecedent of a pronoun *it* is another pronoun *it*. When a pronoun is selected as the antecedent, the system has access to that pronoun's own antecedent in a fully transitive fashion so that an NP is always returned as the antecedent of a pronoun, even when this is accessed via one or more intervening pronouns. Given that pronominal mentions of entities may reflect the salience of their antecedents, pronouns are awarded a bonus of +1.

- *Syntactic parallelism*: The pre-processing software (FDG Parser) used by MARS also provides the syntactic role of the NP complements of the verbs. This indicator increases the chances that an NP with the same syntactic role as the current pronoun will be its antecedent by awarding it a boosting score of +1.
- *Frequent candidates*: This indicator was motivated by observations during annotation of coreference that texts frequently contain a narrow 'spine' of references, with perhaps fewer than three entities being referred to most frequently by pronouns throughout the course of the document. This indicator awards a boosting score (+1) to the three NPs that occur most frequently as competing candidates of all pronouns in the text.

Five of the original indicators are computed in a different manner by MARS. In the case of the indicator *lexical reiteration*, in addition to counting the number of explicit occurrences of an NP, MARS includes pronouns previously resolved to that NP. The conditions for boosting it remain the same.

Collocation match was originally implemented to boost candidates found in the same paragraph as the pronoun, preceding or following a verb identical or morphologically related to a verb which the pronoun precedes or follows. This indicator was modified so that in the first step, for every appearance of a verb in the document, the immediately preceding and immediately following heads (PHEAD and FHEAD in the FDG output below respectively) of NP arguments are written to a file. In the case of prepositions, the immediately following NP argument is written. As an illustration, after processing the text

> Do not touch the battery terminals with metal objects such as paper clips or keychains. Doing so can cause burns or start a fire. Carry batteries only within the printer or within their original packaging. Leave *the battery* inside the printer until you need to charge or replace *it*.

a file is generated with the following information for the verb *replace*[40]:

> VERB replace PHEAD you FHEAD it
> VERB replace PHEAD battery FHEAD cover
> VERB replace PHEAD printer FHEAD cartridge
> VERB replace FHEAD cartridge

VERB replace PHEAD You FHEAD cartridge
VERB replace FHEAD battery
VERB replace PHEAD battery FHEAD it
VERB replace PHEAD You FHEAD battery
VERB replace PHEAD problem FHEAD battery
VERB replace PHEAD you FHEAD battery
VERB replace PHEAD this FHEAD cartridge
VERB replace PHEAD Ink FHEAD Cartridge
VERB replace PHEAD that FHEAD cartridge
VERB replace FHEAD Cartridge
VERB replace FHEAD Ink

MARS consults this data file when executing *collocation match*. Resolving the pronoun *it* in the last sentence of the text above, the NP *the battery* is awarded a boosting score of +2 because the pronoun is in the FHEAD position with respect to the lemma of the verb *replace* and the lemma of the head of *the battery* also appears in the FHEAD position with respect to that verb in the database. Thus, the indicator applies on the basis of information taken from the whole document, rather than information only found in the paragraph.[41]

First NPs has been renamed *obliqueness*. Following centering theory (see Chapter 3, section 3.1) where grammatical function is used as an indicator of discourse salience, MARS awards subject NPs a score of +2, objects a score of +1, indirect objects no bonus, and to NPs for which the FDG Parser is unable to identify a function a penalising score of –1.[42]

A clause splitter has not yet been incorporated into MARS, so a simplified version of the *referential distance* indicator was implemented, with the distance being calculated only in terms of sentences rather than clauses and sentences.

Regarding the *term preference* indicator, in the first implementation of MARS significant terms were obtained by identifying the words in the text with the ten highest TF.IDF scores. Candidates containing any of these words were awarded the boosting score. In the current implementation, it is the ten NPs that appear with greatest frequency in the document that are considered significant. All candidates matching one of these most frequent NPs are awarded the boosting score.

MARS includes a program that automatically classifies each occurrence of *it* as an instance of nominal anaphora, as non-nominal anaphora, or as pleonastic (Evans 2000). The classification rate is currently reported to be 78.74%. Table 7.9 gives more details on the accuracy of this classification; further details on the pronoun classifying program are also outlined in Chapter 2, section 2.2.1.1.

Kennedy and Boguraev's (1996) syntax filters (see section 7.3.3.2) that act as knowledge-poor approximations of Lappin and Leass's (1994) syntax constraints (see section 5.3.1, Chapter 5) were also implemented in the latest version of MARS. These constraints are applied before activating the antecedent indicators and after the gender and number agreement tests.[43]

The algorithm implemented by MARS closely follows the original one (Mitkov 1998b), but phase 2 and phase 5 (see below) represent a significant point

of divergence between MARS and the original algorithm. MARS operates in five phases (Mitkov et al. 2001). In *phase 1*, the text to be processed is parsed syntactically, using Conexor's FDG Parser (Tapanainen and Järvinen 1997), which returns the parts of speech, morphological lemmas, syntactic functions, grammatical number and, most crucially, dependency relations between tokens in the text, which facilitates complex noun phrase (NP) extraction.

In *phase 2*, anaphoric pronouns are identified and non-anaphoric and non-nominal instances of *it* are filtered using the machine learning method described by Evans (2000). In its current implementation, MARS is only intended to resolve third person pronouns and possessives of singular and plural number that demonstrate identity-of-reference nominal anaphora.

In *phase 3*, for each pronoun identified as anaphoric, candidates are extracted from the NPs in the heading of the section in which the pronoun appears, and from NPs in the current and preceding two sentences (if available) within the paragraph under consideration. Once identified, these candidates are subjected to further morphological and syntactic tests.

Extracted candidates are expected to obey a number of constraints if they are to enter the *set of competing candidates*, i.e. the candidates that are to be considered further. Firstly, candidates are required to agree with the pronoun with respect to number and gender, as was the case in the original algorithm. Secondly, they must obey the syntactic filters (i) and (ii) as specified in section 7.3.3.2 above.

In *phase 4*, preferential and impeding factors (a total of 14) are applied to the sets of competing candidates. On application, each factor applies a numerical score to each candidate, reflecting the extent of the system's confidence about whether the candidate is the antecedent of the current pronoun.

Finally, in *phase 5*, the candidate with the highest composite score is selected as the antecedent of the pronoun. Ties are resolved by selecting the most recent highest-scoring candidate.

7.4.3 Optimisation of MARS

The scores of the antecedent indicators as proposed in Mitkov's original method were derived on the basis of empirical observations and have never been regarded as definite or optimal. The idea of optimising these scores so that they provide the best success rate of the algorithm has been under consideration since the early stages of development; later on, the antecedent indicators as implemented in MARS were optimised using a genetic algorithm.[44]

Given that the score of a candidate for antecedent is computed by adding the scores of each of the indicators, the algorithm can be represented as a function with 14 parameters, each one representing an antecedent indicator:

$$\text{score}_k = \sum_{i=1}^{i=14} x_{k_i}$$

where score_k is the composite score assigned to the candidate k and x_{k_i} is the score assigned to the candidate k by the indicator i. The goal of an optimisation

procedure (search method) would be to find the set of indicator scores for which the composite score is maximum for the antecedents and lower for the rest of the candidates. In other words, the optimisation seeks to find the set of indicators for which the success rate of the anaphora resolution algorithm would be maximal. Experiments with memory-based learning and the perceptron method were conducted to optimise MARS, but both of them performed poorly with success rates lower than the non-optimised version. By contrast, a genetic algorithm was found to be more suitable and led to improvement in performance. The optimisation of MARS is discussed in detail by Orasan and Evans (2000) and Orasan et al. (2000).

7.4.4 Evaluation of MARS

MARS was evaluated on eight different files, from the domains of computer hardware and software technical manuals, featuring 247 401 words and 2263 anaphoric pronouns[45] (Table 7.9). Of the anaphoric pronouns, 1709 were intrasentential anaphors and 554 were intersentential. Each text was annotated coreferentially in accordance with the methodology briefly outlined in Chapter 6, section 6.5 and by Mitkov et al. (2000).

The overall success rate of MARS was 59.35% (1343/2263). After using a genetic algorithm (Orasan et al. 2000), the success rate rose to 61.55% (1393/2263). Success rate is defined as the ratio of the number of anaphoric pronouns that MARS resolves correctly to the number of all anaphoric pronouns in the text. In 238 cases the antecedents were not on the list of candidates due to pre-processing errors.

Table 7.10 gives details on the evaluation of MARS – covering the standard version and the 'optimised' version in which the genetic algorithm was used to obtain the set of scores leading to optimal performance (and as a result an improvement of up to around 7% can be seen). As a result of errors at the level

Table 7.9 The characteristics of the texts used for evaluation of MARS

Text	Words	Anaphoric pronouns	Non-nominal anaphoric/ pleonastic *it*	Classification accuracy for *it*
ACC	9 753	157	22	81.54
CDR	10 453	83	7	92.86
BEO	7 456	70	22	83.02
MAC	15 131	149	16	89.65
PSW	6 475	75	3	94.91
WIN	2 882	48	3	97.06
SCAN	39 328	213	22	95.32
GIMP	155 923	1 468	313	83.42
Total	247 401	2 263	408	85.54

Table 7.10 Success rates for the different versions of MARS

		MARS											
		Standard				'Optimised'				MAX		Baseline	
Files	Old (2000)	Default	w/o *it* filter	w/o num/ gender agr	w/o syn constr	Default	w/o *it* filter	w/o num/ gender agr	w/o syn constr	Sct	Ptl	Recent	Random
ACC	33.33	51.59	52.87	35.67	49.04	55.41	55.41	43.31	43.31	73.88	96.18	28.02	26.75
BEO	35.48	60.00	60.00	45.71	60.00	67.14	64.28	50.00	67.14	81.43	95.71	35.71	22.86
CDR	53.84	67.47	68.67	51.81	67.47	75.90	74.69	54.22	74.69	78.31	95.18	36.14	43.37
GIMP	–	57.15	60.42	17.57	57.63	57.83	60.83	18.94	57.22	79.70	91.69	37.80	30.72
MAC	53.93	71.81	69.79	60.40	71.14	75.84	77.85	67.11	76.51	83.89	96.64	51.68	44.97
PSW	64.55	82.67	84.00	80.00	82.67	86.67	90.67	80.00	89.33	92.00	97.33	49.33	45.33
SCAN	–	61.50	62.44	46.48	60.56	63.85	64.79	51.64	63.85	79.81	87.32	32.39	30.52
WIN	33.32	52.08	62.50	39.58	52.08	68.75	66.67	60.42	68.75	81.25	87.50	37.50	18.75
Total	45.81	59.35	61.82	29.03	59.35	61.55	63.68	32.04	60.41	80.03	92.27	37.78	31.82

of NP extraction, and therefore possible omission of antecedents, the success rate of MARS cannot reach 100%. In the *MAX* columns the maximum success rates that MARS can obtain are indicated. The column *Sct* represents the maximum possible success rate when a pronoun is considered correctly resolved if the whole NP representing its antecedent is selected as such. As can be seen, this figure does not exceed 92%. Given the pre-processing errors, inevitable in an automatic system, a pronoun was considered to be correctly resolved if only part of the NP which represented its antecedent was identified, and that part included the head of the NP (as proposed in MUC-7). When this partial matching is considered, the maximum success rate can reach the values presented in the column *Ptl*. Two baseline models, presented in the *Baseline* columns, were evaluated, one in which the most recent candidate was selected as the antecedent and one in which a candidate was selected at random – both after agreement restrictions had been applied. The column *Old* displays the performance of a fully automatic implementation of a slightly modified version of the original algorithm.[46] This implementation did not include any additional components such as new or modified indicators or recognition of pleonastic pronouns, but neither did it operate on clauses as in the case of the original algorithm.[47]

MARS was evaluated in four different configurations: *Default*, in which the algorithm was run in its full version as outlined in 7.4.2; *w/o it filter*, where the system was run without attempting to identify pleonastic/non-nominal instances of *it*; *w/o num/gender agr*, where the system was run without applying number and gender agreement constraints between pronouns and candidates, and *w/o syn constr*, where no syntactic constraints were enforced between pronouns and intrasentential candidates. By comparing these columns with the *Default* column, for example, it is possible to see that, overall, MARS gains around 30% in performance as a result of enforcing number and gender agreement between pronouns and competing candidates. The contribution made by the syntactic constraints was not as high as expected due to their reliance on an accurate parse which was not always obtained for the texts processed. For each configuration and each text, the obtained success rate is displayed in the column *Standard*. Additionally, the genetic algorithm was used to find the upper limit of MARS's performance when the optimal set of indicator scores is applied, which is shown in the column '*Optimised*'.[48]

Interestingly MARS's performance was slightly worse (in terms of success rate) when recognition of pleonastic/non-nominal *it* was attempted. This was due to inaccuracies in the classification module with anaphoric instances of *it* being incorrectly filtered. In fact the success rate did not reflect the positive contribution made by the classification module and as a consequence a new measure of performance was proposed (Mitkov et al. 2001).

7.5 Automatic multilingual anaphora resolution

Mitkov's robust, knowledge-poor approach served as a basis for the development of fully automatic pronoun resolution systems for Bulgarian.

7.5.1 *Fully automatic version for Bulgarian*

The adaptation of Mitkov's approach for Bulgarian benefits from a suite of language processing tools specially developed to support the resolution algorithm.[49] The suite consists, in order of processing, of the following modules: tokeniser, sentence splitter, paragraph segmenter, POS tagger,[50] clause chunker, NP grammar and section heading extractor. Each one of the pre-processing modules, except for the NP parser, showed precision and recall in the 90% range.[51] In addition, a hand-crafted[52] small term bank containing 80 terms from the domains of programming languages, word processing, computer hardware and operating systems was made use of.[53] This bank also featured 240 phrases containing these terms.

In addition to the original set of indicators, the Bulgarian version of Mitkov's approach used three new indicators: *adjectival noun phrases*, *proper name preference* and *selectional restriction pattern*.

- *Adjectival noun phrases*: Noun phrases which contain adjectives modifying the head are awarded a score of 1.

Empirical analysis shows that in Bulgarian constructs of that type are more salient than NPs consisting simply of a noun or of an indefinite article and a noun.

- *Proper name preference*: Noun phrases that represent entity names (person, organisation, product, etc.) are more likely to be referred to; such NPs are awarded a score of 2.[54]

It was found that entity names are better candidates for antecedent than terms, hence proper names are given a higher bonus than terms (recall that term preference assigns a score of 1).[55]

- *Selectional restriction pattern*: Noun phrases occurring in the same collocation as the anaphor with respect to a specific verb are given a bonus of 2.

This preference is different from the *collocation match* preference as defined in Mitkov's original approach in that it operates on a wider range of 'selectional restriction patterns' associated with a specific verb[56] and not on exact lexical matching within a paragraph. As an illustration assume that 'Delete file' has been identified as a legitimate collocation, being a frequent expression in a domain-specific corpus, and consider the example

(7.19) Make sure you save *the file* in the new directory. After that you can delete *it*.

Whereas the 'standard' *collocation match* will not be activated here[57] the *selectional restriction pattern* will identify *delete file* as an acceptable construction and will reward the candidate *the file*.

The evaluation of the Bulgarian version suggested a success rate of 75% on a corpus of software manuals containing 221 anaphors. An optimised version[58] of the anaphora resolution system scored a success rate of 78.8% on these texts. Given that the anaphora resolution system operates in a fully automatic mode, this result could be considered very satisfactory. It should be noted that many of

Table 7.11 Evaluation results of the Bulgarian version of Mitkov's knowledge-poor approach

		Success rate		Success rate
Text	Pronouns	Standard	Optimised	Baseline most recent
Software manuals	221	75.0	78.8	58.0
Tourist guides	116	68.1	69.8	65.0
All texts	337	72.6	75.7	60.4

the errors arise from the inaccuracy of the pre-processing modules such as clause splitting and NP extraction.

The anaphora resolution system was also evaluated in the genre of tourist texts. As expected, the success rate dropped to 68.1% which, however, can still be regarded as a very good result, given the fact that neither manual pre-editing of the input text nor any post-editing of the output of the pre-processing tools was undertaken. The main reason for the decline of performance is that some of the original indicators, such as term preference, immediate reference and sequential instructions of the knowledge-poor approach, are genre specific.

The performance of the approach was compared with that of a baseline model which selects as antecedent the most recent NP which matches the anaphor in gender and number. Table 7.11 presents a summary of the evaluation results.

The explanation for the higher success rate of the Bulgarian version of Mitkov's approach as opposed to MARS is the fact that Bulgarian is gender sensitive and the gender (masculine, feminine and neuter) constraints successfully filter out the majority of unacceptable candidates.

7.6 Summary

This chapter has presented Mitkov's knowledge-poor approach to pronoun resolution, in both its original version and its latest fully automatic version, referred to as MARS. The evaluation results show that due to the inevitable errors in automatic pre-processing, MARS's performance is lower than the original algorithm which operates on post-edited outputs from a part-of-speech tagger and NP extractor. The chapter also discusses the multilingual nature of Mitkov's algorithm, which has been successfully adapted to languages such as Polish, Arabic, French and Bulgarian. Similarly to MARS, the algorithm for Bulgarian works in fully automatic mode but yields a higher success rate, the main reason being the gender discrimination in this language. Whereas the aforementioned multilingual developments are based on extensions of Mitkov's original algorithm to other languages, this chapter has also discussed another multilingual project where the English and French versions of Mitkov's algorithm benefit from a bilingual corpus and mutually enhance their performance.

Notes

1 Following established conventions, in this chapter the approach is often referred to as *Mitkov's approach*, or *the knowledge-poor approach*.
2 The approach has become better known through a later updated publication (Mitkov 1998b).
3 Called *antecedent indicators* (see section 7.1.2).
4 MARS (Mitkov's Anaphora Resolution System) is an acronym coined by Richard Evans.
5 The approach handles only pronominal anaphors whose antecedents are noun phrases.
6 Typical of the genre of user guides.
7 *Portable StyleWriter*. User's guide. Apple Computers, 1994.
8 The negative score applies for versions of the algorithm which use a search scope of three or more sentences.
9 Example adapted from Dagan et al. (1995).
10 A sentence splitter would have already segmented the text into sentences, a POS tagger would have determined the parts of speech and a simple phrasal grammar would have detected the noun phrases.
11 In this project cataphora was not treated.
12 Note that this restriction may not always apply in languages other than English (e.g. German).
13 Computed as the ratio (correctly resolved anaphors) / (number of all anaphors).
14 The approach was also evaluated for other languages such as Polish, Bulgarian and Arabic.
15 Or a more appropriate comprehensive sampling procedure.
16 This figure was obtained on a comparatively small set of data; recent unpublished tests have confirmed the good critical success rates of the approach.
17 The recent version of the knowledge-poor approach referred to as MARS was also compared to a baseline model which randomly selects the antecedent from all candidates surviving the agreement restrictions (see section 7.4).
18 First noun phrases 1 + indicating verbs 1 + lexical reiteration 0 + section heading 0 + collocation 0 + immediate reference 0 + sequential instructions 0 + term preference 1 + indefiniteness 0 + prepositional noun phrases 0 + referential distance 2 = 5.
19 First noun phrases 0 + indicating verbs 0 + lexical reiteration 0 + section heading 0 + collocation 0 + immediate reference 0 + sequential instructions 0 + term preference 1 + indefiniteness 0 + prepositional noun phrases –1 + referential distance 2 = 2.
20 The MARS version of Mitkov's approach was compared both with Baldwin's approach and with Kennedy and Boguraev's (1996) parser-free method. Chapter 8, section 8.6 provides more details on that evaluation.
21 This table reports on the results of the original knowledge-poor approach: it does not include the evaluations conducted for MARS.
22 See, however, the evaluation workbench (Chapter 8, section 8.6) when MARS is compared with Kennedy and Boguraev's (1996) approach and once again with Baldwin's CogNIAC.
23 The approach was recently implemented for French (see section 7.2.4) as part of a bilingual project based on the so-called 'mutual enhancement' methodology (see section 7.3). Mitkov's approach was also implemented as part of a fully automatic system in Bulgarian (see section 7.5.1).
24 There are other forms of definiteness in Arabic which are not discussed here since they are not typical of technical manuals.
25 The critical success rate for English was measured on a subset of the evaluation data. No evaluation of the critical success rate was conducted for the Polish direct version; also, no baseline models were implemented for Arabic.

26 The implementation for French as well as the implementation of the mutual enhancement bilingual strategy were carried out by Catalina Barbu.
27 These exceptions were not found in the bilingual corpus.
28 These two examples are from Cornish (1986).
29 The term *equivalent* is used to denote the French translation of an English word or the English translation of a French word.
30 The expression 'to be resolved directly' refers to the cases where there is only one (singular or plural) candidate for antecedent.
31 Since the algorithm for English performs with a higher success rate, it has been decided that the algorithm for French needs a larger margin (4 as opposed to 3) if its output were to be preferred.
32 See note 4 above.
33 MARS was implemented and fine-tuned by Richard Evans.
34 FDG stands for Functional Dependency Grammar.
35 This statement refers to anaphora resolution systems and not to the coreference resolution systems implemented for MUC-6 and MUC-7.
36 Moreover, Dagan and Itai (1991) undertook additional pre-editing such as removing sentences for which the parser failed to produce a reasonable parse, cases where the antecedent was not an NP, etc. Kennedy and Boguraev (1996) manually removed 30 occurrences of pleonastic pronouns (which could not be detected by their pleonastic recogniser) as well as 6 occurrences of *it* which referred to a VP or prepositional constituent.
37 Fully automatic anaphora resolution means that there is no human intervention at any stage: such intervention is sometimes large-scale, such as manual simulation of the approach, and sometimes smaller-scale, as in the cases where the evaluation samples are stripped of pleonastic pronouns or anaphors referring to constituents other than NPs.
38 The best accuracy reported in robust parsing of unrestricted texts is around the 86% mark; the accuracy of identification of non-nominal pronouns is under the 80% mark though Paice and Husk reported 92% for identification of pleonastic *it*.
39 The program for automatic identification of non-nominal anaphora was developed by Richard Evans (Evans 2000). See also Chapter 2, section 2.2.1.1 for more details.
40 In fact the file is bigger and features information on all appearances of each verb in the document.
41 Recently a generalisation of the collocation match indicator has been experimented with using WordNet hierarchical relations. In this experiment if concepts were involved in a specific pattern, then their superordinates were included too. The deployment of WordNet did not produce the expected improvement of performance due to the facts that initially no word sense disambiguator was made use of and that the evaluation data contained many domain-specific concepts not occurring in WordNet. As I put the finishing touches to this book, new experiments involving a word sense disambiguation program and statistical significance measures for patterns are under way.
42 Note that FDG proposes grammatical functions for most words. The POS tagger used in Mitkov's original version was not able to identify syntactic functions, and first NPs in sentences or clauses were used as approximations of subjects.
43 A gazetteer of first names was initially used for checking gender agreement. A recent experiment making use of automatic WordNet-based procedures for gender identification (Orasan and Evans 2001) and identification of animate entities (Evans and Orasan 2000) did not result in improvement of the performance of MARS, mainly due to the fact that many of the senses appearing in the genre of technical manuals were not present in the WordNet ontology.
44 The optimisation through genetic algorithm was carried out by Constantin Orasan.

45 Pronouns that exhibit nominal identity-of-reference anaphora; MARS's operation is restricted to this class of anaphors.
46 Note that Mitkov's original algorithm did not operate in fully automatic mode since the outputs of the POS tagger and the NP extractor were post-edited.
47 In this implementation there was no access to information about clauses.
48 The quotes are used to draw attention to the fact that in this case, in order to explore the limitations of MARS, the genetic algorithm was used as a search algorithm rather than as a general optimisation method.
49 The implementation of the Bulgarian version and the development of the pre-processing tools for Bulgarian were carried out by Hristo Tanev.
50 The POS tagger was based on the BULMORPH morphological analyser (Krushkov 1997) enhanced by disambiguation rules.
51 For further details on the development and performance of the pre-processing tools see Tanev and Mitkov (2000).
52 This was done for experimental purposes. In future applications the incorporation of automatic term extraction techniques is envisaged.
53 See section 7.4.2 for the ways in which MARS automatically computes the indicator *term preference*.
54 While this preference appears to be 'universal', so far it has been implemented in the Bulgarian version only.
55 In this project the set of proper names and that of terms are considered to be disjoint.
56 At the time of writing, these patterns are extracted from a list of frequent expressions involving the verb and domain terms in a purpose-built term bank, but generally they can be automatically collected from large domain-specific corpora.
57 The original *collocation match* indicator as reported in Mitkov (1998b) would have been activated on a text such as 'Make sure you do not delete the file yet. After that, of course, you can delete it'.
58 The optimisation made use of genetic algorithms in a manner similar to that described in Orasan et al. (2000) and section 7.4.3.

CHAPTER EIGHT

Evaluation in anaphora resolution

Evaluation is the driving force for progress in research and development, and is essential for every NLP system. Evaluation provides a means both of assessing the individual performance of a system and of determining where it stands compared to other approaches and applications. The evaluation culture now prevailing in NLP is connected with the move towards 'engineering' solutions where a system is evaluated by running it against a practical task, such that this task can be used to measure objectively the system's performance against competing systems. The growing interest and research in evaluation have also been inspired by the availability of annotated corpora.

Most of this chapter is based on my recent work on evaluation in anaphora resolution.[1] To begin, I shall argue that it is necessary to distinguish the evaluation of an algorithm and the evaluation of an anaphora resolution system, and shall discuss the measures of recall and precision. Following this, I shall propose a package of measures and comparative evaluation tasks for anaphora resolution. The chapter will then proceed to discuss the evaluation of anaphora resolution systems and the reliability of the evaluation results. Finally, an evaluation workbench for anaphora resolution will be presented and other proposals will be outlined.

8.1 Evaluation in anaphora resolution: two different perspectives

I maintain (in Mitkov 2001b) that evaluation in anaphora resolution should be addressed from two different perspectives depending on whether the evaluation only focuses on the anaphora resolution algorithm or if it covers the performance of the anaphora resolution system. I propose a distinction between *evaluation of anaphora resolution algorithms* and *evaluation of anaphora resolution systems*. By *anaphora resolution system* I refer to a whole implemented system that processes input at various levels such as morphological, syntactic, semantic, discourse, etc., and feeds the analysed text to the anaphora resolution algorithm.

A natural way to test an anaphora resolution algorithm is to let it run in an 'ideal environment' without taking into consideration any possible errors or complications which occur at various pre-processing stages. In contrast, when

evaluating an anaphora resolution system, one will certainly have to face a drop in performance due to the impossibility of analysing natural language with absolute accuracy. A number of anaphora resolution systems either operate on human-controlled inputs (e.g. pre-analysed corpora or human-corrected outputs from pre-processing modules) or are manually simulated, which suggests that the evaluation they report is only concerned with the anaphora resolution algorithm itself. On the other hand, there are systems which fully process the text before it is sent to the anaphora resolution algorithm, and their evaluation is usually concerned with the evaluation of the entire anaphora resolution system.[2] Based on this distinction, the evaluations reported in sections 7.1.5, 7.2.2, 7.2.3 and 7.3.5 dealt with the performance of the algorithm, whereas the evaluations described in 7.4.4 and 7.5.1 addressed the performance of the anaphora resolution system.

With this distinction acknowledged, it is desirable that the evaluation of an anaphora resolution algorithm be performed alongside the evaluation of the anaphora resolution system of which it is part. In fact, it is possible that an anaphora resolution system that performs poorly is still based on a very effective algorithm. In this case better pre-processing tools should be considered.

8.2 Evaluation in anaphora resolution: consistent measures are needed

The Message Understanding Conferences (MUCs) introduced the measures *recall* and *precision* for coreference resolution. These measures have been adopted by a number of researchers for evaluation of anaphora resolution algorithms or systems. I argue that these measures, as defined, are not satisfactory or sufficiently clear when applied to the evaluation of anaphora resolution algorithms (Mitkov 2000, 2001b). I base my arguments on the following definitions of recall and precision.

Definition 1 (Aone and Bennett 1995)

$$\text{Recall} = \frac{\text{Number of correctly resolved anaphors}}{\text{Number of all anaphors identified by the program}}$$

$$\text{Precision} = \frac{\text{Number of correctly resolved anaphors}}{\text{Number of anaphors attempted to be resolved}}$$

Definition 2 (Baldwin 1997)[3]

$$\text{Recall} = \frac{\text{Number of correctly resolved anaphors}}{\text{Number of all anaphors}}$$

$$\text{Precision} = \frac{\text{Number of correctly resolved anaphors}}{\text{Number of anaphors attempted to be resolved}}$$

To begin, Definitions 1 and 2 describe precision in the same way, but they compute recall differently for anaphora resolution systems: Aone and Bennett include only the anaphors identified by the program, whereas Baldwin considers 'all

anaphors'.[4] Aone and Bennett's definition of recall considers only anaphors identified by the program and not all anaphors, thus preventing this measure from being sufficiently indicative of the resolution success. In fact, the program could end up identifying only a certain number of anaphors that are easy to resolve and the recall obtained would not provide a realistic picture of the performance. In addition, for robust algorithms that always propose an antecedent, Definition 1 would not be able to distinguish between recall and precision since 'the number of all anaphors identified by the program' would be equal to the 'number of anaphors attempted to be resolved'.

Next, while in the definition of precision the set of 'anaphors attempted to be resolved' makes sense for certain algorithms which leave pronouns unresolved (e.g. pronouns which are ambiguous or cannot be resolved by the algorithm), systems that only attempt to resolve pronouns with a single candidate would obtain unfairly high precision.

In view of the inconsistencies arising from the definition and use of recall and precision in evaluating algorithms, I propose instead the measure *success rate* which simply reflects the resolution success of an algorithm against all anaphors (as marked by human annotators) in the evaluation corpus (see 8.3.1). Since in this case the success rate focuses on the performance of a specific algorithm, it is assumed that the input to the algorithm is correct.

8.3 Evaluation package for anaphora resolution

Using my knowledge-poor approach (see Chapter 7) as a testbed, I propose an evaluation package for evaluating anaphora resolution algorithms consisting of (i) performance measures, (ii) comparative evaluation tasks and (iii) component measures. The first cover the overall performance of the algorithm and the second compare the algorithm with other approaches, whereas the third look at the efficiency of the separate components of the algorithm. These measures are transferable to the evaluation of anaphora resolution systems, but the figures obtained in this case will reflect the performance of the whole system and not just the resolution module.

The performance measures are *success rate*, *critical success rate* and *non-trivial success rate*. The comparative evaluation tasks include evaluation against *baseline models*, comparison with *similar approaches* and comparison with *well-established algorithms*. The measures applied to evaluate separate components of the algorithm are *decision power* and *relative importance*.

8.3.1 Evaluation measures covering the resolution performance of the algorithm

The proposed measures are illustrated and have been tested on pronominal anaphors, but they can equally be applied to lexical noun phrase anaphors. I restrict the validity of these measures to nominal anaphora, which is the most extensively studied and best understood type in Computational Linguistics.

The **success rate** for an anaphora resolution algorithm

$$\text{Success rate}_{\text{Anaphora resolution algorithm}} = \frac{\text{Number of successfully resolved anaphors}}{\text{Number of all anaphors}}$$

reflects the resolution success of an algorithm against all anaphors[5] in the evaluation corpus and is normally expressed as a percentage. Since this measure focuses on the performance of the algorithm and not on any pre-processing modules, the exact success rate will be obtained if the input to the anaphora resolution algorithm is either post-edited by humans or extracted from an already tagged corpus.[6]

The measure **non-trivial success rate** applies only to the anaphors which have more than one candidate for antecedent, removing those preceded by only one NP in the search scope of the algorithm (and therefore having only one candidate) since their resolution would be trivial.

The measure **critical success rate** applies only to those 'tough' anaphors which still have more than one candidate for antecedent after gender and number filters. This measure can be very indicative in that it can point to misleading results based on the evaluation of data containing only very easy-to-resolve anaphors (e.g. anaphors that can be resolved directly after gender agreement checks).

More formally, let N be the set of all anaphors involved in an evaluation, and S the set of anaphors which have been successfully resolved. Further, let K be the set of anaphors which have only one candidate for antecedent (and which therefore are correctly resolved in a trivial way), M the set of anaphors which are resolved on the basis of gender and number agreement and let $n = \text{card}(N)$[7], $s = \text{card}(S)$, $k = \text{card}(K)$ and $m = \text{card}(M)$. Clearly $s \leq n$, $k \leq s$, $k + m \leq s$, $k \geq 0$, $m \geq 0$, $s \geq 0$. The following relations hold[8]:

(8.1) success rate ≥ non-trivial success rate ≥ critical success rate

since

$$\text{success rate} = \frac{s}{n}, \quad \text{non-trivial success rate} = \frac{s-k}{n-k},$$

$$\text{critical success rate} = \frac{s-k-m}{n-k-m}$$

and

$$\frac{s}{n} \geq \frac{s-k}{n-k} \geq \frac{s-k-m}{n-k-m}, \quad k \geq 0, m \geq 0, s \geq 0.$$

As an illustration, consider evaluation data containing 100 anaphors. Assume that 20 of these anaphors have only one candidate for antecedent and that the antecedents of a further 10 anaphors can be determined only on the basis of gender and number agreement. Furthermore, let us assume that the algorithm resolves 80 of the anaphors correctly. The success rate would then be 80/100 = 80%, the non-trivial success rate would be 60/80 = 75% and the critical success rate 50/70 = 71.4%.

The non-trivial success rate is indicative of the performance of the algorithm in that it removes anaphors that have no competing candidates for antecedents

from the evaluation. The critical success rate is an important criterion for evaluating the efficiency of the factors employed by the anaphora resolution algorithms in 'critical cases' where agreement constraints alone cannot point to the antecedent.[9] It is logical to assume that good anaphora resolution algorithms have high critical success rates which are close to the overall success rates. In fact, it is really the critical success rate that matters: high critical success rate typically implies high overall success rate.

In the case of Mitkov's algorithm the critical success rate exclusively accounts for the performance of the antecedent indicators since it is associated with anaphors whose antecedents can be tracked down only with the help of the antecedent indicators.

8.3.2 Comparative evaluation tasks

The performance of a specific approach can be compared to a representative set of other algorithms and models, indicating where the approach stands in the state of the art of anaphora resolution. Three classes of comparative tasks are presented here: evaluation against baseline models, evaluation against approaches that share a similar 'philosophy' and evaluation against well-established (benchmark) approaches in the field.

The **evaluation against baseline models** is important to provide information as to how effective an approach is, by comparing it with unsophisticated, basic models. This type of evaluation justifies the usefulness of the approach developed: however high the success rate may be, it may not be worth while developing a specific approach unless it demonstrates clear superiority over simple baseline models. I compared my knowledge-poor approach with (i) a baseline model which checks agreement in number and gender and, where more than one candidate remains, picks out as antecedent the most recent subject matching the gender and number of the anaphor, and (ii) a baseline model which selects as antecedent the most recent noun phrase that matches the gender and number of the anaphor (Mitkov 1998a, 1998b).

The most recent version of Mitkov's knowledge-poor approach, referred to as MARS, was also compared to a baseline model which randomly selects the antecedent from all candidates surviving the agreement restrictions (Table 7.10). An even weaker baseline model would be to select randomly any candidate before any agreement checks.

The comparison with other similar methods (if available) or with other well-known approaches helps to discover what the new approach brings to the current state of play in the field. As an illustration, a **comparison to similar approaches** included running Breck Baldwin's CogNIAC algorithm (Baldwin 1997) on part of the evaluation texts (Table 7.3) on which Mitkov's approach had already been run (Mitkov 1998a). CogNIAC was chosen here because Mitkov's and Baldwin's approaches share common principles – both are regarded as knowledge-poor and use POS taggers rather than parsers. The MARS version of Mitkov's approach was compared both with Baldwin's approach and with Kennedy and Boguraev's (1996) parser-free method. Section 8.6 provides more

details on that evaluation. This type of evaluation has also been used by Tetreault (1999) who compares his centering-based pronoun resolution approach with similar methods such as Strube's S-list approach.[10]

Finally, with regard to the **comparison with well-established approaches**, Hobbs's method has been used for a benchmark evaluation by a number of researchers (Baldwin 1997; Mitkov 1998b; Tetreault 1999; Walker 1989). The BFP algorithm (Brennan et al. 1987) has also been used for comparison (Tetreault 1999).

However, I should point out that it is difficult to draw direct comparisons between resolution algorithms using any of the aforementioned methods. Even if the evaluations are performed over the same test data, the pre-processing tools used in the resolution systems may introduce enough variation or error to cast a shadow of uncertainty over any direct comparisons. The evaluation workbench, described in section 8.6, provides a solution to this problem.

8.3.3 Evaluation of separate components of the anaphora resolution algorithm

It is important to evaluate the performance of the separate components of anaphora resolution algorithms because this type of assessment provides useful insights as to how the approach may be further improved. In particular the evaluation of each resolution factor gives an idea of its significance or contribution and provides a basis upon which the factor scores can be adjusted[11] with a view to attaining an overall improvement in the approach. I carried out an evaluation of each antecedent indicator of my knowledge-poor algorithm and concluded that there are two measures of significance: the *decision power*, which reflects the influence of each indicator on the final choice of antecedent and the *relative importance*, which is regarded as the relative contribution of a specific factor in that it is computed by measuring the drop in performance if this indicator is removed. In the following discussion, these measures will be illustrated on the set of antecedent indicators outlined in Chapter 7, section 7.1.2, although it should be noted that they can be computed for any set of anaphora resolution factors.

Decision power is the measure of the influence of each factor (or indicator in the case of Mitkov's approach) on the final decision, its ability to 'impose' its preference in line with, or contrary to, the preference of the remaining factors (indicators).[12] I define the decision power (DP_K) of a boosting indicator K in the following way:

$$DP_K = \frac{S_K}{A_K}$$

where S_K is the number of cases where the candidate to which the indicator K has been applied has been selected as the antecedent, while A_K is the number of all applications of this indicator. For the penalising indicators *prepositional noun phrase* and *indefiniteness* this figure is calculated as

$$DP_K = \frac{\text{Non-}S_K}{A_K}$$

Table 8.1 Decision power values for the antecedent indicators used in Mitkov's knowledge-poor approach (Chapter 7)

Indicator	Decision power	Comments
Immediate reference	1	Very decision-powerful, points always to the correct candidate
Prepositional noun phrase	0.922	Very decision-powerful and discriminating
Collocation	0.909	Very decision-powerful and discriminating
Section heading	0.619	Fairly decision-powerful, but alone cannot impose the antecedent
Lexical reiteration	0.585	Sufficiently decision-powerful
First NP	0.493	Averagely decision-powerful
Term preference	0.357	Not sufficiently decision-powerful
Referential distance	0.344	Not sufficiently decision-powerful

where Non-S_K is the number of cases where the candidate to which the indicator K has been applied has not been selected as the antecedent; A_K again is the number of all applications of this indicator. The *immediate reference* emerges as the most 'influential' indicator, followed by *prepositional noun phrases* and *collocation pattern preference* (Table 8.1). The relatively low figures for the majority of (seemingly very useful) indicators should not be regarded as a surprise: one should bear in mind firstly that in most cases a candidate is picked (or rejected) as an antecedent on the basis of applying a number of different indicators and, secondly, that most anaphors have a relatively high number of candidates for antecedent.

Another way of measuring the importance of a specific factor (indicator) would be to evaluate the approach with this factor 'switched off'.[13] This measure is called **relative importance** since it shows how important the presence of a specific factor is. Relative importance (RI_K) for a given indicator K is defined as

$$RI_K = \frac{SR - SR_{-K}}{SR}$$

where SR_{-K} is the success rate obtained when the indicator K is excluded, and SR is the success rate (with all the indicators on). In other words, this measure expresses the non-absolute, relative contribution of this indicator to the 'collective efforts' of all indicators, showing how much the approach would lose out if a specific indicator were removed. It should be noted that being relatively important does not mean decision-powerful and confident, and vice-versa. For instance, it was found that *referential distance* has the highest value (4.6% or 5.7% after the genetic algorithm was applied) for relative importance (Mitkov et al.

2001),[14] whereas this factor is among the least confident ones. One possible explanation comes from the fact that indicators such as *immediate reference* and *collocation pattern preference* are applied relatively seldom and even though they impose their decision very strongly towards the correct antecedent, they do not score very highly as 'relatively important' factors given their infrequent intervention. Finally, due to the complicated interactions of all indicators, there is no direct correlation between these two measures.

8.4 Evaluation of anaphora resolution systems

The **success rate** of anaphora resolution systems is defined in a similar way to that for anaphora resolution algorithms. However, the success rate for anaphora resolution systems reflects, in addition to the resolution rate of the algorithm implemented, the overall performance of the system, which is critically affected by its ability to carry out successful pre-processing. The correct identification of noun phrases, which are regarded as candidates for antecedents of nominal anaphora, is a crucial part of the pre-processing. Errors and variations in NP identification can drastically affect the success rate of an anaphora resolution system. The success rate of a specific anaphora resolution system is expressed as the ratio:

$$\text{Success rate}_{\text{Anaphora resolution system}} = \frac{\text{Number of successfully resolved anaphors}}{\text{Number of all anaphors}}$$

where *Number of all anaphors* is all anaphoric occurrences in the evaluation text as identified by humans. This definition assumes that the identification of anaphors (and therefore the identification of non-anaphoric NPs including non-anaphoric pronouns) is the responsibility of the system. Since the pre-processing is expected to be automatic, it is likely that the system may miss some anaphors or candidates for antecedents, which could result in a reduction in the success rate.

It is proposed that in addition to measuring the success rate of the anaphora resolution system, it would be useful to calculate the success rate of the anaphora resolution algorithm by running it on perfectly analysed inputs (Fukumoto et al. 2000). Such a measure sheds light on the limitations of a specific algorithm, provided that the pre-processing is 100% correct.

On the other hand, Mitkov et al. (2002) observe that the current definition of success rate does not capture cases where the program incorrectly tries to resolve instances of non-nominal anaphora.[15] For programs handling nominal anaphora, it is important to be able to judge the efficiency of the program in removing instances of non-nominal anaphora and as opposed to incorrectly attempting to resolve these instances to NPs. To this end a measure called **resolution etiquette**, which reflects this efficiency, is proposed. This measure is defined as follows:

$$\text{Resolution etiquette}_{\text{Anaphora resolution system}} = \frac{A' + B'}{A + B}$$

where A' is the number of correctly resolved nominal anaphors (out of a total of A) and B' is the number of pronouns, definite descriptions or proper names which are correctly filtered as instances of non-nominal anaphora (out of a total of B).[16] The resolution etiquette captures the contribution made to the system by both recognition modules for non-nominal and pleonastic pronouns and the anaphora resolution module itself, and is intended to describe a system's ability to 'behave appropriately' in response to a set of anaphors.

The measures *non-trivial success rate* and *critical success rate* can be applied to anaphora resolution systems as well. It should be noted, however, that the inequalities (8.1) may not hold in a fully automatic processing environment and therefore in this case these measures may not be as indicative as they are for the evaluation of algorithms. As an illustration, consider the scenario when an anaphora resolution system extracts no candidates for an anaphor due to pre-processing errors. The 'standard' success rate includes this set and none of them is correctly resolved, resulting in a drop in the success rate. In computing critical success rate, however, these 'always wrong' anaphors are excluded because they do not have more than one candidate after agreement filters have been applied, and so at times there can be a higher score for critical success rate than for 'standard' success rate.

Comparison with *baseline models* is particularly important when evaluating anaphora resolution systems. Table 7.10 in Chapter 7 shows the results from comparing MARS with a baseline model which selects as antecedent the most recent NP matching the anaphor in gender and number, and with a baseline model which picks a randomly generated NP from the list of candidates as antecedent.

The question that still remains is how to evaluate systems that are *almost* 'automatic' in the sense that they may involve some (but not full) human intervention – for instance, the elimination of anaphors whose antecedents are VPs and other non-NP constituents in the case of anaphora resolution systems that handle nominal anaphora only. One way of ensuring a fair comparison would be to run such systems in a 'fully automatic mode' and provide these results as well.

8.5 Reliability of the evaluation results

A major issue in the evaluation of an anaphora resolution algorithm or an anaphora resolution system is the reliability of the results obtained. One mandatory question is how definitive the evaluation results can be considered. To start, it has to be pointed out that the majority of anaphora resolution systems report results from tests on one genre only. Next, whether the evaluation is restricted to only one genre or not, the validity of evaluation largely depends on the size, representativeness and statistical significance of the evaluation corpus. The results are expected to be more reliable if the evaluation data is very large, covering not hundreds of anaphors but many thousands: it has already been seen that even in the same genre, results may differ if the samples are not large enough (Table 7.1, Chapter 7). Theoretically speaking, the success rate or other

evaluation measures could be regarded as definitive only if the approach were tested on all naturally occurring texts, which is an unrealistic task. Nevertheless, this consideration highlights the advantages of carrying out the evaluation task automatically. Automatic evaluation requires a large corpus with annotated coreferential links, against which the output of the anaphora resolution systems is to be matched. Chapter 6 provides more information on the existence and the development of coreferentially annotated corpora, with a view to using them in the evaluation process.

An alternative method to obtain more reliable results would be to employ comprehensive sampling procedures. It might be worth while experimenting not only with the selection of random samples, but also with selecting them in such a way that no two anaphors are located within a window of N (e.g. $N = 100$) sentences (Mitkov 2001b). It is believed that such a sampling process will produce statistically more significant results.[17]

The issue as to how reliable or realistic the obtained performance figures are largely depends on the nature of the data used for evaluation. Some evaluation data may contain anaphors more difficult to resolve, such as anaphors that are (slightly) ambiguous and require real-world knowledge for their resolution, or anaphors that have a high number of competing candidates, or that have their antecedents far away, etc., whereas other data may have most of their anaphors with single candidates for antecedent. Therefore it is suggested that in addition to the evaluation results, information should be provided as to how difficult to resolve the anaphors in the evaluation data are.[18] To this end more research is needed to come up with suitable measures for quantifying the average 'resolution complexity' of the anaphors in a certain text. In the meantime, I propose that simple statistics such as the number of anaphors with more than one candidate, and more generally, the average number of candidates per anaphor, or statistics showing the average distance between the anaphors and their antecedents, would be more indicative of how 'easy' or 'difficult' the evaluation data is and should be provided in addition to the information on the numbers or types of anaphors occurring in the evaluation data. Barbu and Mitkov (2001) as well as Tanev and Mitkov (forthcoming) include in their evaluation data information about the average number of candidates per anaphoric pronoun (computed to be as high as 12.9 for English) and information about the average distance from the pronoun to its antecedents in terms of sentences, clauses or intervening NPs.

The next section addresses the problem of comparing the evaluation results in anaphora resolution by postulating that comparison on the same data is not sufficient; what also matters is comparison on the basis of the same preprocessing tools.

8.6 Evaluation workbench for anaphora resolution

In order to secure a 'fair', consistent and accurate evaluation environment, and to address some of the problems identified above, I proposed the development

of an **evaluation workbench for anaphora resolution** (Mitkov 2000) which allows the comparison of anaphora resolution approaches sharing common principles or similar pre-processing (e.g. POS tagger, NP extractor, parser). The workbench enables the 'plugging in' and testing of anaphora resolution algorithms on the basis of the same pre-processing tools and data. This development is a time-consuming project, given that most of the algorithms may have to be re-implemented, but it is expected to produce a better picture as to the advantages and disadvantages of the different approaches. Developing one's own evaluation environment (and even re-implementing some of the 'benchmark' algorithms) also alleviates the formidable difficulties associated with obtaining the code of the original programs. Another advantage of the evaluation workbench can be seen in the fact that all incorporated approaches operate in a fully automatic mode.

While the workbench is an open-ended architecture which allows the inclusion of new algorithms, three approaches extensively cited in the literature were first selected for comparative evaluation: Kennedy and Boguraev's parser-free version of Lappin and Leass's RAP (Kennedy and Boguraev 1996; see also Chapter 5, section 5.4), Baldwin's pronoun resolution method CogNIAC which uses limited knowledge (Baldwin 1997; see also Chapter 5, section 5.5) and Mitkov's knowledge-poor pronoun resolution approach (Mitkov 1998b; see also Chapter 7). All three of these algorithms share a similar pre-processing methodology: they do not rely on a parser to process the input and instead use POS taggers and NP extractors; none of the methods makes use of semantic or real-world knowledge. Kennedy and Boguraev's and Baldwin's algorithms were re-implemented, and the standard, non-optimised version of MARS was used to represent Mitkov's algorithm. Since the original version of CogNIAC is non-robust and resolves only anaphors that obey certain rules, for fairer and comparable results the 'resolve-all' version as described by Baldwin (1997) was implemented.

The current version of the evaluation workbench[19] employs, similarly to MARS, one of the best performing 'super-taggers' in English – Conexor's FDG Parser (Tapanainen and Järvinen 1997). This shallow parser provides information on the dependency relations between words which allows the extraction of complex NPs. It also gives morphological information and the syntactic roles of words. Although the FDG Parser does not identify the noun phrases in the text, the dependencies established between words have served in the implementation of a noun phrase extractor.

The algorithms receive a list of all NPs in the text and build their own list of candidates for antecedents. In the case of Mitkov's approach, candidates from the current and previous three sentences are considered, Baldwin's CogNIAC looks at candidates from the current paragraph, whereas Kennedy and Boguraev processes candidates from the whole text. The lists are generated by running an XML parser over the file resulting from the noun phrase extractor and selecting only the nominal anaphors. Each entry includes information on the word form, the lemma of the word or of the head of the noun phrase, the starting position in the text, the ending position in the text, the part of speech, the

Table 8.2 Comparative evaluation carried out with the help of the evaluation workbench

File	Number of pronouns	Success rate		
		MARS	CogNIAC	Kennedy and Boguraev
PSW	77	79.74	72.1	79.8
MAC	148	66.06	60.8	67.1
WIN	51	56.86	55.9	58.7
BEO	67	45.16	45.0	46.3
CDR	83	64.83	61.3	66.3
Total	426	62.53	59.02	63.64

grammatical function, the index of the sentence that contains the candidate and the index of the verb whose argument is the candidate.

The workbench incorporates an automatic scoring system operating on an XML input file where the correct antecedent for every anaphor has been marked.[20] The results are visually displayed on the screen and can also be saved on file. For easier visual comparison, each anaphor is displayed together with the antecedents proposed by each of the algorithms.

The comparative evaluation was based on a corpus of technical texts manually annotated for coreference. The corpus contains more than 50 000 words, with 19 305 noun phrases and 484 anaphoric pronouns. Files used were: 'Beowulf HOW TO' (referred to in Table 8.2 as BEO), 'Linux CD-Rom HOW TO' (CDR), 'Macintosh Help file' (MAC), 'Portable StyleWriter Help File' (PSW) and 'Windows Help file' (WIN).

Table 8.2 shows the success rates of the three anaphora resolution algorithms on the above files. The overall success rate calculated for the 426 anaphoric pronouns found in the texts was 62.5% for MARS, 59.02% for CogNIAC and 63.64% for Kennedy and Boguraev's method.

Besides the evaluation system, the workbench also incorporates a basic statistical calculator of the anaphoric occurrences in the input file. The parameters calculated are: the total number of anaphors, the number of anaphors in each morphological category (personal pronoun, noun, reflexive, possessive), the number of inter- and intrasentential anaphors, average number of candidates per anaphor and average distance from the anaphors to their antecedents. More details on the implementation of the evaluation workbench and on the evaluation results are reported in Barbu and Mitkov (2000, 2001).

While the workbench is based on the FDG shallow parser at the moment,[21] the environment is being updated in such a way that two different modes will be available: one making use of a shallow parser (for approaches operating on partial analysis) and one employing a full parser (for algorithms making use of full analysis). Future versions of the workbench will include access to semantic information through WordNet in order to accommodate approaches incorporating knowledge of this type.

8.7 Other proposals

Donna Byron (2001) maintains that a number of additional kinds of information should be included in the evaluation in order to make the performance of algorithms for pronoun resolution more transparent. To start with, in addition to recall and precision, Byron puts forward the measure *resolution rate* defined as follows:

$$\text{Resolution rate} = \frac{\text{Number of pronouns resolved correct}}{\text{Number of referential pronouns}}$$

Referential pronouns are those that refer anaphorically, cataphorically or deictically; therefore the set of referential pronouns includes cataphoric and deictic[22] pronouns but excludes the pronouns defined as pleonastic in Chapter 1, section 1.4.1. Therefore, the resolution rate goes beyond the scope of anaphoric pronouns, since it includes pronominal instances of deixis and cataphora; at the same time it does not account for cases where unnecessary resolution of pleonastic pronouns is attempted.

Byron is concerned that most pronoun resolution studies do not detail exactly what types of pronouns (e.g. personal, reflexive, gendered, singular, etc.) they resolve. Therefore, she proposes that the *pronoun coverage* be explicitly reported. Next, she would like to see more information on which types of pronouns have been *excluded* from a specific experiment. Byron explains that it has been common to exclude (i) set constructions which are required to interpret pronouns with a split antecedent ('Pat went to Kim's apartment and they went dancing'), quoted speech ('Mr. Van Dyke described the incident saying "The guy ran right out in front of me" ') or cataphora, (ii) pronouns with no antecedents in the discourse such as deictic and generic pronouns, (iii) pronouns which have antecedents different from NPs such as clauses, or pronouns representing examples of indirect anaphora,[23] and (iv) pronouns excluded due to idiosyncratic reasons imposed by the domain/corpus. In addition to making explicit the pronoun coverage and exclusion categories, Byron suggests that all evaluations of pronoun resolution methods should provide details on the evaluation corpus and on the evaluation set size, and report not only recall/precision but also resolution rate. She proposes that this information be presented in a concise and compact format (table) called standard disclosure (Byron 2001).

Stuckardt (2001) argues that evaluation of anaphora resolution systems should take into account several factors beyond simple accuracy of resolution. In particular, both developer-oriented (e.g. related to the selection of optimal resolution factors) and application-oriented (e.g. related to the requirement of the application, as in the case of Information Extraction where a proper name antecedent is needed) evaluation metrics should be considered.[24] In fact I can argue further that a way to measure the usefulness of an anaphora resolution system is to see by how much it could enhance the performance of a specific NLP application.

Bagga (1998) proposes a methodology for evaluation of coreference resolution systems which can be directly transferred to anaphora resolution. He classifies coreference according to the processing required for resolution, and proposes

that evaluation be carried out separately for each of the following classes (listed in ascending order of processing): appositives, predicate nominals, proper names, pronouns, quoted speech pronouns, demonstratives, exact matches, substring matches, identical lexical heads, synonyms and anaphors that require external world knowledge for their resolution.

Several other papers have addressed the evaluation of coreference resolution systems such as Vilain et al. (1995), Bagga and Baldwin (1998a), Popescu-Belis and Robba (1998) and Trouilleux et al. (2000). The reader is referred to the original publications for more details.

8.8 Summary

This chapter has argued that evaluation of anaphora resolution algorithms and anaphora resolution systems should be done separately and has thus proposed a package of evaluation measures and tasks. Fair comparison of methods requires that evaluation be done not only on the same data, but also on the basis of the same pre-processing tools. A special evaluation environment has been developed to carry out this type of comparative evaluation.

Notes

1 This chapter focuses on evaluation in anaphora resolution. As pointed out in Chapter 5, section 5.10, the evaluation for anaphora resolution is not the same as that for coreference resolution.

2 For further discussion on automatic anaphora resolution (as opposed to non-automatic), see Chapter 7, section 7.4.1.

3 When evaluating his pronoun resolution approach, Baldwin (1997) defines recall as the 'number of correctly resolved anaphors' divided by 'the number of instances of coreference' which I understand as divided by 'the number of all anaphors' given the class of anaphora he tackles (see Chapter 5, section 5.5, for more on his pronoun resolution algorithm). This definition is in line with that used by Gaizauskas and Humphreys (1996). In fact, Gaizauskas and Humphreys (1996) define recall as 'a measure of how much of what a system is supposed to find, is actually found'. This translates into the formula of recall above as part of Definition 2. The definition of precision that follows corresponds to the original formulation of this measure by Baldwin (1997) and is equivalent to the ones used by Gaizauskas and Humphreys (1996).

4 Baldwin does not specify if 'number of instances of coreference', which I understand as 'all anaphors', refers to all anaphors as marked by humans, or if it refers to all anaphors as identified by the program. I have taken the first interpretation as the more probable.

5 As marked by humans. See, however, the comments in section 9.2.4 on the danger of the evaluation figures' being compromised if human evaluation is not reliable.

6 On the other hand the success rate of an anaphora resolution system reflects the performance of the whole system; in this case the text to be processed is not normally expected to be analysed by humans.

7 By card(A) is meant the cardinality of the set A, i.e. the number of elements that this set contains.

8 Note that these relations hold in an 'ideal environment', when the input to the anaphora resolution algorithm has been correctly analysed. For different outcomes in the evaluation of anaphora resolution systems see section 8.4.
9 Factor-based algorithms typically employ a number of factors after gender and number checks. Factors can be preferences or constraints.
10 See Chapter 5, section 5.10.
11 For preference-based approaches where the preference is expressed numerically.
12 In other words, decision power serves as a measure for the 'confidence' of each indicator.
13 Similar techniques have been used by Lappin and Leass (1994).
14 This finding has been independently confirmed by previous studies (Lappin and Leass 1994); Tanev and Mitkov (2000) establish that this factor has the highest relative importance value (5.2%) for Bulgarian too.
15 The definitions of recall and precision do not capture these cases either.
16 In Mitkov et al. (2002) this measure is computed only on the basis of pronouns (including anaphoric and non-anaphoric).
17 This would bring us closer to the assumption that the samples are independent in that anaphors which are far from each other are less likely to be correlated in terms of referential or, more generally, linguistic features and that the resolution of one anaphor is not influenced by that of another (e.g. the resolution of anaphors is not facilitated by common local clues and is not affected by carrying over errors).
18 The critical success rate addresses this issue to a certain extent in the evaluation of anaphora resolution algorithms by providing the success rate for the anaphors that are more difficult to resolve.
19 Implemented by Catalina Barbu.
20 The annotation scheme currently recognised by the system is MUC, but support for the MATE annotation scheme is being developed.
21 It is worth mentioning that experiments with different pre-processing tools are under way with a view to achieving optimal performance. At the time of completing the book, the Xerox language processing suite was being incorporated as an alternative pre-processing option.
22 Termed 'exophoric' in the paper. Following Halliday and Hasan (1976), Byron distinguishes between exophora and deixis.
23 For reasons of consistency, some of the original terms used by Byron have been replaced with equivalent terms as introduced in Chapter 1.
24 See Mitkov's (2001b) work (summarised in section 8.3.3) related to the metrics addressing the anaphora resolution factors.

CHAPTER NINE

Outstanding issues

9.1 Anaphora resolution: where do we stand now?

After considerable initial research, followed by years of relative silence in the early 1980s, anaphora resolution has attracted the attention of many researchers in the last ten years and a great deal of successful work on the topic has been produced. Discourse-oriented theories and formalisms such as DRT and centering inspired new research on the computational treatment of anaphora. The drive towards corpus-based, robust NLP solutions further stimulated interest for alternative and/or data-enriched approaches. Last, but not least, application-driven research in areas such as automatic abstracting and information extraction independently identified the importance of (and boosted the research in) anaphora and coreference resolution.

Much of the earlier work in anaphora resolution heavily exploited domain and linguistic knowledge (Carbonell and Brown 1988; Carter 1987a; Rich and LuperFoy 1988; Sidner 1979) which was difficult both to represent and to process, and required considerable human input. However, the pressing need for the development of robust and inexpensive solutions to meet the demands of practical NLP systems encouraged many researchers to move away from extensive domain and linguistic knowledge and to embark instead upon knowledge-poor anaphora resolution strategies. A number of proposals in the 1990s deliberately limited the extent to which they rely on domain and/or linguistic knowledge (Baldwin 1997; Dagan and Itai 1990, 1991; Kennedy and Boguraev 1996; Mitkov 1996, 1998b; Nasukawa 1994) and reported promising results in knowledge-poor operational environments.

The drive towards knowledge-poor and robust approaches was further motivated by the emergence of cheaper and more reliable corpus-based NLP tools such as POS taggers and shallow parsers, alongside the increasing availability of corpora and lexical resources such as WordNet. In fact, the availability of corpora, both raw and annotated with coreferential links, provided a strong impetus to anaphora resolution with regard to both training and evaluation. Corpora, especially when annotated, are an invaluable resource not only for empirical research but also for automated learning methods (e.g. machine learning methods) and for the evaluation of implemented approaches. From simple

co-occurrence rules (Dagan and Itai 1990) through training decision trees to identify anaphor–antecedent pairs (Aone and Bennett 1995) to genetic algorithms which optimise the resolution factors (Orasan et al. 2000), the successful performance of more and more approaches was made possible through the availability of suitable corpora.

Even though the last ten years have seen considerable advances in the field of anaphora resolution, there are still a number of outstanding issues that either remain unsolved or need further attention and, as a consequence, represent major challenges to the further development of the field. A fundamental question that requires further investigation is how far the performance of anaphora resolution algorithms can go and what are the limitations of knowledge-poor methods. In particular, more research should be carried out into the factors influencing the performance of these algorithms. Another significant problem for automatic anaphora resolution systems is that the accuracy of the pre-processing is still too low and, as a result, the performance of such systems remains far from ideal. As a further consequence, only a few anaphora resolution systems operate in fully automatic mode: most of them rely on manual pre-processing or use pre-analysed corpora. One of the impediments for the evaluation or for the fuller employment of machine learning (ML) techniques is the lack of widely available corpora annotated for anaphoric or coreferential links. More work towards the proposal of consistent and comprehensive evaluation is necessary; so, too, is work in the multilingual context. The remaining part of this chapter will briefly discuss issues in need of further attention.

9.2 Issues for continuing research

9.2.1 The limits of anaphora resolution

A fundamental question that needs further research concerns the limits of anaphora resolution algorithms and the trade-off between low effort and high performance. Methods that heavily exploit knowledge of syntax, semantics and local focusing and which resort to a limited amount of world or domain knowledge, such as Carter's algorithm (1986, 1987a), have been reported to achieve success rates of up to 93%. On the other hand, knowledge-poorer methods such as Lappin and Leass's (1994), Baldwin's (1997), Mitkov's (1998b) and Ge et al.'s (1998) algorithms have scored in the high 80s on specific evaluation sets. In spite of the comparatively high results obtained by some knowledge-poor algorithms and the claim that certain types of anaphor can be successfully resolved without real-world knowledge,[1] the lack of semantic, domain or real-world knowledge imposes serious limitations. Recent research (Palomar et al. 2001b) suggests that knowledge-poor approaches have their limits and in order to achieve a success rate of 75% or more, some semantic knowledge in the form of selectional restrictions, e.g. derivable from an ontology such as WordNet, is essential. Another especially important factor required for the disambiguation of a great number of anaphors, as pointed out in Chapter 2 (section 2.2.3.4), is real-world knowledge.

In fact this privileged factor shows even more clearly the limitations of automatic anaphora resolution given the unrealistic task of representing and acquiring real-world or common-sense knowledge.

The limits of anaphora resolution algorithms most certainly differ from language to language. Whereas in English often semantic or real-world knowledge is needed to correctly interpret anaphors referring to objects and events, in Romance languages and especially in Slavonic languages[2] a number of cases can be safely resolved on the basis of gender agreement alone and therefore the necessity for real-world knowledge in these languages may not be so severely felt.

In light of these limitations and with a view to improving performance, more research is needed into the factors on which anaphora resolution algorithms are based. One basic question concerns the set of core or optimal factors that have to be present in every anaphora resolution algorithm. Other outstanding issues include investigation into the impact, genre- and language-specificity and mutual dependency of the factors. For a brief related discussion, see Chapter 2, section 2.2.3.4, and Mitkov (1997a).

9.2.2 Pre-processing and fully automatic anaphora resolution

A real-world anaphora resolution system vitally depends on the efficiency of the pre-processing tools which analyse the input before feeding it to the resolution algorithm. Inaccurate pre-processing could lead to a considerable drop in the performance of the system, however accurate an anaphora resolution algorithm may be.[3] The accuracy of today's pre-processing is still unsatisfactory from the point of view of anaphora resolution. Whereas POS taggers are fairly reliable, full or partial parsers are not. Named entity recognition is still a challenge, requiring extensions such as product name recognition, which is vital for a number of genres. While some recent progress in areas such as identification of pleonastic pronouns and, in general, instances of non-nominal anaphora (Cardie and Wagstaff 1999; Evans 2000), identification of non-anaphoric definite descriptions (Bean and Riloff 1999; Vieira and Poesio 2000b) and recognition of animacy (Evans and Orasan 2000) and gender (Orasan and Evans 2001) has been reported, these tasks and other important pre-processing tasks, such as term recognition, have a long way to go. For instance, the best accuracy reported in robust parsing of unrestricted texts is around the 87% mark (Collins 1997); the accuracy of identification of non-nominal pronouns does not normally exceed 80% (Evans 2000, 2001).[4] Other tasks may be more accurate but still far from perfect. The state of the art of NP chunking, which does not include NPs with post-modifiers, is 90–93% in terms of recall and precision.[5] The best-performing named entity taggers achieve an accuracy of about 96% when trained and tested on news about a specific topic, and about 93% when trained on news about a topic and tested on news about another topic (Grishman 2002).

Another point worth noting is that whereas *'standard' pre-processing programs* such as part-of-speech taggers, shallow parsers, full parsers, etc., are being constantly developed and improved (and yet there could be formidable problems in

getting hold of such software from the public domain!), anaphora resolution *task-specific pre-processing tools*, such as programs for identifying non-anaphoric pronouns or definite NPs, or programs for animacy or gender recognition, have received considerably less attention.

As a result of the above limitations, the *majority of anaphora resolution systems do not operate in fully automatic mode*. In fact, research in anaphora resolution has so far suffered from a bizarre anomaly in that, until recently, hardly any fully automatic operational systems had been reported: almost all described approaches relied on some kind of pre-editing of the text that was fed to the anaphora resolution algorithm[6]; some of the methods were only manually simulated (for a brief discussion on the topic of automatic anaphora resolution see Chapter 7, section 7.4.1). Recent reports (Soon et al. 1999; Barbu and Mitkov 2000; Fukumoto et al. 2000; Orasan et al. 2000; Palomar et al. 2000; Tanev and Mitkov 2000) show that more and more researchers attempt fully automatic resolution albeit at the expense of lower success rate. In fact the evaluations suggest higher success rates for Spanish (76%)[7] and Bulgarian (74%)[8] than for English[9]: one possible explanation for these results is that these two languages are much more gender-discriminative than English and a considerable number of anaphors are resolved after applying gender constraints only.

9.2.3 The need for annotated corpora

Corpora annotated with anaphoric or coreferential links are still a rare commodity, and those that do exist are not of a large size (Chapter 6, section 6.2), despite being much needed for different methods in anaphora/coreference resolution systems. Corpora of this kind have been used in the training of machine learning algorithms (Aone and Bennett 1995) and statistical approaches to anaphora resolution (Ge et al. 1998).[10] In other cases, they were used for optimisation of existing approaches (Orasan et al. 2000) and their evaluation (Mitkov et al. 1999). The automatic training and evaluation of anaphora resolution approaches require that the annotation cover anaphoric or coreferential chains and not just single anaphor–antecedent pairs. This is because the resolution of a specific anaphor would be considered successful if any preceding non-pronominal element of the anaphoric chain associated with that anaphor were identified.

The need for *annotated corpora* is an outstanding issue which brings about additional issues. The act of annotating corpora follows a specific *annotation scheme*, an adopted methodology as to how to encode linguistic features in a text. The annotation scheme ideally has to deliver wide coverage and should be clear and simple to use. It appears, however, that wide coverage and reliable markup are not compatible desiderata. Once an annotation scheme has been proposed to encode linguistic information, *annotation tools* have to be developed to apply this scheme to corpus texts, making the annotating process faster and more user-friendly. In addition, the process of annotation will be more efficient if a specific *annotation strategy* is employed. The annotation of corpora at anaphoric or coreferential level suffers from the lack of sufficient inter-annotator agreement (Chapter 6, section 6.5) and therefore the development of a good annotation

strategy, which can be a crucial factor for improving the agreement, is an issue that urgently requires further attention. In fact, erroneous annotation can be especially detrimental in that anaphors correctly resolved by the program could be returned as incorrect attempts when matched against wrongly annotated data. Finally, while a well thought-out annotation strategy is a key prerequisite for better agreement, additional efforts are needed to further improve the other two components of the annotation process: the annotating scheme and the annotating tool.

9.2.4 Other outstanding issues

In spite of the recent progress in *evaluation in anaphora resolution*, it is felt that the proposals still fall short of providing a comprehensive and clear picture of this task. There are still a number of outstanding issues related to the reliability of the evaluation results that need further attention. The question as to how reliable or realistic the obtained performance figures are depends largely on the nature of the data used for evaluation. Evaluation results should be less *relative* with regard to a specific evaluation dataset or with regard to another approach, and should be more *absolute*. Therefore, in addition to the evaluation results, information (and measures) should be provided as to how difficult the anaphors are to resolve in the evaluation data. For further discussion of that issue, see Mitkov (2001b) or Chapter 8 of this book. Finally, in a recent presentation (Mitkov 2001c), I drew attention to several 'traps' that we may find ourselves caught in when analysing evaluation. Evaluation results could be compromised if the annotated corpora contain incorrect markups due to human errors, if the original documents feature spelling errors and/or ungrammaticalities, or if the results are not reported in a 'transparent' or 'honest' manner.

An issue which merits further attention and emerges from the *multilingual context* of recent NLP work as a whole is the development of multilingual anaphora resolution systems. Whereas initial work on multilingual anaphora resolution has been based on extension of original approaches (usually in English) to other languages (Azzam et al. 1998a; Mitkov and Stys 1997; Mitkov et al. 1998; Tanev and Mitkov 2000), one of the truly *multilingual* challenges is to exploit multilingual tools and resources for enhancing the efficiency of anaphora resolution. Good first examples of such work (albeit preliminary) are the projects reported by Harabagiu and Maiorano (2000) and Mitkov and Barbu (2000).[11]

Another outstanding issue is the *coverage of varieties of anaphora*. The majority of projects focus on pronoun resolution; a good number of projects also address resolution of definite descriptions (some including indirect anaphora) or zero pronouns. However, the resolution of non-nominal anaphora, which is arguably more difficult to handle, is almost completely ignored.

Finally, the work on anaphora resolution should provide a suitable *service to the research community*. While papers can be easily obtained[12] and web demos are beginning to emerge,[13] more has to be done in the way of aiding researchers working in this field; experience, software and data produced should be readily shared. By way of example, against the background of scarce annotated data, it

would be particularly important for the existing resources to be shared by the anaphora community. Anaphora resolution programs should be freely available for testing and for integration into larger NLP systems. I believe that positive recent developments[14] are a step in the right direction.

Notes

1 Baldwin (1997).
2 Slavonic languages have a three-gender system: masculine, female and neuter.
3 Mitkov et al. (2002) show that the success rate of the fully automatic version of Mitkov's approach (MARS) is up to 25% lower than the version which uses post-edited entries to the algorithm, even though MARS used one of the most efficient shallow parsers for English (FDG).
4 However, Paice and Husk (1987) reported 92% for identification of strictly pleonastic *it* in a narrow domain.
5 Personal communication with John Carroll.
6 Apart from the coreference resolution systems implemented for MUC-6 and MUC-7.
7 See Palomar et al. (2000).
8 See Tanev and Mitkov (2000) and Chapter 7, section 7.5.1.
9 See the evaluation of MARS, Chapter 7, section 7.4.4.
10 In fact Ge et al. (1998) make use of a completely parsed corpus which in addition to coreferential links, has parts of speech, noun phrases, sentences, etc., marked.
11 See also Chapter 7, section 7.3.
12 A preliminary list of downloadable papers is now available at http://www.wlv.ac.uk/~le1825/download.htm (the list is updated on a regular basis).
13 See http://www.wlv.ac.uk/sles/compling.
14 The Research Group in Computational Linguistics at the University of Wolverhampton has recently offered its corpora annotated for coreferential links as well as its programs for anaphora resolvers free to the research community (see http://www.wlv.ac.uk/sles/compling).

References

Abraços, J. and Lopes, J.G. (1994) 'Extending DRT with a focusing mechanism for pronominal anaphora and ellipsis resolution'. *Proceedings of the 15th International Conference on Computational Linguistics (COLING'94)*, 1128–1132. Kyoto, Japan.

Al-Kofani, K., Grom, B. and Jackson, P. (1999) 'Anaphora resolution in the extraction of treatment history language from court opinions by partial parsing'. *Proceedings of the 17th International Conference on Artificial Intelligence and Law*, 138–146. Oslo, Norway.

Allen, J. (1995) *Natural language understanding*. The Benjamin/Cummings Publishing Company Inc.

Alshawi, H. (1987). *Memory and context for language interpretation*. Cambridge: Cambridge University Press.

Alshawi, H. (Ed.) (1992) *The core language engine*. Cambridge, MA: MIT Press.

Angelova, G., Kalaydjiev, O. and Hahn, W.v. (1998) 'The gain of failures: using side-effects of anaphora resolution for term consistency checks'. *Proceedings of the 8th International Conference 'Artificial Intelligence: Methodology, Systems, Applications' (AIMSA-98)*, 1–13. Sozopol, Bulgaria. Springer Lecture Notes in Artificial Intelligence, Vol. 1480.

Aone, C. and McKee, D. (1993) 'A language-independent anaphora resolution system for understanding multilingual texts'. *Proceedings of the 31st Annual Meeting of the ACL (ACL'93)*, 156–163. Columbus, OH, USA.

Aone, C. and Bennett, S. (1994) 'Discourse tagging tool and discourse-tagged multilingual corpora'. *Proceedings of the International Workshop on Sharable Natural Language Resources (SNLR)*, 71–77. Nara, Japan.

Aone, C. and Bennett, S. (1995) 'Evaluating automated and manual acquisition of anaphora resolution strategies'. *Proceedings of the 33rd Annual Meeting of the ACL (ACL'95*), 122–129. Cambridge, MA, USA.

Aone, C. and Bennett, S. (1996) 'Applying machine learning to anaphora resolution'. In Wermter, S., Riloff, E. and Scheler, G. (Eds.) *Connectionist, statistical and symbolic approaches to learning for Natural Language Processing*, 302–314. Berlin: Springer.

Ariel, M. (1990) *Accessing noun phrase antecedents*. London: Routledge.

Asher, N. (1987) 'A typology for attitude verbs and their anaphoric properties'. *Linguistics and Philosophy*, 10, 125–197.

Asher, N. (1993) *Reference to abstracts objects in discourse*. Dodrecht: Kluwer Academic.

Asher, N. and Wada, H. (1988) 'A computational account of syntactic, semantic and discourse principles for anaphora resolution'. *Journal of Semantics*, 6, 309–344.

Azzam, S., Humphreys, K. and Gaizauskas, R. (1998a) 'Coreference resolution in a multilingual information extraction'. *Proceedings of the Workshop on Linguistic Coreference*. Granada, Spain.

Azzam, S., Humphreys, K. and Gaizauskas, R. (1998b) 'Evaluating a focused-based approach to anaphora resolution'. *Proceedings of the 17th International Conference on Computational Linguistics (COLING'98/ACL'98)*, 74–78. Montreal, Canada.

Azzam, S., Humphreys, K. and Gaizauskas, R. (1999) 'Using coreference chains for text summarisation'. *Proceedings of the ACL'99 Workshop on Coreference and its Applications*, 77–84. College Park, Maryland, USA.

Bagga, A. (1998) 'Evaluation of coreferences and coreference resolution systems'. *Proceedings of the First International Conference on Language Resources and Evaluation (LREC'98)*, Vol. 1, 563–566. Granada, Spain.

Bagga, A. and Baldwin, B. (1998a) 'Algorithms for scoring coreference chains'. *Proceedings of the Workshop on Linguistic Coreference*. Granada, Spain.

Bagga, A. and Baldwin, B. (1998b) 'Entity-based cross-document coreferencing using the vector space model'. *Proceedings of the 18th International Conference on Computational Linguistics (COLING'98/ACL'98)*, 79–85. Montreal, Canada.

Baldwin, B. (1997) 'CogNIAC: high precision coreference with limited knowledge and linguistic resources'. *Proceedings of the ACL'97/EACL'97 workshop on Operational Factors in Practical, Robust Anaphora Resolution*, 38–45. Madrid, Spain.

Baldwin, B. and Morton, T. (1998) 'Dynamic coreference-based summarization'. *Proceedings of the Third International Conference on Empirical Methods in Natural Language Processing (EMNLP–3)*, 1–6. Granada, Spain.

Baldwin, B., Reynar, J., Collins, M., Eisner, J., Ratnaparki, A., Rosenzweig, J., Sarkar, A. and Bangalore, S. (1995) 'Description of the University of Pennsylvania system used for MUC-6'. *Proceedings of the Sixth Message Understanding Conference (MUC-6)*, 177–191. Columbia, Maryland, USA.

Barbu, C. (2000) 'FAST – towards a semi-automatic annotation of corpora'. *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC)*, 501–505. Athens, Greece.

Barbu, C. (2001) 'Automatic learning and resolution of anaphora'. *Proceedings of the International Conference 'Recent Advances in Natural Language Processing' (RANLP'2001)*, 22–27. Tzigov Chark, Bulgaria.

Barbu, C. and Mitkov, R. (2000) 'Evaluation environment for anaphora resolution'. *Proceedings of the International Conference on Machine Translation and Multilingual Applications (MT2000)*, 18-1–18-8. Exeter, UK.

Barbu, C. and Mitkov, R. (2001) 'Evaluation tool for rule-based anaphora resolution methods'. *Proceedings of the 39th Annual Meeting and 10th Conference of the European Chapter of the Association for Computational Linguistics (ACL'2001)*, 34–41. Toulouse, France.

Barlow, M. (1998) 'Feature mismatches and anaphora resolution'. *Proceedings of the Discourse Anaphora and Anaphora Resolution Colloquium (DAARC'2)*, 34–41. Lancaster, UK.

Bean, D. and Riloff, E. (1999) 'Corpus-based identification of non-anaphoric noun phrases'. *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, 373–380. College Park, Maryland, USA.

Biber, D., Conrad, S. and Reppen, R. (1998) *Corpus linguistics: investigating language structure and use*. Cambridge: Cambridge University Press.

Bird, S., Day, D., Garofolo, J., Henderson, J., Laprun, C. and Liberman, M. (2000) 'ATLAS: a flexible and extensible architecture for linguistic annotation'. *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC)*, 1699–1706. Athens, Greece.

Bobrow, D.G. (1964) 'A question-answering system for high school algebra word problems'. *AFIPS Conference Proceedings*, 26, 591–614.

Boguraev, B. (1979) *Automatic resolution of linguistic ambiguities*. TR-11, University of Cambridge Computer Laboratory, Cambridge.

Boguraev, B. and Kennedy, C. (1997) 'Salience-based content characterisation of documents'. *Proceedings of the ACL'97/EACL'97 Workshop on Intelligent Scalable Text Summarisation*, 3–9. Madrid, Spain.

Bolinger, D. (1977) *Pronouns and repeated nouns*. Bloomington, Indiana: Indiana University Linguistics Club.

Botley, S. (1999) *Corpora and discourse anaphora: using corpus evidence to test theoretical claims*. PhD thesis, University of Lancaster, UK.

Botley, S. and McEnery, A. (1991) 'A graphical representation scheme for anaphoric links in natural language texts'. *Proceedings of the 13th Colloquium of the British Computer Society Information Retrieval Specialist Group*, 127–140. Huddersfield, UK.

Breck, E., Burger, J., Ferro, L., House, D., Light, M. and Mani, I. (1999) 'A sys called Qanda'. *Proceedings of the 8th Text Retrieval Conference (TREC-8)*, 499–505. Gaithersburg, Maryland, USA.

Brennan, S., Friedman, M. and Pollard., C. (1987) 'A centering approach to pronouns'. *Proceedings of the 25th Annual Meeting of the ACL (ACL'87)*, 155–162. Stanford, CA, USA.

Brown, G. and Yule, G. (1983) *Discourse analysis*. Cambridge: Cambridge University Press.

Bruneseaux, F. and Romary, L. (1997) 'Codage des références et coréférences dans les dialogues homme-machine'. *Proceedings of the Joint International Conference of the Association for Computers and the Humanities and the Association for Literary and Linguistic Computing (ACH-ALLC'97)*, 15–17. Ontario, Canada.

Byron, D. (2001) 'The uncommon denominator'. *Computational Linguistics*, 27 (4), 569–577.

Canning, Y., Tait, J., Archibald, J. and Crawley, R. (2000) 'Replacing anaphora for readers with acquired dyslexia'. *Proceedings of the Discourse, Anaphora and Reference Resolution Conference (DAARC2000)*, 49–58. Lancaster, UK.

Carbonell, J.G. (1980) 'Towards a process model of human personality traits'. *Artificial Intelligence*, 15 (1, 2), 49–74.

Carbonell, J.G. and Brown, R. (1988) 'Anaphora resolution: a multi-strategy approach'. *Proceedings of the 12th International Conference on Computational Linguistics (COLING'88)*, Vol. I, 96–101. Budapest, Hungary.

Carden, G. (1982) 'Backwards anaphora in discourse context'. *Journal of Linguistics*, 18, 361–387.

Cardie, C. and Pierce, D. (1998) 'Error-driven pruning of treebank grammars for base noun phrase identification'. *Proceedings of the 36th Annual Meeting of the ACL and COLING-98*, 218–224. Montreal, Canada.

Cardie, C. and Wagstaff, K. (1999) 'Noun phrase coreference as clustering'. *Proceedings of the 1999 Joint SIGDAT Conference on Empirical Methods in NLP and Very Large Corpora*, 82–89. University of Maryland, USA.

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

Carter, D. (1986) *A shallow processing approach to anaphor resolution*. PhD thesis, University of Cambridge.

Carter, D. (1987a) *Interpreting anaphora in natural language texts*. Chichester: Ellis Horwood.

Carter, D. (1987b) 'Common sense inference in a focus-guided anaphor resolver'. *Journal of Semantics*, 4, 237–246.

Carter, D. (1990) 'Control issues in anaphor resolution'. *Journal of Semantics*, 7, 435–454.

Carvalho, A. (1996) 'Logic grammars and pronominal anaphora'. *Proceedings of the 'Discourse Anaphora and Anaphor Resolution' Conference (DAARC'96)*, 106–122. Lancaster, UK.

Charniak, E. (1972) *Toward a model of children's story comprehension*. AI TR-266, Massachusetts Institute of Technology Artificial Intelligence Laboratory.

Charniak, E. (2001) 'Unsupervised learning of name structure from coreference data'. *Proceedings of the North American Association for Computational Linguistics (NAACL2001)*, 48–54. New Brunswick, NJ, USA.

Chen, H.H. (1992) 'The transfer of anaphors in translation'. *Literary and Linguistic Computing*, 7 (4), 231–238.

Chomsky, N. (1981) *Lectures on government and binding*. Foris, Dodrecht.

Chomsky, N. (1995) *The minimalist program*. MIT Press.

Cohen, P. (1984) 'The pragmatics of referring and the modality of communication'. *Computational Linguistics*, 10, 97–146.

Collins, M. (1996) 'A new statistical parser based on biagram lexical dependencies'. *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics (ACL'96)*. Santa Cruz, California.

Collins, M. (1997) 'Three generative, lexicalised models for statistical parsing'. *Proceedings of the 35th Annual Meeting of the ACL (ACL'97)*, 16–23. Madrid, Spain.

Connoly, D., Burger, J. and Day, D. (1994) 'A machine learning approach to anaphoric reference'. *Proceedings of the International Conference 'New Methods in Language Processing' (NeMeLaP-1)*, 255–261. Manchester, UK.

Cormack, S. (1993) *Focus and discourse representation theory*. PhD thesis, Centre for Cognitive Sciences, University of Edinburgh, Edinburgh, UK.

Cormack, S. (1998) 'Incremental pronoun resolution in Discourse Representation Theory'. *Proceedings of the Discourse Anaphora and Anaphora Resolution Colloquium (DAARC'2)*, 82–95. Lancaster, UK.

Cornish, F. (1986) *Anaphoric relations in English and French. A discourse perspective*. London: Croom Helm.

Cornish, F. (1996) ' "Antecedentless" anaphors: deixis, anaphora, or what? Some evidence from English and French'. *Journal of Linguistics*, 32 (1), 19–41.

Cornish, F. (1999) *Anaphora, discourse and understanding. Evidence from English and French*. Oxford: Oxford University Press.

Coulson, M. (1995) 'Anaphoric reference'. In Green, J. and Coulson, M. (Eds.) *Language understanding: current issues*. Buckingham: Open University Press.

Cristea D. and Dima, G. (2000) 'Anaphora and cataphora what's in there?' Paper presented at the 5th TELRI European Seminar 'Corpus Linguistics: How to Extract Meaning from Corpora'. Ljubljana, Slovenia.

Cristea, D., Ide, N. and Lomary, L. (1998) 'Veins theory: a model of global discourse cohesion and coherence'. *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and of the 18th International Conference on Computational Linguistics (COLING'98/ ACL'98)*, 281–285. Montreal, Canada.

Cristea, D., Ide, N., Marcu, D. and Tablan, V. (2000) 'An empirical investigation of the relation between discourse structure and coreference'. *Proceedings of the 19th International Conference on Computational Linguistics (COLING'2000)*, 208–214. Saarbrücken, Germany.

Crowe, J. (1996) 'Shallow techniques for the segmentation of news reports'. *Proceedings of the AISB Workshop on Language Engineering for Contents Analysis and Information Retrieval*. Brighton, UK.

Daelemans, W., Zavarel, J., Slot, K. and Bosch, A. (1999) *TiMBL: Tilburg Memory Based Learner, version 2.0*. Reference guide, ILK technical report 99-01. ILK 99-01, Tilburg University.

Dagan, I. and Itai, A. (1990) 'Automatic processing of large corpora for the resolution of anaphora references'. *Proceedings of the 13th International Conference on Computational Linguistics (COLING'90)*, Vol. III, 1–3. Helsinki, Finland.

Dagan, I. and Itai, A. (1991) 'A statistical filter for resolving pronoun references'. In Feldman, Y.A. and Bruckstein, A. (Eds.) *Artificial intelligence and computer vision*, 125–135. Elsevier Science Publishers (North-Holland).

Dagan, I., Justeson, J., Lappin, L., Leass, H. and Ribak, A. (1995) 'Syntax and lexical statistics in anaphora resolution'. *Applied Artificial Intelligence*, 9, 633–644.

Dahl, D. (1986) 'Focusing and reference resolution in PUNDIT'. *Proceedings of the 5th National Conference on Artificial Intelligence*. Philadelphia.

Dahl, D. and Ball, C. (1990) *Reference resolution in PUNDIT*. Research Report CAIT-SLS-9004. Paoli: Center for Advanced Information Technology.

Dahlbäck, N. (1992) 'Pronoun usage in NLI-dialogues – a wizard of Oz study'. *Papers from the Third Nordic Conference on Text Comprehension in Man and Machine*, 27–42. Linköping, Sweden.

Davies, S., Poesio, M., Bruneseaux, F. and Romary, L. (1998) *Annotating coreference in dialogues: proposal for a scheme for MATE*. First draft. Available at http://www.hcrc.ed.ac.uk/~poesio/MATE/anno_manual.html

Day, D., Aberdeen, J., Hirschman, L., Kozierok, R., Robinson, P. and Vilain, M. (1997) 'Mixed-initiative development of language processing systems'. *Proceedings of the Fifth Conference on Applied Natural Language Processing (ANLP97)*, 153–164. Washington DC, USA.

Day, D., Goldschen, A. and Henderson, J. (2000) 'A framework for cross-document annotation'. *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC)*, 199–203. Athens, Greece.

DeCristofaro, J., Strube, M. and McCoy, K. (1999) 'Building a tool for annotating reference in discourse'. *Proceedings of the ACL'99 Workshop on the Relation of Discourse/Dialogue Structure and Reference*, 54–62. College Park, Maryland, USA.

Denber, M. (1998) *Automatic resolution of anaphora in English*. Internal Report. Eastman Kodak Co.

Di Eugenio, B. (1990) 'Centering theory and the Italian pronominal system'. *Proceedings of the 13th International Conference on Computational Linguistics (COLING'90)*, 270–275. Helsinki, Finland.

Dillon, G. (1977) *Introduction to contemporary linguistic semantics*. New Jersey: Prentice Hall.

Dunker, G. and Umbach, C. (1993) *Verfahren zur Anaphernresolution in KIT-FAST*. Internal Report KIT-28. Technical University of Berlin.

Evans, R. (2000) 'A comparison of rule-based and machine learning methods for identifying non-nominal it'. *Natural Language Processing – NLP2000*. Lecture notes in Artificial Intelligence, 233–242. Springer Verlag.

Evans, R. (2001) 'Applying machine learning toward an automatic classification of it'. *Literary and Linguistic Computing*, 16 (1), 45–57.

Evans, R. and Orasan, C. (2000) 'Improving anaphora resolution by identifying animate entities in texts'. *Proceedings of the Discourse, Anaphora and Reference Resolution Conference (DAARC2000)*, 154–162. Lancaster, UK.

Fellbaum, C. (Ed.) (1998) *WordNet: an electronic lexical database*. London: The MIT Press.

Ferrández, A. and Peral, S. (2000) 'A computational approach to zero-pronouns in Spanish'. *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics* (ACL'2000), 166–172. Hong Kong.

Ferrández, A., Palomar, M. and Moreno, L. (1997) 'Slot unification grammar and anaphora resolution'. *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP'97)*, 294–299. Tzigov Chark, Bulgaria.

Ferrández, A., Palomar, M. and Moreno, L. (1998) 'Anaphora resolution in unrestricted texts with partial parsing'. *Proceedings of the 17th International Conference on Computational Linguistics (COLING'98/ACL'98)*, 385–391. Montreal, Canada.

Ferrández, A., Palomar, M. and Moreno, L. (1999) 'An empirical approach to Spanish anaphora resolution'. *Machine Translation*, 14 (3–4), 191–216.

Firbas, J. (1992) *Functional sentence perspective in written and spoken communication*. Cambridge: Cambridge University Press.

Fischer, I., Geistert, B. and Goerz, G. (1996) 'Incremental anaphora resolution in a chart-based semantics construction framework using l-DRT'. *Proceedings of the International Colloquium*

on Discourse Anaphora and Anaphora Resolution. Lancaster (DAARC), 235–244. Lancaster, UK.

Fligelstone, S. (1992) 'Developing a scheme for annotating text to show anaphoric relations'. In Leitner, G. (Ed.) *New directions in English language corpora: methodology, results, software developments*, 153–170. Berlin: Mouton de Gruyter.

Foley, W. and Van Valin, R. (1984) *Functional syntax and universal grammar*. Cambridge: Cambridge University Press.

Fox, B. (1987) *Discourse structure and anaphora*. Cambridge Studies in Linguistics, 48. Cambridge: Cambridge University Press.

Fraurud, K. (1988) 'Pronoun Resolution in unrestricted text'. *Nordic Journal of Linguistics*, 11, 47–68.

Fukumoto, F., Yamada, H. and Mitkov, R. (2000) 'Resolving overt pronouns in Japanese using hierarchical VP structures'. *Proceedings of the Workshop on Corpora and NLP*, 152–157. Monastir, Tunisia.

Gaizauskas, R. and Humphreys, K. (1996) 'Quantitative evaluation of coreference algorithms in an information extraction system'. Paper presented at the DAARC-1 conference, Lancaster, UK. Reprinted in Botley, S. and McEnery, A. (Eds.) (2000) *Corpus-based and computational approaches to discourse anaphora*, 143–167. Amsterdam: John Benjamins.

Galaini, Chikh Mustafa (1992) *Jami'u al-durus al-arabiah* (Arabic lesson collection). Beirut: Manshurat al-maktabah al-asriyah (Modern Library).

Garside, R., Fligelstone, S. and Botley, S. (1997) 'Discourse annotation: anaphoric relations in corpora'. In Garside, R., Leech, G. and McEnery, A. (Eds.) *Corpus annotation: linguistic information from computer text corpora*, 66–84. London: Addison Wesley Longman.

Ge, N. (1998) *Annotating the Penn Treebank with coreference information*. Internal report, Department of Computer Science, Brown University.

Ge, N., Hale, J. and Charniak, E. (1998) 'A statistical approach to anaphora resolution'. *Proceedings of the Workshop on Very Large Corpora*, 161–170. Montreal, Canada.

Gelbukh, A. and Sidorov, G. (1999) 'A dictionary-based algorithm for indirect anaphora resolution'. *Proceedings of VEXTAL'99*, 169–173. Venice, Italy.

Geldbach, S. (1997) 'Pronominale Referenzen in der maschinellen Übersetzung: ein Verfahren zur Anaphernresolution', *Sprache und Datenverarbeitung*, 21 (2), 76–85.

Geldbach, S. (1999) 'Anaphora and translation discrepancies in Russian–German MT'. *Machine Translation*, 14 (3–4), 217–230.

Gordon, P., Grosz, B. and Gilliom, L. (1993) 'Pronouns, names and the centering attention in discourse', *Cognitive Science*, 17 (3), 311–347.

Grishman, G. (1986) *Computational linguistics*. Cambridge: Cambridge University Press.

Grishman, G. (2002) 'Information extraction'. In Mitkov, R. (Ed.) *Oxford handbook of computational linguistics*. Oxford: Oxford University Press.

Grosz, B. (1977a) 'The representation and use of focus in a system for understanding dialogs'. *Proceedings of the 5th International Joint Conference on Artificial Intelligence (IJCAI'77)*, 67–76. Cambridge, Massachusetts.

Grosz, B. (1977b) *The representation and use of focus in dialogue understanding*, Technical report No. 151, SRI International, Menlo Park, California.

Grosz, B. and Sidner C. (1986) 'Attentions, intentions and the structure of discourse'. *Computational Linguistics*, 12, 175–204.

Grosz, B., Joshi, A. and Weinstein, S. (1983) 'Providing a unified account of definite noun phrases in discourse'. *Proceedings of the 21st Annual Meeting of the Association for Computational Linguistics (ACL'83)*, 44–50. Cambridge, Massachusetts.

Grosz, B., Joshi, A. and Weinstein, S. (1986) *Towards a computational theory of discourse interpretation*. Unpublished manuscript.

Grosz, B., Aravind J. and Weinstein, S. (1995) 'Centering: a framework for modelling the local coherence of discourse'. *Computational Linguistics*, 21 (2), 203–225.

Guindon, R. (1988) 'A multidisciplinary perspective on dialogue structure in user-advisor dialogues'. In Guindon, R. (Ed.) *Cognitive science and its applications for human–computer interaction*. Hillsdale, NJ: Erlbaum.

Günther, F. and Lehmann, H. (1983) 'Rules for pronominalisation'. *Proceedings of the First Conference of the European Chapter of the Association for Computational Linguistics*, 144–151. Pisa, Italy.

Haegeman, L. (1994) *Introduction to government and binding theory*. Oxford: Blackwell.

Hahn, U. and Strube, M. (1997) 'Centering-in-the-large: computing referential discourse segments'. *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics*, 104–111. Madrid, Spain.

Halliday, M. and Hasan, R. (1976) *Cohesion in English*. London: Longman.

Harabagiu, S. and Maiorano, S. (1999) 'Knowledge-lean coreference resolution and its relation to textual cohesion and coherence'. *Proceedings of the ACL'99 Workshop on the Relation of Discourse/Dialogues Structure and Reference*, 29–38. University of Maryland.

Harabagiu, S. and Maiorano, S. (2000) 'Multilingual Coreference Resolution'. *Proceedings of ANLP-NAACL2000*, 142–149. Seattle, Washington.

Harabagiu, S., Bunescu, R. and Maiorano, S. (2001) 'Text and knowledge mining for coreference resolution'. *Proceedings of the 2nd Meeting of the North American Chapter of the Association of Computational Linguistics (NAACL-2001)*, 55–62. Pittsburgh, PA, USA.

Harbert, W. (1995) 'Binding theory, control, and pro'. In Webelhuth, G. (Ed.) *Government and binding theory and the minimalist program*, 177–240. Oxford: Blackwell, UK.

Hartrumpf, S. (2001) 'Coreference resolution with syntactic-semantic rules and corpus statistics'. *Proceedings of the ACL 2001 workshop on Computational Natural Language Learning*, 137–144. Toulouse, France.

Hawkins, J. (1978) *Definiteness and indefiniteness*. London: Croom Helm.

Hearst, M. (1994) 'Multi-paragraph segmentation of expository text'. *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics*, 9–16. Las Cruces, New Mexico, USA.

Hearst, M. (1997) 'TextTiling: segmenting text into multi-paragraph subtopic passages'. *Computational Linguistics*, 23 (1), 33–64.

Hinds, J. (1978) 'Anaphora in discourse'. *Current Enquiries into Language and Linguistics*, 22, 180–222. Edmonton: Linguistic research.

Hirschman, L. (1997) *MUC-7 coreference task definition*. Version 3.0.

Hirschman, L. and Chinchor, N. (Eds.) (1997) *MUC-7 Proceedings*. Science Applications International Corporation.

Hirschman, L., Robinson, P., Burger J. and Vilain, M. (1997) 'Coreference annotation and evaluation'. Paper presented at the *SALT Evaluation Workshop*. Sheffield, UK.

Hirschman, L., Robinson, P., Burger, J. and Vilain, M. (1998) 'The role of annotated training data'. *Proceedings of the Workshop on Linguistic Coreference*. Granada, Spain.

Hirst, G. (1981) *Anaphora in natural language understanding*. Berlin: Springer Verlag.

Hitzeman, J. and Poesio, M. (1998) 'Long distance pronominalisation and global focus'. *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and of the 18th International Conference on Computational Linguistics (COLING'98/ACL'98)*, 550–556. Montreal, Canada.

Hobbs, J. (1976) *Pronoun resolution*. Research Report 76–1. New York: Department of Computer Science, City University of New York.

Hobbs, J. (1978) 'Resolving pronoun references'. *Lingua*, 44, 339–352.

Hobbs, J., Douglas, A., Bear, J., Israel, D., Kameyama, Stickel, M. and Tyson, M. (1996) 'FASTUS: a cascaded finite-state transducer for extracting information from natural language text'.

In Roche, E. and Shabes, Y. (Eds.) *Finite-state devices for Natural Language Processing*. Cambridge, Massachusetts: MIT Press.

Huddleston, R. (1984) *Introduction to English grammar*. Cambridge: Cambridge University Press.

Hudson-D'Zmura, S. (1988) *The structure of discourse and anaphor resolution: the discourse center and the roles of nouns and pronouns*. PhD thesis, University of Rochester.

Hutchins, J. and Somers, H. (1992) *An introduction to Machine Translation*. London: Academic Press.

Ingria, R. and Stallard, D. (1989) 'A computational mechanism for pronominal reference'. *Proceedings of the 27th Annual Meeting of the ACL*, 262–271. Vancouver, British Columbia.

Jackendoff, R. (1977) *X-bar syntax: a study of phrase structure*. Cambridge, Massachusetts: MIT Press.

Jensen, K. (1986) *PEG 1986: a broad-coverage computational syntax of English*. Technical report, IBM T.J. Watson Research Center.

Joshi, A. and Kuhn, S. (1979) 'Centered logic: the role entity centered sentence representation in natural language inferencing'. *Proceedings of the 6th International Joint Conference on Artificial Intelligence (IJCAI)*, 435–439. Tokyo, Japan.

Joshi, A. and Weinstein, S. (1981) 'Control of inference: role of some aspects of discourse structure – centering'. *Proceedings of the 7th International Joint Conference on Artificial Intelligence (IJCAI-81)*, 385–387. Vancouver, Canada.

Kameyama, M. (1995) *Zero anaphora: the case of Japanese*, PhD thesis. Stanford University.

Kameyama, M. (1996) 'A property-sharing constraint in centering'. *Proceedings of the 24th Annual Meeting of the Association for Computational Linguistics (ACL'86)*, 200–206. New York, USA.

Kameyama, M. (1997) 'Recognizing referential links: an information extraction perspective'. *Proceedings of the ACL'97/EACL'97 Workshop on Operational Factors in Practical, Robust Anaphora Resolution*, 46–53. Madrid, Spain.

Kameyama, M. (1998) 'Intrasentential centering: a case study'. In Walker, M., Joshi, A. and Prince, E. (Eds.) *Centering theory in discourse*, 89–112. Oxford: Clarendon Press.

Kamp, H. (1981) 'A theory of truth and semantic representation'. In Groenendijk, J., Janssen, T. and Stokhof, M. (Eds.) *Formal methods in the study of language*, 277–322. Mathematical Centre tract 135. Amsterdam.

Kamp, H. and Reyle, U. (1993) *From discourse to logic*. Dordrecht: Kluwer.

Kantor, R. (1977) *The management and comprehension of discourse connection by pronouns in English*, PhD thesis. Department of Linguistics, Ohio University.

Karlsson, F., Voutilainen, A., Heikkilä, J. and Anttila, A. (Eds.) (1995) *Constraint grammar: a language-independent system for parsing free text*. Berlin/New York: Mouton de Gruyter.

Karttunen, L. (1969) Pronouns and variable. *CLS* 5.

Kehler, A. (1997a) 'Current theories of centering and pronoun interpretation: a critical evaluation'. *Computational Linguistics*, 23 (3), 467–475.

Kehler, A. (1997b) 'Probabilistic coreference in information extraction'. *Proceedings of the 2nd Conference on Empirical Methods in Natural Language Processing (EMNMLP-2)*, 163–173. Providence, Rhode Island, USA.

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

Kibble, R. (2001) 'A reformulation of rule 2 of centering theory'. *Computational Linguistics*, 27 (4), 579–587.

Kibble, R. and van Deemter, K. (2000) 'Coreference annotation: whither'. *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC'2000)*, 1281–1286. Athens, Greece.

Klapholz, D. and Lockman, A. (1975) 'Contextual reference resolution'. *American Journal of Computational Linguistics*, microfiche 36.

Krushkov, H. (1997) *Modelling and building of machine dictionaries and morphological processors* (in Bulgarian). PhD thesis, University of Plovdiv.

Kuno, S. (1972) 'Functional sentence perspective: a case study from Japanese and English'. *Linguistic Inquiry*, 3, 269–320.

Kuno, S. (1975) 'Three perspectives in the functional approach to syntax'. In Grossman, R., San, L. and Vance, T. (Eds.) *Papers from the parasession on functionalism*, 276–336. Chicago: Chicago Linguistic Society.

Leass, H. and Schwall, U. (1991) *An anaphora resolution procedure for machine translation*. IBM Germany Science Center. Institute for Knowledge Based Systems, Report 172.

Langacker, R. (1969) 'On pronominalisation and the chain of command'. In Reibel, D. and Schane, S. (Eds.) *Modern studies in English*, 160–186. Englewood Cliffs: Prentice Hall.

Lappin, S. and Leass, H. (1994) 'An algorithm for pronominal anaphora resolution'. *Computational Linguistics*, 20 (4), 535–561.

Leech, G. and Garside, R. (1991) 'Running a grammar factory: the production of syntactically analysed corpora or treebanks'. In Johannsson, S. and Stenstrom, A. (Eds.) *English computer corpora: selected papers and research guide*, 15–32. Berlin: Mouton De Gruyter.

Lockman, A. (1978) *Contextual reference resolution*. PhD thesis, Faculty of Pure Science, Columbia University.

Lyons, J. (1977) *Semantics*. Cambridge: Cambridge University Press.

Mann, W. and Thompson, S. (1988) Rhetorical structure theory: toward a functional theory of text organization. *Text*, 8 (3), 243–281.

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

Martínez-Barco, P., Muñoz, R., Azzam, S., Palomar, M. and Ferrández, A. (1999) 'Evaluation of pronoun resolution algorithm for Spanish dialogues'. *Proceedings of VEXTAL'99*, 325–332. Venice, Italy.

McCarthy, J. and Lehnert, W. (1995) 'Using decision trees for coreference resolution'. *Proceedings of the 14th International Conference on Artificial Intelligence (IJCAI-95)*, 1050–1055. Montreal, Canada.

McCord, M. (1989) *A new version of slot grammar*. Research Report RC 14506, IBM Research Division, Yorktown Heights, New York.

McCord, M. (1990) 'Slot grammar: a system for simpler construction of practical natural language grammars'. In Studer, R. (Ed.) *Natural language and logic: international scientific symposium*, 118–145. Lecture Notes in Computer Science. Berlin: Springer Verlag.

McCord, M. (1993) 'Heuristics for broad-coverage natural language parsing'. *Proceedings, APRA Human Language Technology Workshop*, University of Pennsylvania.

McEnery, A. (2002) 'Corpora'. In Mitkov, R. (Ed.) *The Oxford handbook of computational linguistics*. Oxford: Oxford University Press.

McEnery, A., Tanaka, I. and Botley, S. (1997) 'Corpus annotation and reference resolution'. *Proceedings of the ACL'97/EACL'97 Workshop on Operational Factors in Practical, Robust Anaphora Resolution*, 67–74. Madrid, Spain.

Minsky, M. (Ed.) (1968) *Semantic information processing*. Cambridge, Massachusetts: MIT Press.

Minsky, M. (1975) 'A framework for representing knowledge'. In Winston, P.H. (Ed.) *The psychology of computer vision*. New York: McGraw-Hill.

Mitkov, R. (1994a). 'An integrated model for anaphora resolution'. *Proceedings of the 15th International Conference on Computational Linguistics (COLING'94)*, 1170–1176. Kyoto, Japan.

Mitkov, R. (1994b) 'A new approach for tracking center'. *Proceedings of the International Conference 'New Methods in Language Processing' (NeMeLaP-1)*, 150–154. Manchester, UK.

Mitkov, R. (1995a) *Anaphora resolution in Natural Language Processing and Machine Translation*. Working paper. Saarbrücken: IAI.

Mitkov, R. (1995b) 'An uncertainty reasoning approach for anaphora resolution'. *Proceedings of the Natural Language Processing Pacific Rim Symposium (NLPRS'95)*, 149–154. Seoul, Korea.

Mitkov, R. (1996) 'Pronoun resolution: the practical alternative'. Paper presented at the *Discourse Anaphora and Anaphor Resolution Colloquium (DAARC)*, Lancaster, UK. Also appeared in Botley, S. and McEnery, A. (Eds.) (2000) *Corpus-based and computational approaches to discourse anaphora*, 189–212. Amsterdam/Philadelphia: John Benjamins.

Mitkov, R. (1997a) 'Factors in anaphora resolution: they are not the only things that matter. A case study based on two different approaches'. *Proceedings of the ACL'97/EACL'97 Workshop on Operational Factors in Practical, Robust Anaphora Resolution*, 14–21. Madrid, Spain.

Mitkov, R. (1997b) 'How far are we from (semi-)automatic annotation of anaphoric links in corpora'. *Proceedings of the ACL'97/EACL'97 Workshop on Operational Factors in Practical, Robust Anaphora Resolution*, 82–87. Madrid, Spain.

Mitkov, R. (1998a) 'Evaluating anaphora resolution approaches'. *Proceedings of the Discourse Anaphora and Anaphora Resolution Colloquium (DAARC'2)*, 164–172. Lancaster, UK.

Mitkov, R. (1998b). 'Robust pronoun resolution with limited knowledge'. *Proceedings of the 17th International Conference on Computational Linguistics (COLING'98/ACL'98)*, 869–875. Montreal, Canada.

Mitkov, R. (1999a) *Anaphora resolution: the state of the art*. Working paper. University of Wolverhampton, Wolverhampton.

Mitkov, R. (1999b) 'Towards automatic annotation of anaphoric links in corpora'. *International Journal of Corpus Linguistics*, Vol. 4(2), 261–280.

Mitkov, R. (1999c) 'Multilingual anaphora resolution'. *Machine Translation*, 14 (3–4), 281–299.

Mitkov, R. (2000) 'Towards more comprehensive evaluation in anaphora resolution'. *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC'2000)*, Vol. III, 1309–1314. Athens, Greece.

Mitkov, R. (2001a) 'Outstanding issues in anaphora resolution'. In Al. Gelbukh (Ed.) *Computational linguistics and intelligent text processing*, 110–125. Springer.

Mitkov, R. (2001b) 'Towards a more consistent and comprehensive evaluation of anaphora resolution algorithms and systems'. *Applied Artificial Intelligence: An International Journal*, 15, 253–276.

Mitkov, R. (2001c) 'Evaluation in anaphora resolution'. Presentation at the EUROLAN 2001 Summer School on *Creation and exploitation of annotated language resources*. CD-ROM. University of Iasi, Iasi, Romania.

Mitkov, R. (Ed.) (2002) *The Oxford handbook of computational linguistics*. Oxford: Oxford University Press.

Mitkov, R. and Stys, M. (1997) 'Robust reference resolution with limited knowledge: high precision genre-specific approach for English and Polish'. *Proceedings of the International Conference 'Recent Advances in Natural Language Processing' (RANLP'97)*, 74–81. Tzigov Chark, Bulgaria.

Mitkov, R. and Schmidt, P. (1998) 'On the complexity of pronominal anaphora resolution in Machine Translation'. In Martín-Vide, C. (Ed.) *Mathematical and computational analysis of natural language*, 207–222. Amsterdam/Philadelphia: John Benjamins.

Mitkov, R. and Barbu, C. (2000) 'Improving pronoun resolution in two languages by means of bilingual corpora'. *Proceedings of the Discourse, Anaphora and Reference Resolution Conference (DAARC2000)*, 133–137. Lancaster, UK.

Mitkov, R., Choi, S.K. and Sharp, R. (1995) 'Anaphora resolution in Machine Translation'. *Proceedings of the Sixth International Conference on Theoretical and Methodological Issues in Machine Translation (TMI'95)*, 87–95. Leuven, Belgium.

Mitkov, R., Lee, K.H., Kim, H. and Choi, K.S. (1997) 'English-to-Korean Machine Translation and the problem of anaphora resolution'. *Journal of Literary and Linguistics Computing*, 12 (1), 23–30.

Mitkov, R., Belguith, L. and Stys, M. (1998) 'Multilingual robust anaphora resolution'. *Proceedings of the Third International Conference on Empirical Methods in Natural Language Processing (EMNLP-3)*, 7–16. Granada, Spain.

Mitkov, R., Orasan, C. and Evans, R. (1999) 'The importance of annotated corpora for Natural Language Processing: the case of anaphora resolution and clause splitting'. *Proceedings of the TALN'99 Workshop on Corpora and NLP*, 60–69. Cargese, France.

Mitkov, R., Evans, R., Orasan, C., Barbu, C., Jones, L. and Sotirova, V. (2000) 'Coreference and anaphora: developing annotating tools, annotated resources and annotation strategies'. *Proceedings of the Discourse, Anaphora and Reference Resolution Conference (DAARC2000)*, 49–58. Lancaster, UK.

Mitkov, R., Evans R. and Orasan, C. (2002). 'A new, fully automatic version of Mitkov's knowledge-poor pronoun resolution method'. In Al. Gelbukh (Ed.). *Computational Linguistics and Intelligent Text Processing*, 168–186. Springer.

Mooney, R. (2002) 'Machine learning'. In Mitkov, R. (Ed.) *The Oxford handbook of computational linguistics*. Oxford: Oxford University Press.

Morgan, J. (1968). 'Some strange aspects of it'. Papers from the fourth regional meeting, Chicago Linguistic Society, April 1968, 81–93.

Mori, T., Matsuo, M. and Nakagawa, H. (1997) 'Constraints and defaults of zero pronouns in Japanese instruction manuals'. *Proceedings of the ACL'97/EACL'97 Workshop on Operational Factors in Practical, Robust Anaphora Resolution*, 7–13. Madrid, Spain.

Morton, T. (1999) 'Using coreference for question answering'. *Proceedings of the ACL'99 Workshop on Coreference and its Applications*, 85–89. College Park, Maryland, USA.

Morton, T. (2000) 'Coreference for NLP applications'. *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics (ACL)*, 173–180. Hong Kong.

Muñoz, R. (2001) *Tratamiento y resolución de las descripciones definidas y su applicación en sistemas de extracción de información*. PhD thesis, University of Alicante.

Muñoz, R. and Palomar, M. (2000) 'Processing of Spanish definite description with the same head'. *Proceedings of NLP'2000*, 212–220. Patras, Greece.

Muñoz, R. and Palomar, M. (2001) 'Semantic-driven algorithm for definite description resolution'. *Proceedings of the International Conference 'Recent Advances in Natural Language Processing' (RANLP'2001)*, 180–186. Tzigov Chark, Bulgaria.

Muñoz, R., Montoyo, A., Llopis, F. and Suárez, A. (1998) 'Reconocimiento de entitades en el sistema EXIT'. *Procesamiento del Lenguaje Natural*, 23, 47–53.

Muñoz, R., Saiz-Noeda, M., Suárez, A. and Palomar, M. (2000) 'Semantic approach to bridging reference resolution'. *Proceedings of the International Conference 'Machine Translation and Multilingual Natural Language Processing' (MT2000)*, 17-1–17-7. Exeter, UK.

Murata, M. and Nagao, M. (2000) 'Indirect reference in Japanese sentences'. In Botley, S. and McEnery, A. (Eds.) *Corpus-based and computational approaches to discourse anaphora*, 211–226. Amsterdam/Philadelphia: John Benjamins.

Nakaiwa, H. and Ikehara, S. (1992) 'Zero pronoun resolution in a Japanese-to-English machine translation system by using verbal semantic attributes'. *Proceedings of 3rd Conference on Applied Natural Language Processing (ANLP'92)*, 201–208. Trento, Italy.

Nakaiwa, H. and Ikehara, S. (1995) 'Intrasentential resolution of Japanese zero pronouns in a machine translation system using semantic and pragmatic constraints'. *Proceedings of the International Conference on Theoretical and Methodological Issues in Machine Translation (TMI'95)*, 96–105. Leuven, Belgium.

Nakaiwa, H., Yokoo, A. and Ikehara, S. (1994) 'A system of verbal semantic attributes focused on the syntactic correspondence between Japanese and English'. *Proceedings of the 15th International Conference on Computational Linguistics (COLING'94)*, 672–678. Kyoto, Japan.

Nakaiwa, H., Shirai, S., Ikehara, S. and Kawaoka, T. (1995) 'Extrasentential resolution of Japanese zero pronouns using semantic and pragmatic constraints'. *Proceedings of the AAAI 1995 Spring Symposium Series: Empirical Methods in Discourse Interpretation and Generation*, 99–105. Stanford, California, USA.

Nakaiwa, H., Bond, F., Uekado, T. and Nozawa, Y. (1996) 'Resolving zero pronouns in texts using textual structure'. *Proceedings of the International Conference 'New Methods in Language Processing' (NeMLaP-2)*, 25–36. Ankara, Turkey.

Nariyama, S. (2000) *Referent identification for ellipted arguments in Japanese*. PhD thesis. The University of Melbourne, Australia.

Nasukawa, T. (1994) 'Robust method of pronoun resolution using full-text information'. *Proceedings of the 15th International Conference on Computational Linguistics (COLING'94)*, 1157–1163. Kyoto, Japan.

Orasan, C. (2000) 'CLinkA – a coreferential links annotator'. *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC'2000)*, 491–496. Athens, Greece.

Orasan, C. and Evans, R. (2000) 'Experiments in optimising the task of anaphora resolution'. *Proceedings of ICEIS 2000*, 191–195. Stafford, UK.

Orasan, C. and Evans, R. (2001) 'Learning to identify animate references'. *Proceedings of the Workshop Computational Natural Language Learning 2001 (CoNLL-2001)*, 129–136. Toulouse, France.

Orasan, C., Evans, R. and Mitkov, R. (2000) 'Enhancing preference-based anaphora resolution with genetic algorithms'. *Proceedings of NLP'2000*, 185–195. Patras, Greece.

Paice, C. and Husk, G. (1987) 'Towards the automatic recognition of anaphoric features in English text: the impersonal pronoun "it" '. *Computer Speech and Language*, 2, 109–132.

Palomar, M., Saiz-Noeda, M., Muñoz, R., Suárez, A., Martínez-Barco, P. and Montoyo, A. (2001a) 'PHORA: A NLP system'. *International Conference on Intelligence Text Processing and Computational Linguistics (CicLing-2001)*, 126–139. Lecture Notes in Computer Science 2004. Berlin: Springer.

Palomar, M., Ferrández, A., Moreno, L., Martínez-Barco, P., Peral, J., Saiz-Noeda, M. and Muñoz, R. (2001b) 'An algorithm for anaphora resolution in Spanish texts'. *Computational Linguistics*, 27 (4), 545–567.

Passonneau, R. (1996) *Instructions for applying Discourse Reference Annotation for Multiple Applications (DRAMA)*. Unpublished Internal Document.

Passonneau, R. and Litman, D. (1997) 'Discourse segmentation by human and automated means'. *Computational Linguistics* 23 (1), 3–139.

Peral, J., Palomar, M. and Ferrández, A. (1999) 'Coreference-oriented interlingual slot structure and Machine Translation'. *Proceedings of the ACL'99 Workshop on Coreference and its Applications*, 69–76. College Park, Maryland, USA.

Pérez, C.R. (1994) 'Estudio de la incidencia de differentes fuentes de la información en el establecimiento de relaciones anafóricas'. *Bulletín de la Sociedad Española para el Procesamiento del Lenguaje Natural*, No. 14, March.

Poesio, M. (2000) 'Semantic analysis'. In Dale, R., Moisl, H. and Somers, H. (Eds.) *Handbook of natural language processing*. 93–122. New York and Basel: Marcel Dekker.

Poesio, M. and Vieira, R. (1998) 'A corpus-based investigation of definite description use'. *Computational Linguistics*, 24 (2), 183–216.

Poesio, M., Vieira, R. and Teufel, S. (1997) 'Resolving bridging references in unrestricted text'. *Proceedings of the ACL'97/EACL'97 Workshop on Operational Factors in Practical, Robust Anaphora Resolution*, 1–6. Madrid, Spain.

Poesio, M., Cheng, H., Henschel, R., Hitzeman, J., Kibble, R. and Stevenson, R. (2000) 'Specifying the parameters of centering theory: a corpus-based evaluation using text from

application-oriented domains'. *Proceedings of 38th Annual Meeting of the Association for Computational Linguistics (ACL)*. 400–407. Hong Kong.

Popescu-Belis, A. (1998) 'How corpora with annotated coreference links improve reference resolution'. *Proceedings of the First International Conference on Language Resources and Evaluation*, 567–572. Granada, Spain.

Popescu-Belis, A. and Robba, I. (1997) 'Cooperation between pronoun and reference resolution for unrestricted texts'. *Proceedings of the ACL'97/EACL'97 Workshop on Operational Factors in Practical, Robust Anaphora Resolution*, 30–37. Madrid, Spain.

Popescu-Belis, A. and Robba, I. (1998) 'Three methods for evaluating reference resolution'. *Proceedings of the Workshop on Linguistic Coreference*. Granada, Spain.

Preuß, S., Schmitz, B., Hauenschild, C. and Umbach, C. (1994) 'Anaphora resolution in Machine Translation'. In Ramm, W. (Ed.) *Studies in Machine Translation and Natural Language Processing*, (Vol. 6 'Text and content in Machine Translation: aspects of discourse representation and discourse processing'), 29–52. Luxembourg: Office for Official Publications of the European Community.

Prince, E. (1981) 'Toward a taxonomy of given-new information'. In Cole, P. (Ed.) *Radical pragmatics*, 223–255. New York: Academic Press.

Quinlan, J.R. (1986) 'Induction for decision trees'. *Machine Learning*, 1 (1), 81–106.

Quinlan, J.R. (1993) *C4.5: Programs for machine learning*. San Mateo, California: Morgan Kaufmann.

Quirk, R., Greenbaum, S., Leech, G. and Svartvik, J. (1985) *A comprehensive grammar of the English language*. Longman.

Reinhart, T. (1976) *The syntactic domain of anaphora*. PhD thesis, MIT, Cambridge, Massachussets.

Reinhart, T. (1983a) *Anaphora and semantic interpretation*. London: Croom Helm.

Reinhart, T. (1983b) 'Coreference and bound anaphora: a restatement of the anaphora questions'. *Linguistics and Philosophy*, 6, 47–88.

Rich, E. and LuperFoy, S. (1988) 'An architecture for anaphora resolution'. *Proceedings of the Second Conference on Applied Natural Language Processing (ANLP-2)*, 18–24. Texas, USA.

Riloff, E. and Jones, R. (1999) 'Learning dictionary for information extraction by multi-level bootstrapping'. *Proceedings of the 16th National Conference on Artificial Intelligence (AAAI-99)*, 474–479. Orlando, Florida, USA.

Rocha, M. (1997) 'Supporting anaphor resolution with a corpus-based probabilistic model'. *Proceedings of the ACL'97/EACL'97 Workshop on Operational Factors in Practical, Robust Anaphora Resolution*, 54–61. Madrid, Spain.

Rocha, M. (1999) 'Coreference resolution in dialogues in English and Portuguese'. *Proceedings of the ACL'99 Workshop on Coreference and its Applications*, 53–61. College Park, Maryland, USA.

Rolbert, M. (1989) *Résolution de formes pronominales dans l'interface d'interrogation d'une base de données*. Thèse de doctorat. Faculté des Sciences de Luminy.

Ross, J. (1967) 'On the cyclic nature of English pronominalisation'. In *To honour Roman Jakobson, 1669–1682*. The Hague: Mouton.

Russell, B. (1905) 'On denoting'. *Mind*, 14, 479–493. Reprinted in Marsh, R.C. (Ed.) *Logic and knowledge*. London: George Allen and Unwin.

Saggion, H. and Carvalho, A. (1994) 'Anaphora resolution in a machine translation system'. *Proceedings of the International Conference 'Machine Translation, 10 years on'*, 4.1–4.14. Cranfield, UK.

Saiz-Noeda, M., Peral, J. and Suáres, A. (2000) 'Semantic compatibility techniques for anaphora resolution'. *Proceedings of the Workshop on Corpora and NLP*, 43–48. Monastir, Tunisia.

Sanford, A. and Garrod, S. (1989) 'What, when, and how? Questions of immediacy in anaphoric reference resolution'. *Language and Cognitive Processes*, 4 (3/4), 235–262.

Sidner, C. (1979) *Toward a computational theory of definite anaphora comprehension in English*. Technical report No. AI-TR-537. Cambridge, Massachussetts: MIT Press.

Sidner, C. (1983) 'Focusing in the comprehension of definite anaphora'. In Brady, M. and Berwick, R. (Eds.) *Computational models of discourse*. Cambridge, Massachussetts: MIT Press. Also published in Grosz, B., Jones, K.S. and Webber, B. (Eds.) *Readings in Natural Language Processing*. Morgan Kaufmann Publishers, Inc. (1986).

Smith, G. (1991) *Computers and human language*. New York: Oxford University Press.

Soon, W.M., Ng, H.T. and Lim, C.Y. (1999) 'Corpus-based learning for noun phrase coreference resolution'. *Proceedings of the 1999 Joint SIGDAT Conference on Empirical Methods in NLP and Very Large Corpora*, 285–291. University of Maryland, USA.

Soon, W.M., Ng, H.T. and Lim, C.Y. (2001) 'A machine learning approach to coreference resolution of noun phrases'. *Computational Linguistics*, 27 (4), 521–544.

Strube, M. (1998) 'Never look back: an alternative to centering'. *Proceedings of the 17th International Conference on Computational Linguistics (COLING'98/ACL'98)*, 1251–1257. Montreal, Canada.

Strube, M. and Hahn, U. (1996) 'Functional centering'. *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics*, 270–277. Santa Cruz, California, USA.

Stuckardt, R. (1996) 'An interdependency-sensitive approach to anaphor resolution'. *Proceedings of the International Colloquium on Discourse Anaphora and Anaphora Resolution. Lancaster (DAARC)*, 400–413. Lancaster, UK.

Stuckardt, R. (1997) 'Resolving anaphoric references on deficient syntactic descriptions'. *Proceedings of the ACL'97/EACL'97 Workshop on Operational Factors in Practical, Robust Anaphora Resolution*, 30–37. Madrid, Spain.

Stuckardt, R. (2001) 'Design and enhanced evaluation of a robust anaphor resolution algorithm'. *Computational Linguistics*, 27 (4), 479–506.

Stys, M. and Zemke, S. (1995). 'Incorporating discourse aspects in English–Polish MT: towards robust implementation'. *Proceedings of the International Conference 'Recent Advances in Natural Language Processing' (RANLP'95)*, 95–102. Tzigov Chark, Bulgaria.

Suri, L. and McCoy, K. (1994) 'RAFT/RAPR and centering: a comparison and discussion of problems related to preceding complex sentences'. *Computational Linguistics*, 20 (2), 301–317.

Tanaka, I. (2000) *The value of an annotated corpus in the investigation of anaphoric pronouns, with particular reference to backwards anaphora in English*. PhD Thesis, University of Lancaster.

Tanev, H. and Mitkov, R. (2000) 'LINGUA – a robust architecture for text processing and anaphora resolution in Bulgarian'. *Proceedings of the International Conference on Machine Translation and Multilingual Applications (MT2000)*, 20-1–20-8. Exeter, UK.

Tanev, H. and Mitkov, R. (2002). Shallow language processing architecture for Bulgarian (forthcoming)

Tapanainen, P. and Järvinen, T. (1997) 'A non-projective dependency parser'. *Proceedings of the 5th Conference of Applied Natural Language Processing (ANLP-5)*, 64–71. Washington, DC, USA.

Tetreault, J. (1999) 'Analysis of syntax-based pronoun resolution methods'. *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL'99)*, 602–605. Maryland, USA.

Tetreault, J. (2001) 'A corpus-based evaluation of centering and pronoun resolution'. *Computational Linguistics*, 27 (4), 507–520.

Tin, E. and Akman, V. (1994) 'Situated processing of pronominal anaphora'. *Proceedings of the KONVENS'94 Conference*, 369–378. Vienna, Austria.

Trouilleux, F., Gaussier, E., Bès, G. and Zaenen, A. (2000) 'Coreference resolution evaluation based on descriptive specificity'. *Proceedings of the Second International Conference on Language Resources and Evaluation*, 1315–1322. Athens, Greece.

Tsujimura, N. (1996) *An introduction to Japanese linguistics*. Massachusetts: Blackwell.

Tutin, A., Antoniadis, G. and Clouzot, C. (1999) 'Annoter le corpus pour le traitement des anaphors'. *Proceedings of the TALN'99 Workshop on Corpora and NLP*, 49–59. Cargese, France.

Tutin, A., Trouilleux, F., Clouzot, C., Gaussier, E., Zaenen, A., Rayot, S. and Antoniadis, G. (2000) 'Annotating a large corpus with anaphoric links'. *Proceedings of the Discourse, Anaphora and Reference Resolution Conference (DAARC2000)*, 28–38. Lancaster, UK.

Uehara, S. (1996) 'Anaphoric pronouns in English and their counterparts in Japanese'. *Proceedings of the Discourse, Anaphora and Reference Resolution Conference (DAARC96)*, 64–75. Lancaster, UK.

Van Deemter, K. and Kibble, R. (1999) 'What is coreference and what should coreference annotation be?' *Proceedings of the ACL99 Workshop on Coreference and its Applications*, 90–96. College Park, Maryland, USA.

Vicedo, J.L. and Ferrández, A. (2000) 'Importance of pronominal anaphora resolution to question answering systems'. *Proceedings of 38th Annual Meeting of the Association for Computational Linguistics (ACL)*, 555–562. Hong Kong.

Vieira, R. and Poesio, M. (2000a) 'Processing definite descriptions in corpora'. In Botley, S. and McEnery, A. (Eds.) *Corpus-based and computational approaches to discourse anaphora*, 189–212. Amsterdam/Philadelphia: John Benjamins.

Vieira, R. and Poesio, M. (2000b) 'An empirically-based system for processing definite descriptions'. *Computational Linguistics*, 26 (4), 525–579.

Vilain, M., Burger, J., Aberdeen, J., Connoly, D. and Hirschman, L. (1995) 'A model-theoretic coreference scoring scheme'. *Proceedings of the 6th Message Understanding Conference (MUC-6)*, 45–52. San Francisco, California, USA.

Voutilainen, A., Heikkilä, J. and Anttila, A. (1992) *A constraint grammar of English: a performance-oriented approach*. Publication No. 21, Helsinki: University of Helsinki.

Wada, H. (1990) 'Discourse processing in MT: problems in pronominal translation'. *Proceedings of the 13th International Conference on Computational Linguistics (COLING'90)*, Vol. 1, 73–75. Helsinki, Finland.

Wakao, T. (1994) 'Reference resolution using semantic patterns in Japanese newspaper articles'. *Proceedings of the 15th International Conference on Computational Linguistics (COLING'94)*, 1133–1137. Kyoto, Japan.

Walker, M. (1989) 'Evaluating discourse processing algorithms'. *Proceedings of the 27th Annual Meeting of the ACL (ACL'97)*, 251–261. Vancouver, Canada.

Walker, M. (1998) 'Centering, anaphora resolution and discourse structure'. In Walker, M., Joshi, A. and Prince, E. (Eds.) *Centering theory in discourse*. Oxford: Clarendon Press.

Walker, M., Iida, M. and Cote, S. (1994) 'Japanese discourse and the process of centering'. *Computational Linguistics*, 20 (2).

Walker, M., Joshi, A. and Prince, E. (1998) 'Centering in naturally occurring discourse: an overview'. In Walker, M., Joshi, A. and Prince, E. (Eds.) *Centering theory in discourse*. Oxford: Clarendon Press.

Webber, B. (1979) *A formal approach to discourse anaphora*. New York: Garland Publishing.

Wilks, Y. (1973) *Preference semantics*. Stanford AI Laboratory memo AIM-206. Stanford University.

Wilks, Y. (1975a) 'Preference semantics'. In Keenan, E. (Ed.) *The formal semantics of natural language*. Cambridge: Cambridge University Press.

Wilks, Y, (1975b) 'An intelligent analyzer and understander of English'. *Communications of the ACM*, 18, 264–274.

Wilks, Y. (1975c). 'A preferential pattern-seeking semantics for natural language interfaces'. *Artificial Intelligence*, 6, 53–74.

Wilks, Y. (1977) 'Good and bad arguments about semantic primitives', *Communication and Cognition*, 10, 181–221.

Williams, S., Harvey, M. and Preston, K. (1996) 'Rule-based reference resolution for unrestricted text using part-of-speech tagging and noun phrase parsing'. *Proceedings of the International Colloquium on Discourse Anaphora and Anaphora Resolution (DAARC)*, 441–456. Lancaster, UK.

Winograd, T. (1972) *Understanding natural language*. New York: Academic Press/Edinburgh: Edinburgh University Press.

Woods, W. (1968) 'Procedural semantics for a question-answering machine'. *AFIPS Conference Proceedings*, 33, FJCC, 457–471.

Woods, W. (1970) 'Transition network grammars for natural language analysis'. *Communications of the ACM*, 13 (10), 591–606.

Woods, W., Kaplan, R. and Nash-Webber, B. (1972) *The LUNAR Sciences Natural Language Information System: final report*. Report 2378. Cambridge, MA: Bolt Beranek and Newman.

Index