Bilingual Pronoun Resolution:

Experiments in English and French

Cătălina Barbu

A thesis submitted in partial fulfilment of the requirements of the University of Wolverhampton for the degree of Doctor of Philosophy

2003

This work or any part thereof has not been presented in any form to the University or to any other body whether for purposes of assessment, publication or for any other purpose (unless previously indicated). Save for any express acknowledgements, references and/or bibliographies cited in the work, I confirm that the intellectual content of the work is the result of my own efforts and of no other person.

Abstract

Anaphora resolution has been a subject of research in computational linguistics for more than 25 years. The interest it aroused was due to the importance that anaphoric phenomena play in the coherence and cohesiveness of natural language. A deep understanding of a text is impossible without knowledge about how individual concepts relate to each other; a shallow understanding of a text is often impeded without resolving anaphoric phenomena. In the larger context of anaphora resolution, pronoun resolution has benefited from a wide interest, with pronominal anaphora being one of the most frequent discourse phenomena. The problem has been approached in a variety of manners, in various languages.

The research presented in this thesis approaches the problem of pronoun resolution in the context of multilingual NLP applications. In the global information society we are living in, fast access to information in the language of one's choice is essential, and this access is facilitated by emerging multilingual NLP applications. As anaphora resolution is an integral part of many NLP applications, it necessarily has to be approached with the view of multilinguality in mind.

The present project describes an original approach to bilingual anaphora resolution, for English and French. The method is based on combining hand crafted rules with automatic learning. Although only designed and tested for English and French, the approach has the potential to be extended to other languages. The steps preceding the practical development of the system consisted in an investigation of common and distinct features of English and French pronominal systems, while the evaluation of the system provided an interesting insight into specific problems relating to bilingual anaphora resolution.

Acknowledgements

I would like to thank Professor Ruslan Mitkov, my director of studies, for his continuos support and interest in my work that he has shown during my PhD studies. To all my colleagues in the Research Group in Computational Linguistics, a big Thank You! for making my whole PhD experience more enjoyable; also, for many helpful discussions that helped shaping my ideas. Special thanks go to Professor Dan Cristea, who first introduced me to the world of Computational Linguistics.

I am also grateful to other people whose input contributed to this work. Many thanks to Monique Rolbert and Mike Jackson, my supervisors, for their constructive comments at various points during my research. Also to my examiners for their patience in reading my thesis and for their helpful suggestions. Many thanks to the University of Wolverhampton, who provided the financial support that made this work possible.

Finally, my gratitude goes to my parents, Viorel and Margareta and to my husband, Rab, whose continuos encouragement helped me complete this thesis.

Abbreviations

AI - Artificial Intelligence

API - Application Programming Interface

AR - Anaphora Resolution

BNC - British National Corpus

CL - Computational Linguistics

CT - Centering Theory

GA - Genetic Algorithm

GB - Government and Binding

ML - Machine Learning

MUC - Message Understanding Conference

NLP - Natural Language Processing

NLG - Natural Language Generation

NP - Noun Phrase

VP - Verb Phrase

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

In the global information society we are living in, fast access to information presented in different languages and different formats is essential. Consistently with the current efforts in the field of Information Technology of dealing with technical aspects of managing, storing and disseminating multilingual documents, Natural Language Processing (NLP) aims at providing methods and tools for understanding and exploiting natural text in a variety of languages. Among the large range of NLP applications, the treatment of anaphora plays a central role in both understanding and generating coherent discourse.

The research presented in this thesis revolves around two concepts: anaphora and multilinguality, whilst dealing with the integration of automatic anaphora resolution in multilingual NLP applications. Anaphora resolution has been extensively studied for English, with a number of original approaches and different implementations of the task; some approaches exist for other languages, including French, Spanish, Japanese, Italian and German, but the NLP research carried out for these languages does not equate to the research pursued for English. In some cases, developing an anaphora resolver for a new language has simply consisted of adapting a method originally

developed for English.

The purpose of this research is the development and analysis of an extensible pronoun resolution system, that can accommodate more than one language. The outcome of the research is an original approach to bilingual anaphora resolution, for English and French. The method is based on combining handcrafted rules with automatic learning. The resulting system (named ARMAL - *anaphora resolution by machine learning*) has been designed specifically for English and French and it has also been evaluated for these two languages in particular. It will be argued, however, that ARMAL has the potential (at least theoretically) to be used for other languages, with minor changes in the design.

1.1 About anaphora

The importance of anaphora resolution lies in the fact that anaphoric phenomena are an essential component of discourse cohesion. According to [Halliday and Hasan1976]:

> Cohesion occurs where the *interpretation* of some element in the text is dependent on that of another. The one *presupposes* the other, in the sense that it cannot be effectively decoded except by recourse to it. When this happens, a relation of cohesion is set up, and the two elements, the presupposing and the presupposed, are thereby at least potentially integrated in the text. [Halliday and Hasan1976]

The use of referential expressions is an extremely common feature of natural language. These referential expressions come in a large variety of forms, according to the context, the semantics of the discourse and the intended message of the text. It is not the aim of this thesis to provide a classification of anaphoric expressions, nor to study the factors that determine the use of different anaphoric expressions in context.

In the following, I will give only a very brief classification of anaphoric relations, that can help define the goals of this research.

In terms of the syntactic relation between the antecedent and anaphor, there are two main categories of anaphoric relations: nominal anaphora and verb-phrase anaphora[1]. *Nominal anaphora* occurs when a nominal anaphor (noun phrase, pronoun) has as antecedent a non-pronominal noun phrase. *Verb phrase anaphora* occurs when the antecedent is a verb phrase[2].

In terms of the semantic relation holding between the anaphor and its antecedent, nominal anaphora can be broadly classified into two categories. *Direct anaphora* holds between two entities linked by a relation of synonymy, identity or specialisation. *Indirect anaphora* (also called in the literature *bridging* and *associative* anaphora) involves two entities linked by a relation of meronymy or holonymy.

This work is concerned with a restricted type of nominal anaphora, mainly the interpretation of pronouns with noun phrase antecedent.

Among the large variety of anaphoric relations, pronominal reference has been the most extensively studied, being sometimes considered representative for the whole range of nominal anaphora types. As opposed to other anaphoric devices[3], pronouns have more than the capacity of referring; they are inherently referential[4]. Therefore, while in interpreting language one can sometimes ignore other types of referential expressions[5], not interpreting pronouns could mean missing important information.

[1] This division only captures the most important types of anaphora. Non-nominal anaphora covers more phenomena than just verb phrase anaphora, but this is probably the most frequent type of non-nominal anaphora.

[2] Also included in this category are the anaphors which have as antecedent a clause, a sentence, or a group of sentences, although strictly speaking these phenomena are different from verb phrase anaphora.

[3] Such as definite descriptions or adverbials, for example, which can be referential but are not always so

[4] It will be avoided for the moment the case of non-referential pronouns, sometimes referred in the literature as *pleonastic* or *non-nominal*. These cases will be covered in Chapter 2

[5] Unless, of course, a deep understanding of the text is required

The interpretation of pronouns is made more difficult by the fact that pronouns offer very little information about themselves. All they convey is some morphological and syntactical information, such as number, gender, person and case.

These considerations justify the interest that researchers showed towards developing automatic systems for anaphora resolution (and in particular for pronominal anaphora) in naturally occurring texts.

1.2 Applications of anaphora resolution in NLP

Being able to identify the antecedents of an anaphor is an important issue in many natural language processing tasks. In machine translation, the identification of the antecedent of a pronoun can be essential for the production of a meaningful and correct target text. Consider for example the following sentence:

(1.1) The monkeys ate the bananas because they were hungry.

A translation word-by-word into French would give perfect results, with the exception of the ambiguity introduced by the use of the pronoun. The pronoun *they* can be translated either as *elles* (third person, feminine, plural) or as *ils* (third person, masculine, plural), depending on the gender of its antecedent. Giving a random or default translation is not an option in this case, since it can lead to a target text with incorrect meaning. In order to generate the correct French pronoun (which is *ils*), we need to be able to identify the correct antecedent of the English pronoun *they*, which is *the monkeys*. If the antecedent is identified incorrectly as being *the bananas*, the error propagates into the French translation, which becomes:

(1.2) Les singes ont mangé les banannes parce-qu'elles avaient faim.

In this sentence, the pronoun *elles* can only be interpreted as referring to *les*

banannes (since this is the only possible antecedent that agrees in gender with the pronoun), therefore the message conveyed is *The monkeys ate the bananas because the bananas were hungry*, which is obviously not the intended meaning.

The interest in automatic anaphora resolution has renewed with the development of text searching technologies, made necessary by the availability of large collections of information on the Internet.

In information extraction and in question answering, one wants to be able to tell when two discourse entities refer to the same real-world object. In automatic summarisation, one wants to be able to produce cohesive abstracts, and anaphoric pronouns are a major cohesion device. Problems with computer produced abstracts appear when a sentence containing an anaphoric pronoun is included in an abstract, but not the sentence containing the antecedent. This can make the resulting abstract ambiguous or unintelligible.

The large range of applications for anaphora resolution was recognised in the MUC (Message Understanding Conference) community, a separate task being especially created for developing fully-automatic coreference resolution systems. The treatment of anaphora is also important in Natural Language Generation, where the aim is to produce discourse that is at the same both time cohesive and coherent. Natural discourse employs a variety of anaphoric devices, which have to be replicated as closely as possible in a computer-produced document. Therefore an opposite process to anaphora resolution has to be performed, i.e. it is the *choice* of appropriate anaphoric devices that is under focus.

So far, the way the treatment of anaphora is embedded in other natural language processing tasks has been presented, which is the main practical use of the systems for anaphora resolution and generation. Although less common, there have been applications of using anaphora resolvers as an end product on their own.

[Canning et al.2000], for example, shows that dyslexic people can be helped in understanding texts where the pronouns have been replaced by their corresponding noun phrase antecedents. In this kind of application, a very high precision anaphora resolver is necessary, even at the expense of a low recall.

1.3 Anaphora resolution in a multilingual environment

The necessity for multilingual applications has become more pressing with the fast increase of information globalisation that was made possible by the expansion of the Internet and by the advances in communication technologies. We live in a global information society, people need to communicate, shop, have rapid access to information, and they have the technology to do it.

10 years ago, English was the main language used for disseminating information on the Internet. 98% of the data published on the Internet was written in English. However, this tendency to make English the de-facto official language of the Internet has decreased in recent years, and will continue to do so. This is partly the consequence of a shift in the information content and user profile: Internet information is not mainly academic anymore, but also commercial, leisure and entertainment related. This implies that information is not targeted anymore at a restricted group of users, but at practically anyone.

While academic information continues to be mainly presented in English, the other types of web content are much more internationalised. Some information is of local importance only (online newspapers, marketing of local companies, governmental bodies or local independent organisations) and therefore needs to be presented in such a way that would attract as large an audience as possible. An important part of this presentation is that people should be able to access the information in their own

language.

This issue becomes especially relevant if one looks at some figures related to the number of internet users. In 2000, 52% of the people with Internet access were English speakers. According to some estimates provided by GlobalSight, a consultancy firm that advises organisations on internationalising websites, by 2003, only one third of the projected 603m Internet users will be English speakers. IDC[6] surveys also show, not surprisingly, that the vast majority of people prefer web sites in their native language, especially for leisure activities, such as online shopping. Moreover, they are more likely to buy things online if they can use their own language. Hence, it is not surprising that the Internet community acknowledges more and more the importance of providing access to information written in languages other than English. Most search engines nowadays provide the facility of having the web pages translated into the language of your choice, although the translation will only be a rough one. However, automated translation is only a step, and other applications only partially connected to translation are needed to provide global access to information on the Web.

Access to this information can be done through different tasks: multilingual information retrieval (finding documents in a second language given a query expressed in a first language), cross-lingual text summarisation, multilingual question answering, comprehension aids, speech recognition and generation, multilingual text to speech synthesis.

Designing natural language processing applications is expensive and time consuming. It requires resources (corpora, preferably large and enriched with annotation, ontologies, lexicons), linguistic formalisms, algorithms. Processing multilingual information adds a new dimension to the difficulty of natural language

[6]IDC is an international leading provider of technology intelligence, industry analysis, market data, and strategic and tactical guidance to builders, providers, and users of information technology (URL: http://www.idc.com)

processing tasks. While it is not possible to alleviate the need for developing multilingual resources (multilingual corpora, grammars, ontologies and lexicons), better efficiency can be obtained in designing multilingual applications. Redesigning applications for every additional language is expensive, slow, and not always necessary. Without looking for language universals, it is still possible to build multilingual applications that share the same design and core rules, while being fine tuned for specific individual languages. It has been previously shown that anaphora resolution is an important part of several natural language processing tasks, that are now approached from a multilingual perspective. Therefore, anaphora resolution has to be and can be expressed as a multilingual task.

1.4 The task

1.4.1 General overview

Anaphora resolution is a complex task, involving the disambiguation of a variety of phenomena: references to verb phrases, nominal phrases, ellipsis, bridging. As already mentioned, the aim of this work is to approach a subclass of nominal anaphora, which is *nominal anaphora by means of pronouns*. Pronominal anaphora is the most studied type of anaphora, which is partly due to the fact that it is the comparatively easiest anaphoric type to identify and it can be tackled automatically.

Restricting the analysis to pronominal anaphora does not mean however that one can ignore the pronouns that refer to non-nominal entities or the pronouns that do not refer at all. Although the resolution of these particular cases is not in the scope of this research, their identification is necessary so that they can be eliminated from further analysis.

Therefore, the system that will be described in this thesis (ARMAL) will take as input a text in natural language, will identify pronominal anaphors and will attempt to solve them. The execution of the system can be traced (by way of exemplification) on the following text:

> Linux's multiple virtual consoles system make *it* practical to use as a multi-tasking operating system by a visually impaired person working directly through Braille.
>
> In principle *it* should be possible to put together a complete, usable Linux system for a visually impaired person for about $500 (cheap & nasty PC + sound card). This compares with many thousands of dollars for other operating systems (screen reader software/ speech synthesiser hardware). *I* have yet to see this . *I* doubt *it* would work in practice because the software speech synthesisers available for Linux aren't yet sufficiently good. [Linux Access How-To]

Given this text as input, the pronoun resolution system will first have to identify the referential pronouns. The text contains five personal pronouns: three instances of the third person singular pronoun *it* and two instances of the first person singular pronoun *I*.

In expository texts, first and second person pronouns are not considered truly anaphoric[7], therefore they can be safely ignored by the system. As far as the third person pronouns are concerned, the first one refers back to a noun phrase (*Linux*), the second one is pleonastic, while the third one refers to a part of a sentence.

Since the aim is to solve only pronouns with noun phrase antecedent, only the first instance of *it* will be attempted to be solved. However, this means that the system

[7] In contrast to spoken registers, where first and second person pronouns are anaphoric, as these types of text can have multiple actors in a conversation. In expository texts, however, there is normally at most one narrator, which can be referred to by a first person pronoun; second person pronouns usually refer to the reader, thus being exophoric.

should be able to identify in the first instance the fact that the other two pronouns do not refer to noun phrases.

1.4.2 Pronouns tackled

With the considerations presented above, the pronouns that the system aims to solve are the following:

- personal pronouns: *he, she, it, they, him, her, them* / *il, elle, lui, ils, elles, l', le, la, les, se, s', eux*

  As previously mentioned, the system will only tackle expository texts, where first person pronouns do not have an anaphoric value. Therefore, personal, possessive and reflexive pronouns of first person will not be taken into account.

- possessive pronouns: *his, hers, theirs* / *le sien, la sienne, les siens, les siennes, le/la leur, les leurs*

- possessive determiners: *his, her, their, its* / *son, sa, ses, leur, leurs*

- reflexives: *himself, herself, itself* / *lui-même, elle-même, eux-mêmes, soi-même*

1.4.3 Aims of the system

The task of identifying pronominal references relies on a number of limited knowledge sources: syntactical, morphological, semantical, positional, that can be combined in a number of rules. The originality of any system is judged by the way it makes use of these knowledge sources. This thesis will present a system based on automatic learning, where the order and the weights of the rules are learned from a set of training data.

The work presented in this thesis is not a piece of research in linguistics, but in language engineering. Therefore, the main stress was on the implementation of a practical system, on its efficiency and effectiveness. The main requirement of the system was to achieve good performance, but also to be fast, robust and reliable, and, last but not least, to be modular enough to enable its integration in a larger NLP application. A special requirement from a multilingual anaphora resolver is to be easily extended (fine tuned) for other languages.

In designing the system, three main assumptions have been made. The first one refers to the character of written texts, and that is that domain specific written texts tend to employ certain patterns of use for referential expressions. If enough data is available for analysis, certain patterns of behaviour of anaphoric expressions can be recorded and used in the analysis of other texts. This assumption led to the choice of machine learning as the basis for the anaphora resolver. The second assumption was that different anaphoric expressions (in the same language) require different resolution strategies. This was reflected in the selection of training and learning features. The third assumption was that languages with similar pronoun systems share a core of anaphoric behaviours, i.e. equivalent anaphoric expressions in those languages[8] are used to refer to the same class of entities in similar circumstances. This assumption was the ground of the bilingual character of the system.

The system as a whole was tested on technical manuals. However, this is not a restriction imposed by the anaphora resolver engine, but of the pre-processing. There are no rules that are specific to a certain text genre to such an extent as to make the system unusable on other types of text. The choice of technical manuals was based on very practical considerations, holding mostly to the availability of data, as elaborated

[8] In this context, *equivalent* is used for anaphoric expressions that are either translations of each other or used in the same context in different languages

in chapter 4.

1.5 Terminology

Most concepts will be defined as the need arises. This section will define some terms that are most frequently used in the thesis.

The notion of **reference** can be defined in two ways. It can be said to be the symbolic relationship that a linguistic expression has with the concrete object or abstraction it represents. [Halliday and Hasan1976] define reference as the relationship of one linguistic expression to another, in which one provides the information necessary to interpret the other:

> There are certain items in every language which have the property of reference [...]; that is to say, instead of being interpreted semantically in their own right, they make reference to something else for their interpretation.

[Milner1982] differentiates between *actual reference* and *virtual reference*, stating that:

> The segment of reality associated with an expression is its actual reference; the body of conditions characterising a lexical unit is its virtual reference.

Following these definitions, the *virtual reference* of a linguistic element is independent of its use in a context, while the *actual reference* can only be determined from the use of that linguistic element in a discourse.

Anaphora is the central notion of this research. In its broad meaning, it designates the relationship that holds between a pro-form called **anaphor** and the entity it points to, called **antecedent**.

The term **possible antecedent** for a pronoun will be used to designate those noun phrases that can theoretically function as antecedents for the pronoun. In most cases, this term will subsume all noun phrases in the accessibility space of the pronoun.

Cataphora is a similar concept to **anaphora**, with the difference that the referential expression precedes its antecedent.

Endophora subsumes the notions of anaphora and cataphora.

Although strictly speaking the terms **anaphora/anaphor** and **cataphora/cataphor** do not overlap, the term **anaphora** will be used to designate both phenomena (therefore endophora), unless they need to be clearly differentiated.

Coreference holds between two entities that refer to the same real world entity. The relation of coreference is an equivalence relation, which partitions the space of coreferential expressions into equivalence classes, called **coreferential chains**[9].

Automatic processing refers to computer execution without any kind of human intervention. This term will mostly be used with the same meaning as **fully-automatic processing**, as opposed to **semi-automatic processing**, which presupposes some kind of human intervention.

Machine learning and **automatic learning** will be used interchangeably to designate the operation of programs or algorithms that learn from experience.

1.6 Aims of the research

The main aim of this research is to design and develop an original method for pronoun resolution in a multilingual environment. The requirements for such a system are to be able to handle at least two languages (English and French) and to be easily extended to other languages.

[9]For each coreferential expression, there is exactly one coreferential chain that contains it

The research presented in this thesis was conducted according to a number of targets, listed below. They are listed in the logical order given by the progress of the research, therefore they do not always follow the order in which they are fulfilled in this thesis.

- 1.1 - overview of the work done so far in the field
- 1.2 - identify the drawbacks and the achievements of existing anaphora resolvers and identify areas that can be improved
- 2.1 - identify the common and distinct features of pronominal systems in English and French
- 2.2 - identify the largest set of common rules for English and French
- 3.1 - identify features to be used in automatic anaphora resolution
- 4.1 - decide the text genre and identify suitable texts
- 4.2 - collect and annotate data for use in training and evaluation
- 5.1 - assess the suitability of machine learning techniques in the resolution of anaphora
- 5.2 - design a system for bilingual anaphora resolution
- 5.3 - select the most suitable learning strategy for the system
- 6.1 - design an evaluation strategy
- 6.2 - develop an evaluation workbench for pronominal anaphora resolution systems

- 6.3 - perform extensive quantitative and qualitative evaluation of the aforementioned system
- 6.4 - assess the possibility of extending the system for other languages
- 7.1 - investigate the possibility of mutual enhancement within a multilingual anaphora resolver
- 8.1 - identify the main difficulties in automatic anaphora resolution
- 8.2 - perform a corpus based analysis of the most common errors

1.7 Outline of the thesis

The thesis is organised into seven chapters, that approach multilingual anaphora resolution gradually: from theoretical foundations (chapter 2) to practical approaches (chapter 3), data annotation (chapter 4), description of an original pronoun resolution system (chapter 5), evaluation (chapter 6), assessment of difficulties in automatic pronoun resolution (chapter 7) and concluding remarks. Each chapter addresses one or more aims of the research.

Chapter 2 has a mostly theoretical content. It is a non-exhaustive overview of pronominal systems of English and French. The focus is on those pronouns that will be tackled by the anaphora resolver. The main aim of this chapter is to identify common features of different classes of pronouns and to single out differences that the two main languages may present in referential constructs.

Although pronominal anaphora is a complex linguistic and semantic phenomenon, this chapter only concentrates on those aspects likely to provide support for the practical task of resolving pronouns. In this respect, the discussion will start from the

surface form of pronouns, moving towards the possible discourse function that such an expression may hold. It will be shown which linguistic and contextual features can be exploited in order to identify an expression as anaphoric or not and, in the first case, to filter its potential antecedents and to identify the correct one.

The following three chapters concentrate on more practical aspects relating to anaphora resolution.

Chapter 3 presents anaphora resolution from a technical viewpoint. It concentrates on the mechanisms and sources of information necessary for solving anaphora automatically. It shows how this automatic processing is done and which tools are available for performing different pre-processing tasks. This chapter also contains a critical overview of the most well known systems for automatic anaphora resolution in English and French, as well as the existing attempts of approaching multilingual anaphora resolution.

Chapter 4 discusses the use of corpora in anaphora resolution and particularly in the application presented in the next chapter; it explains the choice of a particular data set and presents some analysis of the data. The type of data used for training is described, the choice of the corpus is explained and the type of annotation performed is presented.

Chapter 5 describes ARMAL, an original system for anaphora resolution based on a combination of high confidence handcrafted rules and machine learning techniques. This chapter briefly outlines the theoretical fundamentals of machine learning techniques and argues that anaphora resolution is a suitable task for automatic learning. The original system ARMAL is described, explaining the choice of the

learning features, the learning method, and the general execution flow. The last section of this chapter presents a cheaper and faster alternative to the main machine learning method, a genetic algorithm. It elaborates on the advantages and disadvantages of such a method (which will become more apparent in Chapter 6) and explains why, despite its clear advantages, the method cannot be used successfully in pronoun resolution.

Chapter 6 is dedicated to the problem of evaluation in anaphora resolution. It begins with discussing current issues in evaluation methodologies and continues with the description of the evaluation strategy employed. It also describes the workbench used for comparative evaluation. The systems described in Chapter 5 (both ARMAL and the genetic algorithm) are thoroughly evaluated using a series of different techniques and several types of evaluation results are presented. The performance of each of the languages involved is assessed separately and classes of errors produced by the system are identified.

Chapter 7 discusses difficulties in multilingual anaphora resolution. The discussions and results presented in this section are derived from practical experiments with several pronoun resolution algorithms. First, the main types of general errors are identified and a corpus based analysis of their frequency is performed. Secondly, the problem of morphological disagreement is studied in depth and a large scale corpus analysis is presented, showing the distribution and frequency of the phenomena across different text genres. Thirdly, errors specific to each language investigated are identified.

The **Conclusions** chapter contains a brief overview of the previous chapters and identifies the way the aims of the thesis have been fulfilled.

Chapter 2

Anaphoric expressions in English and French

2.1 Overview

The Collins dictionary of the English language defines a pronoun as ”one of a class of words that serves to replace a noun phrase that has already been or is about to be mentioned in the sentence or context”[1]. This definition describes in fact the most common use of pronouns, that of anaphoric and cataphoric reference to a noun phrase, but ignores other types of anaphoric reference that appear in naturally occurring utterances[2].

It has been acknowledged the fact that different types of anaphoric expressions require different kind of processing in order to be resolved. If this assertion is true for English, it is even more important when trying to resolve anaphora in a multilingual

[1] Collins English Dictionary, 21st Century Edition, pp.1237

[2] *Le Grand Larousse de la Langue Française* captures the more complex definition of a pronoun: ”mot qui peut representer un mot exprimé à un autre endroit de l’enoncé (nom, adjectif, autre pronom, phrase) ou designer un des participants a la communication” (”a word which can represent another word expressed at another place in the discourse (noun, adjective, pronoun, sentence) or designate one of the participants in the communication”)

environment, where each language involved adds an extra dimension to the treatment of various anaphoric forms.

It is beyond the aim of this thesis to provide a classification of anaphoric expressions, or to study the factors that determine the use of different anaphoric expressions in context. This chapter approaches bilingual pronoun resolution from a theoretical point of view, however keeping in mind the practical applicability of the theoretical findings. Although pronominal anaphora is a complex linguistic and semantic phenomenon, this chapter will only concentrate on those aspects likely to provide support for the practical task of resolving pronouns. There are many ways of categorising anaphoric relations: one can start from the syntactic function of the anaphor, from the type of the antecedent, or from the type of the relation. The present discussion will be conducted starting from the most reliable information available a priori to an automatic anaphora resolution system, which is the surface form of the pronouns[3] and will then move towards the possible discourse function that such an expression may hold. It will be shown which linguistic and contextual features can be exploited in order to identify an expression as anaphoric or not and, in the first case, to filter its potential antecedents and to identify the correct one. The idea is to identify the common and distinct features of different anaphoric expressions in English and French in order to construct clusters of anaphoric expressions that can benefit from a similar treatment.

[3]This chapter will not cover however cases of polysemous words, such as the French *l'*, which can be either a definite article or a pronoun. Such cases are supposed to be disambiguated at the lexical analysis level. This discussion assumes that the lexical category of pronouns has been determined prior to attempting any further analysis.

2.2 Pronominal systems

The discussion will concern four classes of pronouns: personal pronouns, possessives (possessive determiners as well as pronouns), reflexives and demonstratives. This reflects the main categories of pronouns that display anaphoric properties and that will be tackled by the pronoun resolution system described in Chapter 5.

2.2.1 Personal pronouns

From the point of view of their frequency of use in written and spoken language, as well as the variety of their anaphoric usages, personal pronouns represent the most important class. English personal pronouns are marked for person, number, case, and third person singular pronouns are additionally marked for gender and animacy. Therefore, the categories that can be distinguished are those listed in the table below[4].

Case	Singular			Plural
	F	M	N	
Subjective	she	he	it	they
Objective	her	him	it	them

Table 2.1: English personal pronouns

French pronouns are morphologically marked for person, gender (masculine, feminine, and neuter), number and case (nominative, accusative, dative, genitive, reflexive), as shown in table 2.2.

Not included in table 2.2 is a French specific personal pronoun, the generic *on*. This pronoun can only be used as a subject, and, although grammatically always a

[4] As discussed before (see Chapter 1), first and second person pronouns are excluded from the investigation.

Case	Singular			Plural		
	Masc	Fem	Refl	Masc	Fem	Refl
Nominative	il	elle	se	ils	elles	se
Accusative obj.	le, l’	la, l’	se	les		se
Indirect obj.	lui		se	leur		se
Oblique	lui		soi	eux	elles	soi

Table 2.2: French personal pronouns

third person masculine singular pronoun, it can designate one or more persons. It can replace either of the first, second or third person pronouns. From a practical point of view, this raises the problem of deciding whether, in a given context, *on* is anaphoric or not. Furthermore, the retrieval of the antecedent of *on* does not benefit from the morpho-syntactical features involved in the resolution of other personal pronouns.

2.2.1.1 Anaphoric usages of personal pronouns

In the following, the terms *anaphora/anaphoric* will be used for describing both types of endophoric reference (anaphora and cataphora), whenever unambiguous, as the distinction between these two types of reference is based solely on the position of the anaphor relative to the antecedent, and has no semantic significance.

The discussion on the anaphoric usages of personal pronouns will start with the following examples:

(2.1) *Lord Romsey* promised *he* would spend 9m on urgent restoration work on his estate.

(2.2) *The second car* hasn’t got time to avoid the person who swerved away to avoid the car that was pulling out and *he* hits it.

(2.3) It can be fun to *raise money for your favourite charity*. You can do *it* on your own or you can get together with family and friends.

These utterances exemplify the main types of pronominal reference. Examples (2.1) and (2.2) feature pronouns referring to textual antecedents in the form of noun phrases. In the third example, the pronoun points back to the previous clause. In the following, the peculiarities of each of these cases will be analysed in more detail.

Pronouns with noun phrase antecedent

Anaphoric pronouns most frequently refer to a noun phrase antecedent. Example (2.1) displays a typical case of **coreference** between the personal pronoun and a noun phrase. *He* can be substituted by *Lord Romsey* with no loss of meaning. Using a pronoun in this case fulfils a stylistic and cohesiveness role, by avoiding the repetition of the full noun phrase. This type of anaphoric usage is the most common for personal pronouns. Coreferentiality with a noun phrase implies agreement in gender and number with the antecedent, with the restrictions discussed in section 2.2.6.

Example (2.2) displays a situation of **bridging** anaphora, where the pronoun *he* can only be interpreted by assuming the knowledge that each car has a driver. The pronoun cannot be directly substituted to any of the textual noun phrases previously appearing in the text, and its interpretation involves a certain amount of mental inferences.

As a general rule, pronouns can refer to any type of noun phrase in a text, including full noun phrases or other pronouns[5].

French pronouns display restricted accessibility to certain expressions. Some of these cases of restricted accessibility have been analysed by Kayne [Kayne1972],

[5]It is not being discussed at this point *the likelihood* of a noun phrase to be referred to by a pronoun, but only *the ability* of a pronoun to refer a noun phrase.

who showed that certain morphemes or expressions can serve as antecedent for *il* in a complex inversion construction, but not when *il* is not permuted. The phrases under discussion are *cela*/*ça*, *ce que, tout, rien*. In the example 2.4 below, (a) is a correct construction featuring a complex inversion, while (b) is unacceptable.

(2.4) (a) Pourquoi cela est-il faux?
(Why is this false?)
(b) Ce que tu dis interesse Max parce-qu'il concerne la linguistique. (*)
(What you are saying interests Max because it concerns the linguistics.)

Similar to *il*, the disjoint *lui* cannot take *cela* as antecedent either. A special case is the short form of the accusative third person pronoun, *l'*, which can refer to *cela*, as in example 2.5:

(2.5) (a) Cela est bienconnu depuis que Max l'a démontré.
(This is well known since Max proved it.)
(b) Cela est bienconnu depuis qu'il a été démontré.
(This is well known since it was proven.)

In French, a special case is represented by the personal pronouns *me, te, se, nous, vous* acting as direct or indirect object for pronominal verbs (for example, *se laver, se reveiller, se cacher*). In these situations, the pronouns always refer to the same object as the subject of the verb and they express reflexivity.

Referential pronouns with non-nominal antecedent

In English, the neuter personal pronoun in singular, *it* can often refer anaphorically to a clause, sentence, or sequence of sentences, as in the examples below:

(2.6) I *go to the gym every day* and I enjoy *it*. (*it* refers

anaphorically to a clause)

(2.7) Switzerland has lots of Alps, but *they are now less densely equipped with skiers' dormitories than those of France and Austria. Many of the big names of Swiss skiing have fewer than 10,000 beds; the total for Val d'Isere/Tignes approaches 50,000.* But *it* is more a consequence of the view we have of Switzerland and of Swiss skiing. (*it* refers to a sequence of sentences)

As opposed to the English *it*, its French counterpart *il* is not capable of anaphorising a sentence; this role is played instead by the demonstratives *ça/cela*, as it will be shown later in the discussion on demonstrative pronouns (section 2.2.3).

French displays a particular feature, where the neutral pronoun *le/l'* can refer an adverbial or an attributive phrase:

(2.8) Ils m'accusent d'etre *fière*, et j'admitte que je *le* suis.
(They accuse me of being *proud* and I admit that I am *so*).

In modern French, the lexical form of the pronoun does not vary, even if the phrase anaphorically referred to is feminine or plural:

(2.9) Nous sommes *amis* et nous *le* seront toujours
(We are friends, and we will always be.)

A particular situation is represented as well by cases where the neutral clitic *le* refers to a prepositional phrase with attributive role:

(2.10) Les ouvriers du port du Havre sont *en grève*, et ils *le* seront pendant un long moment.
(The workers at the Havre docks are on strike, and they will continue to be ∅ for a long while.)

2.2.2 Possessive pronouns and determiners

Possessive pronouns refer to noun phrases only, adding the meaning of possession. English possessives preserve the gender and number of the possessor. Therefore, the same division of forms applies, as noticed for personal pronouns: masculine and feminine singular pronouns and a unique gender plural pronoun.

Sg		Pl
M	F	
his	her, hers	their, theirs

Table 2.3: Morphological features of English possessive pronouns and determiners

As opposed to English, the lexical representation of French third person singular possessives does not offer any indication of the gender or number of the possessor, but only of the possessed object. In this case, the only morphological clue in the identification of the antecedent is the number. See table 2.4 for a full description of French possessive pronouns.

Pronoun	Determiner	Gender of object	Possessor's number	Object's number
le sien	son	M	sg	sg
la sienne	sa	F	sg	sg
les siennes	ses	F	sg	pl
les siens		M	sg	pl
le/la leur	leur	M/F	pl	sg
les leurs	leurs	M/F	pl	pl

Table 2.4: French possessives

2.2.3 Demonstratives

Demonstrative pronouns are used for indicating objects or events in the proximity of the speaker (either spatial proximity, or proximity of thought).

Both English and French demonstratives display different linguistic forms according to the proximity of the object they indicate. The Comprehensive Grammar of the English Language [Quirk et al.1985] defines the reference to an object close to the speaker as *near reference* and the reference to an object far from the speaker as *distant reference*. The only other distinction English makes is between the singular and plural forms of the demonstratives (*this/these, that/those*), while in French they are additionally distinguished by gender (masculine, feminine and neuter). French also displays a generic form of demonstrative (celui, celle, ce), that does not indicate the proximity, and has a restricted use.

	English		French				
	sg	pl	sg			pl	
			m	f	n	m	f
near reference	this	these	celui-ci	celle-ci	ceci	ceux-ci	celles-ci
distant reference	that	those	celui-là	celle-là	cela, ca	ceux-là	celles-là
generic	n/a		celui	celle	ce	-	

Table 2.5: Demonstrative pronouns

The basic referential function of demonstratives is that of deictic reference, which will be discussed in section 2.3.1.2. When demonstratives are endophoric, their antecedent can be either a noun phrase or a larger discourse segment: clause, sentence, a sequence of clauses. This larger segment is called *sentential antecedent*.

When used with any other grammatical function but subject, demonstratives can usually refer anaphorically to inanimated objects only. As a subject, a demonstrative can refer cataphorically to a person as well.

(2.11) *This* is *Mr.Jones*.

The near demonstratives can have both anaphoric and cataphoric references:

(2.12) He asked for *his brown raincoat*, insisting that *this* was his usual coat during the winter months.[6] (anaphoric)

(2.13) He told the story like *this* : *"..."* (cataphoric)

Under normal circumstances, distant demonstratives can only have anaphoric reference, as in the following example:

(2.14) I hear you disliked his latest novel. I read *his first novel*, and *that* was boring, too.

There are however exceptions where, in very limited contexts, distant demonstratives can refer cataphorically, e.g. in expressions of indignation:

(2.15) What do you think of THAT?! Bob smashes up my car, and then expects me to pay for the repairs![6]

Demonstratives have a nominal function when post-modified by a relative clause or other restrictive modifiers:

(2.16) (a) Those who try hard deserve to succeed.

(b) Ceux qui essaient mèritent de reussir.

Only generic demonstratives can be used in French in these contexts, as the French grammar traditionally prohibits the post-modification of nominal demonstratives by any other constructions but relative clauses and adverbials introduced by the particle *de*. Nevertheless, in informal French this rule is often ignored, simple demonstratives being often modified by participles or adverbials:

[6] Examples 2.12 to 2.15 are from [Quirk et al.1985].

(2.17) Tous *ceux* ayant la même maladie...

(All those suffering from the same illness...)

(2.18) Je joins a ma lettre *celle* écrite par le prince.

(I'm attaching to my letter the one written by the prince.)

(2.19) Cette remarque ainsi que toutes *celles* purement grammaticales...

(This remark, as well as all those purely grammatical...)

In English, demonstrative pronouns can be used to anaphorically refer to a verb phrase, clause, or sentence. They are an alternative to the personal pronoun *it*, carrying an extra stress on the anaphor, that the personal pronoun cannot convey:

(2.20) He said *he is going to visit me tomorrow*. I'd like *that*.

(2.21) In 1939, *Germany invaded Poland*. *This* marked the beginning of the Second World War.

As it has already been mentioned before, in French, only the demonstratives *ça*/*cela* can refer to a sentence. In the following example, the variant (a), which features the person pronoun *il*, is unacceptable, the correct construction being (b), which employs a demonstrative:

(2.22) (a) Elle viendra. Il me fait plaisir. (*)

(b) Elle viendra. Ça me fait plaisir.

Same asLike *cela*, *ça* can be used sometimes as deictic, some times as anaphoric or cataphoric.

(2.23) Max a glissé sur une peau de banane. Eve trouve ça drôle.

(Max slipped on a banana skin. Eve finds this funny.)

The use of *ça* as non-anaphoric will be discussed in section 2.3 in the larger context of non-anaphoric pronouns.

2.2.4 Reflexives

Reflexive pronouns indicate the fact that the action of the verb is reflected upon the subject.

From a lexical point of view, English reflexives are constructed by adding the *self* suffix to the accusative form of the personal pronouns (*himself, herself, themselves*). French reflexives are constructed by adding the *même* or *mêmes* suffix to the nominative or accusative form of the personal pronouns (*elle-même, lui-même, nous-mêmes, eux-mêmes*). Contemporary dictionaries and grammars of the French language describe *lui-même* as a form of underlining, including the element *même* which, when following a noun, ”underlines the identity of the object spoken about”[7] and ”linked by an union feature to a disjoint personal pronoun,(...) insists on the identity of the person”.

Reflexives lack independent meaning, therefore they must have an explicit antecedent, which can be either a full noun phrase or a pronoun. The referential property of reflexives is regulated by the theory of Government and Binding, which states that reflexives must be bound in their governing domain. The basic reflexive pronoun corefers with the subject of its clause (overt or implied).

Reflexives can be used either with an emphatic function (underlining the identity of the subject) or with an objective function (the agent and the object of the action indicated by the verb is reflected). The examples below illustrate the two possible usages of reflexives:

(2.24) (a) Peter tuned the piano himself.

(b) Peter a accordé le piano lui-même.

(2.25) (a) Peter is ashamed of himself.

(b) Peter a honte de lui-même.

[7] Le Grand Larousse de la Langue Française, Paris : Librairie Larousse, 1978

In French, reflexives cannot appear as objects in the accusative or dative, where the language usually prohibits non-clitic pronouns:

(2.26) (a) Pierre déteste lui-même. (*)

(b) Pierre se déteste.

This contrasts with the similar English constructions, where reflexive verbs always require a reflexive object:

(2.27) She always prides herself on her academic background.

2.2.5 French indefinite pronouns

The French indefinite pronouns *en* and *y* represent a special category with no counterparts in English. Etymologically adverbials, *en* and *y* have acquired the quality of pronouns equivalent to *de lui, d'elle, d'eux, de cela* and *a lui, a elle, a eux, a cela,* respectively.

En and *y* are not marked for gender, number or person, therefore in certain contexts they can have several possible antecedents, which can be a source of ambiguity. The only linguistic feature that could be used in the tracking of the antecedent of an indefinite pronoun is the case.

[Grevisse1969] states that, when used as pronouns, *en* and *y* represent more often animals, objects or concepts. However, grammarians often wondered if, and under what circumstances *en* and *y* can have human antecedents. [Sanford1965], [Pinchon1972], [Togeby1982], [Rouwet1990] noticed that these pronouns can have human antecedents in any kind of context, they can even have antecedents of first and second person. All sentences below represent examples of perfectly acceptable constructions, and capture various situations where *en* can refer to an animated antecedent:

(2.28) Jean-Jacques a presenté *Emile* a Sophie. Elle est tout de

suite tombée amoureuse de lui. / Elle *en* est tout de suite tombée amoureuse.

(Jean-Jacques introduced Emile to Sophie. She instantly fell in love with him.)

(2.29) Il m'a parlé *de vous* comme je veux qu'on m'*en* parle.

(He talked to me about you as I want them to [talk to me about you].)

(2.30) Ne parlons pas *de moi*. Mais si, parlons-*en*.

(Let's not talk about me. But yes, let's talk [about you]).

However, the ability to refer to animated entities seems to be restricted, since the following represent unacceptable constructions:

(2.31) *Max* espère qu'on *en* parlera a Apostrophe.

(*) Max hopes that they will speak about him to Apostrophe.

(2.32) *Emile* croit que Sophie *y* pense. (*)

(Emile thinks that Sophie thinks of him).

It looks as if *en* and *y* are not generic types of pronominals since they cannot be bound in their minimal governing category. They obey additional constraints, which hold to the syntax of the sentence. Beatrice Lamiroy proposes a formal syntactical explanation in the framework of the GB theory [Lamiroy1990, Lamiroy1991a, Lamiroy1991b]. Inevitably, she notices that there are several types of syntactical contexts where *en* and *y* can have an antecedent in the same complex sentence, therefore her theory fails in quite a number of situations. The following sentences are perfectly acceptable:

(2.33) Il les aimait et il en était aimée.

(He loves them and was loved by them.)

(2.34) Le precepteur d'Émile croit que Sophie en est amoureux.

(Emile's professor thinks that Sophie is in love with him [with Emile].)

(2.35) Ce texte a l'interprétation que Marie en donne.

(This text has the interpretation that Marie gives it.)

(2.36) Comme Emile est beau garcon, Sophie en est tombée amoureuse.

(As Emile is a handsome boy, Sophie fell in love with him.)

In (2.33) *en* has its antecedent in a coordination, in (2.34), inside the subject. In (2.35), the antecedent is the subject, but *en* is inside a relative clause. In (2.36), the antecedent is in an antepositioned circumstantial clause (the opposite is also possible, "Comme elle *en* est amoureuse, Sophie trouve *Emile* plus beau qu'il ne l'est.")

The generic pronoun *on* acts with respect to *en* and *y* in the same way as the personal subject pronouns *je*, *tu*, *nous*. Therefore, in the following constructions, *on* can only act as antecedent for *soi*, and not for *en* or *y*.

(2.37) (a) *On* souhaite rarement que les autres disent du mal de *soi*.

(b) *On* souhaite rarement que les autres *en* disent du mal.(*)

(2.38) (a) *On* mérite rarement que les autres disent du bien de *soi*.

(b) *On* mérite rarement que les autres *en* disent du bien.(*)

In the same type of context, *personne* and *tout le monde* can be successfully used as antecedents for *en*.

(2.39) *Personne* ne mèrite jamais qu'on *en* dise tant de bien.

(2.40) *Tout le monde* mèrite qu'on *en* dise du mal.

Pragmatically speaking, Rouwet's findings do not bring an important contribution to the automatic interpretation of *en* and *y*, due to the impossibility of automatically extracting the semantic content of the constructions.

2.2.6 Grammatical features of anaphoric pronouns

An intuitive assumption that forms the basis of a number of strong constraints in anaphora resolution is that anaphorically linked elements must share similar morphological and semantic features, therefore they must agree in their morphological features. One cannot generally talk about morphological agreement between a pronoun and its antecedent unless they corefer. As a general rule, whenever a pronoun varies in number, gender or person, it agrees in number, gender and person with its coreferential noun or pronoun. However, even in coreference cases situations arise when the coreferential elements do not agree in number or gender.

One of the most frequent cases of gender disagreement in French is due to the differences between grammatical and natural gender. This is a situation common to many languages, including the Indo-European languages[8], which arises when the grammatical gender, assigned arbitrarily, does not match the natural gender of an entity, determined by the properties of the referent. In the example below, *la sentinelle* (*the guard*) has the grammatical gender feminine, but can be referred to by a masculine pronoun if the noun phrase is value loaded and the person referred to is a male:[9]

(2.41) *La sentinelle* était fatiguée. *Il* avait passé toute la nuit en garde.

(*The guard* was tired. *He* had spent the whole night on duty.)

However, if the noun phrase is value free, any pronoun used to refer to it will have to have the same grammatical gender:

(2.42) *La sentinelle* ne doit pas quitter son poste. *Elle* doit y rester à tout temps.

[8] See [Corbett1991] for an extensive survey of genders and gender systems, based on more than 200 languages

[9] Most cases of gender mismatches discussed in the literature concern animate entities

(*The guard* must not leave their post. *They [he or she]* must stay there at all times.)

A similar problem does not arise in English, which does not differentiate nouns by grammatical gender[10]). In colloquial English, there is a certain bias towards using masculine or feminine pronouns according to the prejudice that certain jobs are more likely to be performed by men, respectively women (for example, a doctor is referred to as *he*, while a nurse is referred to by *she*). This situation arises when there is no indication on the gender of the person, and it is usually advisable to avoid the use of prejudiced pronouns by employing alternative constructions instead: *he or she*, *they*, passive voice constructions. Far from making the task of pronoun resolution easier, the use of these constructions introduces a new dimension to the difficulty of the task.

The main problem in finding the antecedent of the phrase *he or she* is its identification as a phrase referring to a single individual, as opposed to two pronouns each referring separately to an individual. Consider the following example:

(2.43) The driver ran from the scene of the accident. The police are holding an inquiry into his or her identity.

(2.44) A deduction is made from *the wife's* or *husband's* insurance benefit to which *he or she* is entitled for any month if he or she is under age 65.

The first example is a clear case of gender underspecification, where *the driver* is referred to by *he or she* in order to avoid overspecifying the gender. In the second example, both *he* and *she* have retrievable antecedents in the text.

[10]Some animate entities however indicate their natural gender through distinct lexical forms. Some names of animals indicate the gender, such as *fox/vixen*, *lion/lioness*, *dog/bitch*, however the masculine form is usually used to denominate an individual belonging to a certain species; some professions are also marked for gender by lexical variations: *actor/actress*, *waiter/waitress*; some nouns have independent, unrelated lexical forms depending on their natural gender: *boy/girl*, *husband/wife*, *son/daughter*. In English, no inanimate entities are marked for gender.

Using a plural pronoun in similar cases raises the problem of number mismatch between the pronoun and its coreferent.

(2.45) Ask *the user* about *their* requirements.

Number disagreement between a pronoun and its antecedent is even a more frequent phenomenon. In a recent experiment [Barbu et al.2002] a corpus based analysis was performed, investigating the occurrences of number disagreement cases. The experiment, performed on sample texts from the British National Corpus containing about 2000 plural pronouns showed that approximately one third of them did not display number agreement with the antecedent, or required semantic inferences for the identification of the semantic number of the antecedent. In the same paper, seven main categories of number mismatch cases were identified. Although the investigation only targeted English texts, the same cases seem to be valid for French as well, but no statistics on their frequency is available. A detailed description of the experiment, as well as results and statistics will be presented in Chapter 7.

2.3 Non-anaphoric pronouns

In English, as well as in French, pronouns can be used with a non-anaphoric function. From a practical point of view, these pronouns have to be identified in order to avoid attempts to resolve them to a noun phrase or verb phrase in the text. Such a wrong resolution may not reflect on the evaluation of the anaphora resolver independently (as it will be seen in Chapter 6, most evaluation measures do not account for such errors), but it may affect applications where anaphora resolution is applied (machine translation, information extraction).

2.3.1 Classes of non-anaphoric pronouns

2.3.1.1 Deictic reference

Deixis is reference by means of an expression whose interpretation is relative to the (usually) extralinguistic context of the utterance, such as who is speaking, the time or place of speaking, the gestures of the speaker, or the current location in the discourse. Resolution of deictic reference is beyond the scope of this project, however it is important to distinguish cases of deixis in order to be able to eliminate them from the resolution process.

Demonstrative pronouns represent typical examples of deictic reference. By their own definition, they can be employed to indicate objects in the environment in which the dialogue is taking place, or the ”context of situation” [Halliday and Hasan1976].

(2.46) Look at *that*!

(2.47) I don’t like *this*.

The same applies to French demonstratives, that can refer deictically more often than personal pronouns.

(2.48) Tu as acheté quelque chose? Oui, j’ai acheté *ça*.

(Did you buy anything? Yes, I bought this.)

Deictic reference is much more common in dialogue then in expository text, therefore it is less important in the context of this project.

2.3.1.2 Non-referential pronouns

Impersonal constructions are characterised by the fact that the subject, although instantiated by a pronoun, is deprived of all referential content.

For Chomsky, it is ”a matter of linguistic fact” that there is no possible value or *denotatum* for the zero or expletive subject of verbs like ”to rain”.

Chomsky and the adepts of GB have oscillated between different analyses of the impersonal subject. Other linguists were forced to reduce such special cases to the general case, to show that, despite appearances, such phrases have the internal structure of a normal sentence, with a predicate and at least one argument. For Chafe [Chafe1970], *it* in *it rains, it is late, it is hot*, refers to ”an all-encompassing state”. Some linguists, among whom most notably Dwight Bolinger [Bolinger1977] sustain that the expletive *it* in ”it’s raining” has a referential value, and an independent meaning. According to Bolinger, ”every word that a language allowed to survive must bring a semantic contribution”. Bolinger builds his argumentation around examples such as: ”It’s so hot that it’s giving me a headache”, claiming that *it* in the first clause has to be *it* in the second clause, therefore anaphoric. This argument is debatable, as the *it* in the secondary clause is more likely to refer to the event in the main clause, than to the impersonal subject.

Other linguists admit the expletive character of *it* but reconstruct at certain sub-adjacent levels of representation, a canonical prepositional structure.

Pleonastic (or expletive) pronouns can appear in both subject (example 2.57) and object (example 2.58) position, with a higher frequency of occurence as subject.

(2.49) *It* snows. / *Il* neige.

(2.50) I hate *it* when it snows on my toast.

Pleonastic pronouns appear in a quite restricted number of situations:

1. **Prop-it constructions**, i.e. as subject of actions that have no existential meaning. Some of the most typical such constructions are:

- Meteorological verbs: *It rains, it snows / il pleut, il neige*
- Expressions indicating the time: *it is late, it is two o’clock / il est tard, il est deux heures*

- Expressions indicating a characteristic of the ambient: *it is hot* / *il fait chaud*

- Expressions of necessity, importance or urgency: *il est important de* , *il est neccessaire de* /*it is important to, it is essential to*

2. **Cleft constructions**, i.e. constructions that allow stressing the idea/concept expressed by a constituent by using a dummy pronoun *it* in subject position and moving the focal constituent after the verb.In these cases, pronouns are used in constructions like ”PRON + to be”, ”ce/c’ + to be”.

(2.51) (a) Now it is they who are trying to manipulate us.

(b) Maintenant c’est eux qui essaient de nous manipuler.

3. **Idiomatic constructions**

If in examples featuring cleft constructions it could be argued that the pronoun *it* has indeed a referential value, referring cataphorically to the following clause or NP, there are idiomatic constructions where *it* is void of any semantic content, as in the following examples:

(2.52) The quicker a site is provided, the better *it* will be for everyone.

(2.53) You know how *it* is with these developments.

In French as well, there are numerous idiomatic constructions of the type ”ça +verb”: *ça marche, ça va, ça colle, ça barde, ça ne rate pas* - where *il* cannot be substituted to *ça*. In these contexts, *ça* can refer to something that is further explicated, as in: ”Comment ça va, la santé/le travail...” raising the question if *ça* is anaphoric or not in these contexts.

2.3.2 French-specific non-anaphoric constructions

The French impersonal *il* has the same form as the third person singular pronoun in masculine, but it is never anaphoric, nor has it a strong form (*lui*) associated. The impersonal *il* is mandatory in active sentences, but its counterpart in the accusative, *le* is unacceptable in non-finite sentences. Therefore, ”Je regarde tonner” is correct, while ”Je le regarde tonner” is not.

In English, however, *it* is always compulsory in similar constructions:

(2.54) I watch *it* rain.

(2.55) I heard *it* thunder.

(2.56) Only the gods can make *it* rain.

In French, some other expressions can be constructed with the expletive *il*:

- il y a + NP

  (2.57) Il y a vingt ans dès que je l’ai vu.

- expressions in ”faire+adjective”: *il fait beau, il fait mauvais*

- expressions in ”faire+noun”: *il fait jour*

Just as in the case of meteorological verbs, these expressions do not admit *le* as subject of non-finite sentences. But, as opposed to ”pleuvoir”, they do not admit a zero subject either.

(2.58) (a)Je crois qu’il est trop tot pour partir.

(b) Je le crois trop tot pour partir.(*)

(I think it is too early to leave.)

(2.59) (a) Je sens qu’il fait grand vent.

(b) Je le sens qu’il fait grand vent.(*)

(I feel there is strong wind.)

In contexts where only *it* can be used in English, in French we can have both *il* and *ça*, for example: *ça pleut, ça barde, ça brouillasse*. *Ça* is a contraction of *cela* that belongs to the familiar style. There are therefore incompatibilities between verbs used in the formal style and *it* and verbs used in the casual style.

Ça can be also used as impersonal, in three types of contexts:

1. Meteorological verbs: *Il pleut, ça pleut, ça brouillasse*
2. With verbs that can only take *ça* as a subject: *ça barde, ça balance*
3. In constructions featuring a sentential argument in sequence:

(2.60) *Il* m'ennuirait beacoup qu'il ne vienne pas. (*)
Ça m'ennuirait beaucoup qu'il ne vienne pas.
(It would annoy me a lot if he didn't come.)

[Corblin1995] makes a distinction between the impersonal uses of *il* and those of *ça*, calling the latter *indistinct reference*. Cadiot [Cadiot1988] investigates the differences between the anaphoric roles of *il* and *ça*. In his opinion, personal pronouns refer to "nominated occurrences" (individuals or classes of individuals) corresponding to "stable cognitive categories". *Ça* does not directly refer to individuals or classes of individuals; it anaphorically refers any expression associated in a discourse to a nominal object. One could say that *ça* refers to its antecedent indirectly, by means of a certain idea that could be associated with that one - taking into account the discourse context and situation.

2.4 Conclusions

The aim of this chapter was to describe the possible anaphoric and non-anaphoric uses of personal, possessive and demonstrative pronouns in both English and French. This investigation led to some interesting observations that may help in the two stages

of automatic pronoun resolution: filtering (rejecting elements that cannot possibly be referred to by a pronoun) and identification of antecedent. We have noticed some cases where the same type of pronoun presents different degrees of accessibility to discourse elements in English as compared to French. This implies that different language specific rules have to be applied in some cases. However, the majority of anaphoric expressions, or at least the most frequently used, have the same behaviour in English and in French, which makes it possible to use the same devices for their resolution.

Chapter 3

Automatic anaphora resolution

3.1 Overview

A computational treatment of anaphoric pronouns has to take into account factors that have been identified by linguists and psycholinguists and model them into a machine-readable format. While, as shown in the previous chapter, different languages have different ways of expressing relationships between discourse entities, there are numerous common factors that contribute to the interpretation of anaphora. It is on these common factors that this chapter will concentrate, identifying those that can be processed automatically. It will be shown how this automatic processing is done and which tools are available for performing different pre-processing tasks. A review of the most important categories of anaphora resolvers developed so far for English and French will also be presented. While there is an important number of original approaches for individual languages, and especially for English, multilingual pronoun resolution can still benefit from an integrated approach.

3.2 Knowledge sources

3.2.1 Types of knowledge sources

Identifying the factors that contribute to the understanding of anaphors has been a topic of interest for linguists, psycholinguists and computational linguists.

The discussion of whether anaphora is a textual or semantic relationship has been carried out ever since anaphora became a topic of research in linguistics. Arguments in the favour and against a purely lexical interpretation have been brought and the overall outcome was that the interpretation of anaphora depends on the combination of linguistic, semantic and pragmatic factors. J.C.Milner [Milner1982] promoted the view that ”anaphors are directly interpretable on the basis of the linguistic context alone, without information related to the designated segments”.

The following example was used by Mehler and Dupoux [Mehler and Dupoux1987] (and subsequently cited by different other authors) to prove the limits of a strictly linguistic analysis and the way the discourse situation can interfere:

(3.1) Apres avoir considèré son dossier, le directeur limogea l’ouvrier, parce qu’*il* était un communiste convaincu.
(After considering his dossier, the director fired the worker because *he* was a convinced communist.)

The pronoun *il* can refer to either *the director* or *the worker*, depending on whether the action is set in a communist country or in the USA, therefore situational context plays a major part in the interpetation of this piece of text.

While semantic and world-knowledge information seems to be a strong clue that can override any other kind of textual interference in the interpretation of anaphora, the importance of semantics as a stand-alone factor has been challenged repeatedly. Consider for example the following sentence:

(3.2) If *a bomb* falls next to you, don't lose *your head*. Put *it* in a bucket and cover it with sand.

Although this classical example has been usually used for arguing the necessity of extra-linguistic knowledge, Kleiber uses it for proving that this kind of knowledge is not enough either. He bases his explanation on the fact that the utterance has got a humorous effect, which can only be explained by the presence of strong strictly linguistic factors that indicate *the head* and not *the bomb* as antecedent, therefore the first interpretation, although incorrect, is valid. Moreover, these linguistic factors are activated before semantic inferences are performed, otherwise the false interpretation would never be mentally built.

However, while linguistics was concerned with the semantics and pragmatics of anaphora, and psycholinguistics concentrated on the mental phenomena and inferences involved in the interpretation of anaphors, computational linguistics is concerned not only with identifying the factors that are generally used in interpreting anaphora, but with extracting and modelling them into a machine-usable format. Also, computational linguistics needs to differentiate between which of these factors can be used as constraints and which can be considered preferences. Therefore, although it is generally agreed that for a perfect interpretation of the text one needs deep semantic and world knowledge, from a purely pragmatic computational linguistics point of view this information is not usable. For the time being this kind of knowledge is impossible to model, which is due to prohibitive expenses involved in building a world knowledge ontology and to the extremely high level of processing necessary for exploiting such ontologies.

In the following, the most important categories of factors that can and have been used in solving pronominal anaphora will be presented.

3.2.2 Morpho-syntactic features

3.2.2.1 Morphological agreement

As pronouns are lexical entities that substitute for a noun phrase, they have to display the same morphological features as the noun phrase they substitute for. However, from a computational point of view this can cause different problems, as this substitution noun phrase - pronoun is not always apparent at textual level. Indeed, pronouns can refer to a conceptual entity that has never appeared in the text, but whose existence can be inferred (as it has been explained in Chapter 1, indirect anaphora). Therefore, numerous cases of gender and number disagreement can appear between a pronominal anaphor and its apparent textual antecedent. An investigation of approximately 2000 plural pronouns extracted from written texts in BNC showed that almost a quarter of plural pronouns do not agree in number to their antecedents (for a more detailed discussion, see section 7.4).

3.2.2.2 Syntax and the role of implicit causality

Syntactic preference

Most of the systems performing anaphora resolution include a preference for a noun phrase in subject position to be antecedent (starting with the first anaphora resolver, Winograd's SHRDLU [Winograd1972]). This preference has been expressed in Hobbs' top-down, left-to-right search [Hobbs1978] and in the ordering of the forward-looking centers in Centering Theory (see section 3.2.4 below). However, it is not always clear if the real preference is for the subject or for the noun phrase appearing first in the sentence, as these phenomena very often overlap. Moreover, it is not clear if penalising a noun phrase in indirect object position is due to its syntactic

function or to its morphological function (a prepositional phrase)[1]. These differences have been the subject of psycholinguistic studies, as it will be shown in section 3.2.5.

Syntactic exclusion

The Theory of Government and Binding (developed by Chomsky in a number of publications throughout the eighties, such as [Chomsky1980],[Chomsky1981], [Chomsky1982]) offers a number of constraints that enable the restriction of the antecedents a pronoun can refer to. The first constraint refers to reflexives and reciprocals, which, lacking independent reference, must be bound to an antecedent:

Principle A: A reflexive must be bound in the minimal domain containing it, its governor and an accessible subject.

The second principle regulates the use of anaphoric pronouns:

Principle B: A pronoun must be free in its governing category.

The third principle refers to the use of nominals:

Principle C: Full expressions must not be bound.

These rules govern the use of reflexive and personal pronouns in certain constructions. They explain why constructions like the following are incorrect or do not convey the intended meaning:

(3.3) *Herself* looked in the mirror. (the reflexive is not bound, so Principle A is violated)

(3.4) Julia likes *her*. (*her* cannot be interpreted as referring to *Julia* because of the free-binding stipulated by Principle B)

(3.5) John likes John. (the repetition of the noun phrase *John* violates Principle

[1]Most rule-based anaphora resolvers consider prepositional phrases less likely to be referred to by pronouns. They also assume that the subject of a clause is more likely to be referred to by a pronoun than a direct or indirect object. Most frequently, prepositional NPs are either indirect objects or complements, thus less likely to be antecedents than non-prepositional subject NPs.

C)

Syntactic parallelism

Syntactic parallelism refers to the phenomenon that occurs when the pronoun and its antecedent have the same syntactic function in different sentences.

Empirical studies showed that sentences displaying this feature read better than sentences containing a pronoun that does not refer to a noun phrase with the same grammatical role. This preference can be explained in the terms of Centering Theory (see next section), but it can be easily overridden by semantic constraints, as proven by the following example, which will be discussed later in this chapter:

(3.6) John took the cake from the table and ate it.

The role of implicit causality

Implicit causality can affect the interpretation of pronouns under a narrower range of circumstances. It can come into play when one event, state or process is given as the reason or cause for another and it is based on the property of some verbs to reflect causality on either their subject or their object.

For exemplification, let's consider the following two sets of sentences:

(3.7) (a) John sold his car to Bill because he had taken up cycling.

(b) John sold his car to Bill because he needed a means of transport.

(3.8) (a) John blamed Bill because he was careless.

(b) John blamed Bill because he needed a scapegoat.

Psycholinguistic studies ([Garnham1992, Gernsbacher1990]) showed that sentences like (3.7a) read better then sentences like (3.7b), and sentences like (3.8b)

read better then sentences like (3.8a). The subordinate clause contains a pronoun that can be interpreted to either of the NPs in the main clause. The subordinate clause provides an explicit cause for the event in the main clause. In addition, the main clause itself (in examples 3.7a and 3.8b) suggests that the reason or cause for the action has something to do with John rather then with Bill. [Grober1991] refers to verbs such as *sell* as NP1 verbs, because they impute causality to the first (or subject) NP in a simple active clause containing them, and to verbs such as *blame* as NP2 verbs, because they seem to impute causality to the second (or object) NP.

If it is important to take into account implicit causality when generating natural language, in order to produce coherent text, it does not seem to play the same role when trying to analyse natural language. The reason behind this lies in the quality of the texts processed. A shallow analysis of the corpus of technical manuals used in this project (see Chapter 4) shows that such documents, although coherent, were not produced with the thought of being easily read in mind, therefore the role of implicit causality is greatly diminished. The coherence of the discourse is mainly realised by semantic constraints.

3.2.3 The role of semantics

It seems that in many cases syntactic and topicality filters are not enough for removing ambiguities. Let's consider as an exemplification the same classical sentence [Carbonell and Brown1988]:

(3.9) John took the cake from the table and ate it.

Simple morphological and syntactical filters cannot fully disambiguate the pronoun *it*, since there are two noun phrases that can equally be antecedents: *the cake* and *the table*. However, the sentence poses no problem of interpretation to human readers, due to the knowledge that *eat* is a verb that requires *food* as object, *cake* is a type of food

and *table* is not a type of food. Wilks [Wilks1973] introduced the idea of preference semantics, whereby dispreferred readings can be allowed when the preferred readings are not present.

As powerful as semantic restrictions seem to be, there are two types of problems associated with their use. First, as Carbonell and Brown pointed out, semantics alone is not enough for eliminating ambiguities. As an alternative to example (4), they produce the following sentence:

(3.10) John took the cake from the table and washed it.

In this case, the semantic restrictions associated with the verb *to wash* are not powerful enough to disambiguate the sentence, if instead of *cake* we have a vegetable or a fruit that is likely to require washing.

The second problem refers to the computational implementation of semantic restrictions. Semantics is inextricably linked to knowledge, therefore using semantic information poses two major problems: encoding real-world knowledge and defining a semantic representation. Neither of these problems is trivial: building world-knowledge ontologies is very time consuming and expensive, requires continuous updating and they can be computationally expensive to consult; for the time being, a consensus has not been achieved on an unique type of semantic representation.

Modelling semantic preferences using collocations

As mentioned before, exploiting selectional constraints is a difficult task, requiring deep understanding of the text and large databases of semantic dependencies. As a cheaper alternative that produces reasonable results, Dagan&Itai [Dagan and Itai1990] introduced the idea of using collocation patterns automatically extracted from large corpora. The basic idea behind their method is that a pronoun can be substituted with its antecedent, therefore the antecedent must satisfy the same selectional constraints.

The model substitutes a pronoun with each of its possible antecedents and computes the frequency of the co-occurrence patterns in a corpus. In this way, a candidate can be rejected if the frequency of its associated co-occurrence patterns is not significant. If more than one candidate satisfies the selectional restrictions and it has a significant number of occurrences in the corpus, other methods have to be used in order to decide on the correct antecedent.

3.2.4 Discourse constraints

While the factors described above relate to characteristics extracted either from the surface form of an anaphor or from its local context, it has to be taken into account the fact that anaphora is a discourse phenomenon, hence it cannot be separated from the analysis of the discourse structure.

Discourse theories introduced the notion of *domain of accessibility* of an anaphor, i.e. the discourse space where a pronoun can find its antecedent. [Sidner1983] and [Grosz et al.1995] refer to *focus* as the portion of discourse that is active at a certain moment; according to them, a pronoun can only refer to those entities that are in focus at the moment when the pronoun is processed. In order to identify the active segment of a discourse, Sidner proposes the *stack model*, that evolves according to the changes in the attentional state of the reader.

Centering theory (CT) [Grosz et al.1995] is a theory of local coherence that models the way the attention shifts within a discourse segment. According to this theory, each discourse unit (which can be a clause, a sentence or another type of discourse segment) has a list of forward-looking centers, partially ordered. This ordering is usually given by the syntactic function of the entities (with different degrees of granularity), but other rules have been employed as well, to model the characteristics of different languages (e.g, textual position, semantic role). Forward looking centers

are likely to be referred to by pronouns in subsequent utterances. In addition, each unit has an unique backward-looking center (Cb), which is the center of attention, the entity the discourse is about. The relationship between the backward-looking center of a unit and the backward-looking and preferred centers of the previous unit determines the fluency of the discourse. If the Cb of the current unit is the same as the Cp of the current unit and as the Cb of the previous unit, it is said that the transition between the two units is a *Continuation* - this situation indicates that the speaker has been discussing about a certain entity and intends to discuss about it in the next unit as well. Whenever the Cb of the current unit is the same as the Cb of the previous unit, but it is not the preferred center in the current unit, the transition is a *Retain*, and it indicates that the speaker has been talking about a certain entity over the past two units, but the topic is likely to change. If the Cb of the current unit is not the same as the Cb of the previous unit, then the transition is a *Shift*, indicating that the speaker has started talking about a new entity[2]. CT stipulates that a *Continuation* transition is preferred over a *Retain*, which is preferred over a *Shift* - the preference being expressed in how easily read the discourse is when employing any of these transitions. The pronominalisation conjecture states that if any of the forward-looking centers of the previous utterance is pronominalised in the current utterance, then the backward-looking center of the current utterance has to be pronominalised as well. CT does not reject a discourse that violates these constraints, it mainly states that if any of the rules is violated, the discourse will be less fluent.

As Centering Theory has been designed as a model of local coherence, applying linearly within the limits of a discourse segment, extensions have been developed to

[2]The *Shift* relation has been further divided into *Smooth-shift* - the Cb of the current unit is the preferred center of the unit and *Abrupt-shift* - if the Cb and the Cp of the current unit differ [Brennan et al.1987] (see 3.4.1). This distinction indicates that the speaker has changed the topic and intends to continue speaking about the new entity, or they have done so momentarily

account for the global discourse coherence.

Veins theory [Cristea et al.1998] describes a model of accessibility based on the rhetorical structure of discourse. The central notion is the *vein*, which intuitively corresponds to an argumentation thread of the discourse. They show [Cristea et al.1999] that the same CT rules apply along the veins, resulting in a substantial reduction of the accessibility space of pronouns.

Despite the large number of methods proposed for the identification of the space of referential accessibility, recent practical experiments [Tetreault2002] shows that improvements in such algorithms do not bring major improvements to the pronoun resolution systems. Tetreault's conclusion is that this is due to the fact that these methods can only bring an improvement on the resolution of long-distance intra-sentential anaphors. However, in natural language this is quite a rare phenomenon, most of the anaphors being inter-sentential.

3.2.5 When and where to apply? - the relative importance of factors

Having identified the factors that help the interpretation of anaphora, the problem that remains is how these factors combine, at which point one is preferred over an other and which is their relative importance. This issue is particularly important for the computational treatment of anaphoric expressions as will be shown later in this chapter.

Psycholinguistic experiments played an important role in solving these problems. Analysing the changes in the reaction time of human subjects in the presence or absence of a certain indicator gives important clues to whether that indicator is particularly important to the interpretation of the text or not. It is possible to study the importance of different factors individually or in the context where they are combined

with other factors.

Several psycholinguistic studies investigated the role of gender cues in the identification of antecedents. [Ehrlish and Rayner1983] showed that gender agreement can eliminate a preference for an early as opposed to a late antecedent. [Garnham1992] and subsequently other works showed that gender agreement does not usually speed up the reading of a text under normal circumstances; however, readers do react to flagrant gender mismatches. [Carreiras1997] analysed the role of gender cues for languages with arbitrary grammatical gender (Italian and French). Their investigation targeted nouns that have a fixed syntactic gender but can be referred to by both masculine and feminine pronouns (the so-called *epicenes*). They showed that pronouns, proper names and clitics were interpreted quicker if they matched the epicene in gender.

The role of number agreement in the interpretation of pronouns has not benefitted of such an extensive analysis. [Garrod and Sanford1982] investigated cases where pronouns followed a conjoined NP in subject position in the previous sentence. They showed that singular and plural pronouns were interpreted equally easily when in subject position, but plural pronouns were interpreted more easily in object position. In a similar direction of research, Sanford and Lockhart [Sanford and Lockhart1990] discovered a small preference for plural subject pronouns over singular subject pronouns following a sentence containing a conjoint NP. Clifton and Ferreira [Clifton and F.Ferreira1987] showed that a sentence containing a plural pronoun was read as quickly when referring to a conjoint NP as when referring to a split antecedent. Their interpretation of the phenomena is that plural pronouns draw their interpretation from a discourse representation, not a surface one. This interpretation is backed up by another experiment ([Carreiras1997]) that showed that plural pronouns were more easily interpreted when their antecedents were spatially close, rather then when their

antecedents were split.

In [Matthews and Chodorow1988] the authors describe an experiment involving deep embedded antecedents and concluded that readers search for the antecedents in a sentence in a left-to-right, top-down, breadth-first order, which is consistent with Hobbs's ([Hobbs1978]) interpretation.

Gernsbacher et al [Gernsbacher1990] performed large scale experiments to investigate the preference for first mentioned NPs over NPs found later in the sentence. They showed that the advantage of first mention is found regardless of the syntactic and semantic role played by a referring expression. This effect is not associated with NPs appearing in subject position, which are first mentions in active sentences with no frontal constituents. These findings contradict previous claims that pronouns in one clause tend to refer to subject NPs in the previous clause.

3.3 Pre-processing tools

This section moves forward from theoretical knowledge sources to practical problems related to their processing - basically underlining the difference between what it is *desirable* and what is *achievable*.

Acquiring information from the knowledge sources described above is not necessarily part of an anaphora resolution algorithm, but plays a central role in automatic anaphora resolution. The availability of high quality pre-processing tools is an important factor in anaphora resolution, given that errors in this stage are carried further and deteriorate the overall success rate of the system. The tasks to be performed in the pre-processing stage are, at least:

1. *Morpho-syntactic analysis of the input text*

This step consists in identifying the parts of speech and, possibly, the grammatical function of the words. A large selection of part-of-speech taggers and shallow parsers are available, with performances close to 99% for part-of-speech taggers and 90% for shallow parsers. Full syntactic parsing does not reach this accuracy, the best tools of this kind achieving around 80% success rate.

2. *Noun phrase identification*

In this step, the word constituents are combined into noun phrases. Correctly identifying the noun phrases is an essential issue in anaphora resolution, as the noun phrases are treated as possible antecedents; failing to identify a noun phrase can lead to failures in resolving a pronoun to that noun phrase, while introducing incorrect NPs expands the searching space for a pronoun, and can lead to incorrect resolution.

3. *Named entity recognition*

This deals with the classification of noun phrases into one of the categories in a predefined set. The most frequent categories are *Person, Date, Time, Number, Organisation, Location*, but these can be further divided into subcategories according to the requirements of certain applications or types of texts. Systems in the MUC competition reported high results in the upper 90s, similar results being achieved by other named entity recognizers, like the one included in GATE[3] [Maynard et al.2001] (up to perfect classification with 100% success rate).

4. *Identifying instances of non-anaphoric pronouns*

Both in English and in French, third person pronouns (il, they, il, le, ils, on) can appear

[3]GATE (General Architecture for Text Engineering) is an open-domain software architecture for the processing of natural language, developed at the University of Sheffield

in non-nominal roles. The task consists in identifying these non-nominal occurrences and removing them from the anaphora resolution process. This can be done using rule-based [Kennedy and Boguraev1996, Lappin and Leass1994, Paice and Husk1987] or machine learning methods [Evans2001]. The accuracy of the classification reported so far for English pleonastic pronouns is around 78%-79% ([Evans2001]).

5. *Term identification*

Identifying terms in a text (words or phrases that are more relevant for a specific domain) means identifying those entities that are more likely to be in the focus, so are more likely to be referred by pronouns. As terms are specific for a domain, if an anaphora resolver is domain-independent but uses terms, the first task is to identify the type of document submitted to processing. Methods for identifying terms in discourse are generally based on statistical approaches (word frequency, TF.IDF).

6. *Extracting semantic information*

Semantic information can be used in a number of ways, including taking advantage of synonymy and hyperonimy, selectional restrictions. It is definitely a valuable resource for anaphora resolution, and not only, because it goes beyond the surface form of the text, providing information about the meaning of the discourse. A resource that can be used for extracting semantic information is WordNet (and its European version, EuroWordNet), a handcrafted collection of words organised in sets of synonyms (synsets), which are hierarchically linked according to their hyperonymic relation. Different senses of a word appear in the order of their frequency.

3.4 Anaphora resolution so far

Methods for anaphora resolution have been implemented both as stand-alone applications and embedded in larger applications of text or speech processing. It is therefore difficult to assess the relative success of such methods, since, being designed for different tasks, they are expected to perform better in the environment they were created for. Moreover, it is difficult to separate methods that tackle pronominal anaphora only from those that deal with pronominal anaphora in the context of coreference, as these domains overlap.

In the following, some of the most successful anaphora resolution methods will be presented; they will be categorised according to the types of knowledge sources they require and the processing mode employed.

3.4.1 Syntax-based methods

Hobbs' naive approach [Hobbs1978] was one of the first algorithms for anaphora resolution and is still one of the best. It employs a left-to-right breadth-first search of the syntactic parsing tree. Implicitly, this searching method assumes a certain salience of the candidates, giving preference to noun phrases in subject position, followed by direct objects, indirect objects, adverbials and adjuncts. The algorithm is augmented with various types of knowledge (morphological agreement, selectional restrictions). Despite its high success rate, this method poses implementation problems, given that deep syntactic parsing is necessary.

Another syntax-based method is that proposed by Lappin and Leass [Lappin and Leass1994]. It employs a variety of intra-sentential syntactic factors and the general idea of the method is to construct coreference equivalence classes that have an associated value based on a set of ten factors. An attempt is then made to resolve

every pronoun to one of the previously introduced discourse referents by taking into account the salience value of the class to which each possible antecedent belongs.

BFP [Brennan et al.1987] is an algorithm for pronoun resolution based on Centering. The general idea is to identify the set of linkages (pronoun, antecedent) that gives the best coherence of a text. The coherence is computed by associating a score to each transition holding between two adjacent units and summing the scores for the text[4]. This kind of approach however can only solve those pronouns that are in Cb position and furthermore it only functions within a discourse segment[5].

3.4.2 Knowledge-poor methods

Traditionally, methods for anaphora resolution are based on factors that contribute to solving pronouns. Sets of possible antecedents are associated to each anaphor and rules are applied on these sets. The rules can be either strong (eliminatory), meaning that if a possible candidate does not satisfy it, it cannot be antecedent for the pronoun under analysis, or weak (preferential) which mainly state that a candidate satisfying the rule is likely to be antecedent for the pronoun. The difference in approaches lies in the rules employed and in the manner they are applied.

Knowledge-poor methods take advantage of those factors involved in anaphora resolution that can be computed reliably. These exclude, for example, the implementation of selectional restrictions or of full syntactic parsers, which are either too expensive to employ or not reliable enough at this time. Therefore, although it is not expected for the knowledge-poor methods to perform as well as more

[4] *Continuation* is considered the ”smoothest” transition and receives a score of 4, followed by *Retain* (3), *Smooth shift*(2) and *Abrupt shift*(1)

[5] A discourse segment is a portion of discourse that has an identifiable communicative goal. A shift in the communicative goal indicates a discourse segment boundary. For practical reasons, a discourse segment can be identified as either a paragraph or a multi-paragraph text span. Different variants of CT employ various definitions of *unit* and *discourse segment*.

complex methods that make use of deeper semantic and syntactic knowledge, they are a computationally cheap alternative, being more appropriate for use in larger text understanding applications.

One of the earliest anaphora resolver is Winograd's SHRDLU [Winograd1972]. It is worth mentioning it here because it was the first such system that defined rules for pronoun resolution. The search for possible antecedents was done in the whole portion of discourse preceding the pronoun and the most salient antecedent was determined on the basis of syntactic and focus preferences.

CogNIAC [Baldwin1997] is a knowledge-poor approach to anaphora resolution based on a set of high confidence rules which are successively applied over the pronoun under processing. The rules are ordered according with their importance and relevance to anaphora resolution. The processing of a pronoun stops when one rule was satisfied. The original version of the algorithm was non-robust, a pronoun being resolved only if one of the rules was applied. The author also describes a robust extension of the algorithm, which employs two more weak rules that have to be applied if all the others failed.

Kennedy and Boguraev [Kennedy and Boguraev1996] describe an algorithm for anaphora resolution based on Lappin&Leass' approach but without employing deep syntactic parsing. The syntactic rules are modelled using heuristics. This work proves that even without employing expensive full parsing, good results can be obtained.

Mitkov's approach [Mitkov1998] is a robust anaphora resolution method for technical texts and is based on a set of boosting and impeding indicators that are applied on each antecedent of a pronoun. A score is calculated based on these indicators and the discourse referent with the highest aggregate value is selected as antecedent. Preliminary evaluation (performed using manual post-editing of the shallow parser output) shows a precision of up to 90%, while subsequent results

([Mitkov et al.2002]) obtained using fully automatic running mode and evaluation are much lower, in the region of 70%.

All these methods deal with personal pronouns referring to nominals only. Moreover, they use generalised rules that apply to any kind of personal pronouns, ignoring the fact that different pronouns may benefit from different approaches to solving them[6].

3.4.3 Machine learning methods

Although various ambiguity tasks in NLP were solved using machine learning, less work was done in applying it to anaphora resolution, leaving space for improvement.

3.4.3.1 Supervised methods

Aone and Benett developed a method for identifying anaphoric relations for a specific domain (business joint ventures). Their system learns decision trees using a set of 66 features, many of these being domain-specific. The machine learning system they developed is not oriented towards solving pronominal anaphora only, but bridging anaphora as well, so the outcome of their system is expressed in both pairs (anaphor, antecedent) and the type of anaphoric relation established between the two.

In [Connolly et al.1997], several machine learning methods are evaluated. Their approach is different compared to the others, in that that they do not learn if a certain candidate is the correct anaphor for a pronoun, but if that candidate is more probable to be antecedent than another noun phrase.

[6] Kennedy&Boguraev's method employs specific syntactic constraints for solving reflexives and has a slightly different way of treating possessives.

3.4.3.2 Unsupervised methods

Cardie describes an unsupervised learning method for coreference resolution based on clustering. As for all unsupervised methods, her approach has the advantage of requiring but a small set of annotated data. However, the results obtained do not reach the level of supervised methods, nor of the classical ones.

Genetic algorithms (see section 5.2.1.6) have not been extensively used for natural language processing tasks, and there have been only two attempts to use them in solving pronominal anaphora - one of the reasons being the difficulty in defining a fitness function. Byron [Byron and Allen1999] presents a genetic algorithm that calculates the best weight assignment for a set of ”voters”. A voter is a module that indicates a preference for a certain noun phrase to be chosen as antecedent for a pronoun. The genetic algorithm identifies the weights by running a pronoun resolution on a training corpus (extract from Penn Treebank and annotated with coreferential links) and identifying the best possible combination of weights for all voters. Although the results reported are good (71% overall success rate), showing a slight improvement of Hobbs' naive algorithm on the same test set, it has to be noticed that only pronouns with simple nominal antecedents were included in the evaluation; therefore, pleonastic pronouns, those referring to verb phrases or clauses, plural pronouns with split antecedents were not solved. Orăsan and Evans [Orăsan et al.2000] used genetic algorithms for finding the best combination of weights for the indicators used within the implementation of Mitkov's knowledge-poor approach [Mitkov1998].

3.4.3.3 Statistical methods

Ge et al. [Ge et al.1998] describe an algorithm for resolving pronominal anaphora based on a statistical framework. They build a probabilistic model by combining

different factors like syntax, morphological features of noun phrases (gender, number, animacy), relative position in text of the antecedent compared to the pronoun , selectional restrictions and mention frequency; the way these factors are combined depends on several independence assumptions. Statistics are collected from a training set (consisting of an annotated excerpt from the Wall Street Journal corpus) in order to find the values of the probabilities involved. The evaluation performed on a testing set from the Wall Street Journal Corpus showed an accuracy of classification of about 85%.

3.4.4 Automatic resolution of other anaphoric types

The previous section presented some of the best known pronoun resolution systems. Although, as mentioned before, pronoun resolution is the most extensively studied area of anaphora resolution, some other types of reference have also been attempted to be solved automatically. This section will briefly discuss some approaches aimed at resolving general coreference, definite descriptions, and indirect anaphora.

3.4.4.1 Coreference resolution in the MUC competitions

The interest in coreference resolution materialised in a special task introduced in the MUC competition starting with their 6th edition. The systems participating in the competition were requested to identify coreference links in texts from a restricted domain (business joint-ventures) and to express their results using the evaluation measures designed for the task (precision and recall).

3.4.4.2 Resolving definite descriptions

Although there have been theoretical works into the role, classification and semantics of definite descriptions ([Heim1982, Neale1990, Prince1982] among the most

important), there are few approaches to solving definite descriptions. Most such attempts do not treat definite descriptions as an individual phenomenon. Some early works by Sidner [Sidner1979, Sidner1983] and Carter [Carter1987] treated definite descriptions in the larger context of resolving anaphoric relations. However, Sidner's focus-tracking algorithm is not viable in real applications, as it requires semantic and world knowledge; moreover, there it does not benefit from evaluation. Carter extended Sidner's approach in a more viable framework and also attempted to evaluate his system on a small amount of manufactured texts.

Among the few approaches targeting definite descriptions, there have to be mentioned the works of Vieira and Poesio. [Vieira and Poesio1999] uses inductive decision trees for the classification of definite description, based on a set of syntactic features computed over a training corpus. In [Vieira1998, Vieira and Poesio2000], a system for the interpretation of definite descriptions is presented that uses syntax-based heuristics for the identification of the antecedents and of the discourse-new elements.

3.4.5 Anaphora resolution for French

All the methods described before were designed and evaluated for English. As mentioned in the Introduction, much more research was put into developing natural language processing tools for English than for any other language - and anaphora resolution is not an exception. However, there have been a number of original anaphora resolution algorithms designed for French, some of them being implemented in automatic systems.

3.4.5.1 Anaphora resolution within the CERVICAL project

The CERVICAL project[7] was developed as part of research on human-computer dialogue treatment.

The method is based on the Theory of Mental Representations [Reboul1997], a global theory of reference which introduces the notion of mental representation (RM - *representation mental*) associated with a referring expression. An RM encodes different types of information: lexical (textual realisation of the RM), logical (relationship with other RMs), visual or encyclopedic. For practical reasons, this quite theoretical definition of a mental representation was simplified: mental representation denotes an entity that appears along a discourse, therefore it can be interpreted as being the ensemble of discourse entities in a coreferential chain. When a new discourse entity is treated, it can be either associated with a mental representation, therefore contributing to its description, or it can introduce a new mental representation. The decision of whether a discourse entity is attached to an already existent RM and the choice of this RM depend on a set of syntactic, morphological and semantic constraints. Reports on the evaluation of the system [Popescu-Belis and Robba1997] showed a promising 62% precision in the resolution of pronominal anaphora; however, the system is not fully automatic, as it receives the referential expressions as input.

3.4.5.2 Trouilleux's pronoun resolution system

In his recent PhD thesis [Trouilleux2001], Trouilleux presents a rule based pronoun resolution system for French. His method relies on identifying a set of rules from corpus observation. The system does present an important drawback, mainly that the system is evaluated on the same corpus used for the selection of the rules. This can lead

[7] The acronym stands for *Communication et référence: vers une informatique collaborant avec la linguistique.*

to biased implementations, and inaccurate evaluation results. However, the reported results are very good, in the mid 80% precision and about the same recall. Given that the system is fully automatic, these results outperform those obtained in English pronoun resolution. For the time being, Trouilleux's system is the best evaluated and documented work in automatic anaphora resolution for French.

3.4.6 Multilingual anaphora resolution methods

The methods previously described followed the traditional trend in NLP of tackling only texts in a specific language - either English or French. More recently, there have been attempts of adapting for other languages methods formerly used for English.

Mitkov's method, for example, was tested for Polish [Mitkov and Stys1997], Arabic [Mitkov et al.1998], Bulgarian [Tanev and Mitkov2000] and French [Mitkov and Barbu2000] and has proven to perform comparatively well. Its multilingual character is due to the language-independent factors and to the lack of certain types of information that are language-specific (such as syntactic features or topicality).

Azzam et. al [Azzam et al.1998] experimented on coreference resolution for English and French using the coreference module integrated in M-LaSIE[8]. Their evaluation, performed on parallel texts containing a very small number of pronouns proved shows lower precision for resolving French pronouns (78%) compared to English (84%), but the results were improved when information about the semantic types of verb roles was added (precision raised to 94%).

These methods are mainly based on adapting methods initially designed for English to other languages. As opposed to this approach, some systems make use of bilingual

[8]Multilingual information extraction system developed at the University of Sheffield as an extension to LaSIE (Large Scale Information Extraction)

parallel corpora in order to mutually enhance the performance of anaphora resolvers in both languages. They take advantage of language-specific features that rend the anaphora resolution process easier in one of the languages, and transfer the information thus acquired to the other language.

[Harabagiu and Maiorano2000] presents an extension of their COCKTAIL coreference resolution system that learns coreference rules from bilingual corpora, therefore being able not only to work in a multilingual environment but to improve its accuracy in each of the languages involved. The system was tested for English and Romanian and showed an improvement of about 4% in precision when the coreference resolver was run on parallel texts over the results obtained when run for each language individually. The main drawback of the latter class of anaphora resolvers is the necessity of running them in a multilingual environment. They require parallel corpora, aligned at word level, a result which cannot be as yet reliably achieved.

Nevertheless, all these works prove that it is possible to design multilingual anaphora resolver cores that are easily adaptable to different languages with small loss of accuracy.

3.5 Conclusions

This chapter has presented different approaches to the problem of anaphora resolution. To summarise the state of the art in automatic anaphora resolution, table 3.1 displays the most successful such systems, along with the type of anaphoric expressions they tackle and their performance.

Concerning multilingual anaphora resolution, the main conclusion is that work in this area is still desirable. The few works carried out in this direction showed promising results but also showed there is still scope for improvement.

Method	Lang	Pronouns solved	Evaluation measure	Accuracy	Evaluation	Genre	Size	Year
Hobbs	En	personal	precision	88.3%	manual	technical	50	1978
Lappin & Leass	En	personal	precision	86%	automatic, corrected input	technical	360	1994
Kennedy & Boguraev	En	personal, poss, refl	precision	78%	automatic	various	306	1998
Mitkov	En	personal	precision, recall	89.7%	automatic, corrected input	technical	104	1998
CogNIAC	En	he/she	precision/ recall	77.9%/ 35%	automatic	news	300	1995
SPAR	En	personal	precision	93%	automatic	various, purpose written	242	1995
Ge	En	personal	precison	84.2%	automatic	WSJ	140	1995
Trouilleux	Fr	personal	precision	80%	automatic	economics	417	1995

Table 3.1: Accuracy of different systems for anaphora resolution

Chapter 4

Corpus annotation and analysis

4.1 Introduction

Corpora are a critical resource with wide applications in most natural language processing applications. However useful raw data may be for certain linguistic observation tasks, corpora enriched with annotation are far more important. Nevertheless, the drawback is that building such resources is expensive, time consuming and in some cases, very difficult.

The first part of this chapter will discuss some general issues related to the development of corpora annotated for anaphoric links and the use of corpora in automatic anaphora resolution.

The second part is dedicated to the description of the corpus especially developed for this chapter, and its use in training and evaluating the system described in Chapter 5. The choice of a particular data set will be explained and some analysis of the data will be presented.

4.2 Anaphoric expressions across different text genres

In designing an NLP system, careful consideration has to be given to the specification of the system's scope and target. Practical issues prohibit us from building good performance NLP applications for unrestricted text, therefore the choice of target texts influences the design of the system and sets an upper limit to the expected performance. The same considerations hold true for anaphora resolution applications.

Text genres differ with respect to the register of referential expressions they employ. In a study involving four text genres (news reportage, academic prose, conversation and speeches), Biber et al [Biber et al.1998] notice clear differences in the number, distribution and types of referential devices. The investigation followed four parameters: status of information (given vs new), the type of reference for given information, the form of the expression and the distance between the anaphor and the antecedent for anaphoric reference).

The study found that news reports contain the highest number of referring expressions, while the academic prose contains the lowest; conversations and public speeches contain a relatively large number of referring expressions.

With regard to the distribution of pronouns, conversations contain the largest number of referring pronouns, but most of them are exophoric, mainly used for referring to the speaker or to the hearer. An interesting finding is that written registers contained almost no exophoric references (apart from some occasional references to the author), therefore most referential pronouns were anaphoric.

Another interesting difference among these text genres appears in the distance between the anaphor and its antecedent. In the spoken data, the antecedent was always much closer to the anaphor (4.5 intervening referring expressions on average) than in the written data (11 intervening referring expressions on average).

These findings help reaching the conclusion that differences in anaphoric devices across registers are important enough to make automatic anaphora resolution a domain-specific task.

4.3 Corpus annotation for anaphora resolution

4.3.1 General issues

Annotated corpora play an important role in at least two areas of automatic anaphora resolution: training of statistical and machine learning methods and evaluation of systems. It is also important to have access to annotated corpora for observing patterns and deducing rules for rule based methods, but this section will only concentrate on the automatic exploitation of corpora.

Manually annotated corpora are essential for both comparative and qualitative evaluation, as it will be discussed in more detail in Chapter 6. As for the use of corpora in training, it will be shown later in Chapter 5 that statistical models for NLP are based on learning patterns of language from large collection of texts. The particular features that will be learned from the corpus are at the discretion of the application, however, in order to perform statistical learning with reliable results, it is necessary for the corpus to contain a significant number of occurrences of each of those features. Therefore, each feature that needs to be learned adds a new dimension to the complexity (and implicitly size) of the corpus. This means that there is a strong correlation between the number of features used for learning and the size of the training corpus, although this correlation has not been expressed mathematically.

Given the expected uses of corpora annotated for anaphoric links[1], their production has to be carefully dealt with, in order to produce high quality data. Manually

[1]The same considerations hold for all types of corpus annotation

annotating corpora for use in anaphora resolution applications is a notoriously difficult and labour intensive task, even when a single phenomenon is targeted. A number of studies [Hirschman1997, Kibble and van Deemter2000, Mitkov et al.2000] show that even when restricted to identity coreference, the decisions as to what and how to annotate are not simple.

[Kibble and van Deemter2000] show that in addition to the numerous ambiguous cases which challenge the annotators, the coreference annotation task suffers from terminological confusions. They criticise the MUC annotation task for mixing up coreference and anaphora relations under the general name of *coreference*, although many of the relations required to be annotated cannot fall under this category. [2]

4.3.2 Annotation schemes

One of the basic issues in corpus annotation is the development of annotation schemes that would encompass all the phenomena that are under observation in that corpus. Other important aspects in the development of annotation schemes are that they should allow easy processing of the corpus (to allow the replication of experiments on the corpus), should be extendible (to allow the marking up of phenomena not initially thought of). There is currently no widely accepted standard in annotating corpora for anaphoric links, a fact proved by the continuous development of new annotation schemes.

During the past 10 years, a number of original annotation schemes have been in use for annotating anaphora. The first one was the UCREL scheme initially developed by Geoffrey Leech (Lancaster University) and Ezra Black (IBM). It allows the marking of different varieties of anaphora including ellipsis, but also non-referential use pronouns.

[2] One point of disagreement for example is the inclusion in coreferential chains of entities which cannot corefer, as in the case of quantified NPs: *every man, most people*.

In addition, annotator uncertainty can be marked. Special symbols added to anaphors and antecedents can encode the direction of reference (i.e. anaphoric or cataphoric), the type of cohesive relationship involved, the antecedent of an anaphor, as well as various semantic features of anaphors and antecedents. Although very comprehensive, this annotation scheme has the drawback of being difficult to process automatically.

In recent years, annotation schemes for anaphoric links have followed the general trend in corpus annotation of employing widely accepted interchange formats like SGML and XML, thus moving a step forward towards standardisation. These formats allow much easier automatic processing, due to the large availability of processing techniques and software.

The MATE scheme for annotating coreference in dialogues [Davies et al.1998] draws on the MUC coreference scheme (see below), adding mechanisms for marking-up further types of information about anaphoric relations as done in the UCREL, DRAMA and Bruneseaux and Romarys schemes. In particular, this scheme allows for the mark up of anaphoric constructs typical in Romance languages such as clitics and of some typical dialogue phenomena. The scheme also provides for the mark up of ambiguities and misunderstandings in dialogue.

Other well known schemes include de Rocha's [de Rocha1997] scheme for annotating spoken Portuguese, Botley's [Botley1999] scheme for demonstrative pronouns, Bruneseaux and Romary's scheme [Bruneseaux and Romary1997], the DRAMA scheme [Passoneau and Litman1997] and the annotation scheme for marking up definite noun phrases proposed by [Poesio and Vieira1998, Vieira and Poesio1999][3].

There have not been mentioned so far the two annotation schemes that will be the basis for the annotation performed in this corpus. The MUC scheme [Hirschman1997]

[3]For a brief outline of the most important annotation schemes, see [Mitkov2002]

has enjoyed large popularity due to its employment in the coreference task of the MUC competitions, despite being the least ambitious scheme in terms of coverage. The scheme has been used by a number of researchers to annotate coreferential links [Gaizauskas and Humphreys2000a, Gaizauskas and Humphreys2000b, Mitkov et al.1999]. In this scheme, 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 the TYPE value 'IDENT' indicates the identity relationship between anaphor and antecedent. The MUC scheme only covers the identity (IDENT) relation for noun phrases and does not include other kinds of relations such as part-of or set membership. 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 mark-up of the tagged element is optional.

The XML-based scheme proposed in [Tutin et al.2000] was developed as part of a large ELRA[4]-funded project dealing with annotating French corpora with anaphoric links. It was designed to deal with coreference and other kinds of discourse relations: "description" (one-anaphora), "sentence" (for sentential antecedents, "member of" (part of), "indefinite". The annotation scheme can handle complex cases like multiple antecedent and discontinuous antecedents.

[4]European Language Resources Association

4.3.3 Annotation tools

Annotation tools are not a requisite for developing annotated corpora, however they do offer the human annotator trouble-free and efficient interaction with the text, hopefully minimising the time necessary to annotate and removing some types of human errors. A good annotation tool should 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 that offers the user an easy-to-navigate environment for manually marking pairs of anaphors-antecedents within the UCREL scheme [Fligelstone1992]. In particular, XANADU allows the user to move around a block of text, displaying circa 20 lines at a time. The user can use a mouse to mark any segment of text to which s/he wishes to add some labelling.

The **DTTool** (Discourse Tagging Tool), [Aone and Bennett1994], has been used for annotating anaphoric relations in Japanese, Spanish and English. It provides a graphical view with different colour coding for different types of anaphors (e.g. 3rd person pronoun, definite NP, proper name, zero pronoun, etc.) and antecedents which are displayed on the screen with arrows linking them. Several tag visualisation modes are also available.

The **Alembic Workbench**, [Day et al.1998], was developed at MITRE and offers, among other facilities, the option to mark-up coreference relations. In the coreference annotation mode the workbench features a window that produces a sorted list of all tagged elements to facilitate the finding of co-referring expressions. The semi-automatic mode extends to simple annotating tasks such as tagging named entities. The Alembic Workbench offers a choice of tag sets, including all those necessary for

the MUC scheme and provides a graphical interface which allows the modification of existing tags and the addition of new ones. Users of the system are also able to construct their own task-specific annotation schemes.

Referee is a discourse annotation and visualisation tool that 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, distance, pronominalisation, definiteness etc.) with each of them and assigning coreference. In segment mode the user can partition the text into arbitrarily nested and overlapping segments, whereas the dialogue mode enables them to code a dialogue by breaking it into turns.

4.4 Corpus annotation for ARMAL

4.4.1 Overview

The annotated corpus produced as a by-product of this research has been designed for use in two stages of the production of the anaphora resolver: training of the machine learning modules and evaluation. As the system targets at least two languages, English and French, data in both languages was needed. Some decisions needed to be taken before the corpus annotation process began. The first decision concerned the choice of data: the register and the text selection. The second decision was related to practical annotation issues, such as the annotation scheme to be employed and the annotation tool to be used.

4.4.2 Choice of data

The choice of technical manuals as target texts for anaphora resolution was based on both theoretical and pragmatic reasons. Section 4.2 showed that restricting the resolution of anaphora to a certain domain is fully justifiable by the differences in anaphoric devices across registers. This is consistent with the current trend in developing domain-specific NLP applications.

Technical manuals contain a relatively high number of pronouns, although their variety and use is quite limited. This could represent a problem, in that a system designed to tackle such texts will probably not be able to deal with unrestricted domain texts. However, it also means that certain anaphoric behaviours can be studied in more depth and solved with better accuracy than general anaphora in free texts. Practically, this translates into the ability of a machine learning module to learn better certain patterns (that occur more frequently in the training corpus), although it will not be able to learn the less frequent ones.

The second reason for choosing technical manuals was that several other anaphora resolvers have been tested on such texts, therefore my system can be more easily compared with other approaches.

Finally, numerous technical documents exist in more than one language, which makes them suitable for use in a multilingual anaphora resolver. Moreover, the Linux documents included in our corpus are public domain, which alleviates the potentially tedious task of acquiring reproduction and distribution permission from the copyright holders.

4.4.3 Peculiarities of technical manuals

As mentioned in the previous section, the corpus used for training and testing consisted of technical manuals, collected on line. The French documents and part of the English documents have been collected specifically for the task of training and evaluating the system presented in this thesis, while the other documents were part of the coreferentially annotated corpus developed at the University of Wolverhampton.

Technical manuals collected on line have a number of peculiarities that can make automatic processing more difficult. They present a highly structured architecture, being split in many small length sections and subsections. This makes necessary for some special decisions to be taken as to the searching space for antecedents and the identification of the elements in focus. Furthermore, most of the texts[5] are not carefully authored, therefore they contain a relatively large number of typos and syntactical mistakes. The most frequent type of syntactical mistakes noticed have been verb-subject disagreement and confusions between the possessive determiner *its* and the contraction *it's*, along with some malformed sentences. These mistakes have been corrected before submitting the texts to processing, as it will be shown further in this chapter.

Apart from these features, which are common to both English and French technical manuals, French technical writing also presents a peculiarity which makes them more difficult to process automatically: lexical borrowing. Technical terms borrowed from English are relatively frequent, even more so in computing literature. This phenomenon is due either to the lack of equivalent French terms (*pipe-line*, for example) or to the fact that certain English terms have gained such widespread use that their translation is unnecessary (such as *software* or *printer*). The first category of

[5]This refers mainly to the Linux documents in particular, which are contributed by individuals. Technical manuals and documentation produced by professionals are much more carefully authored.

terms are usually marked typographically, while the second category are not identified in any way. The problem that borrowed terms raise relate to the impossibility of clearly identifying their morphological type by the shallow parser.

4.4.4 Corpus

4.4.4.1 English texts

The texts selected for English are Linux technical manuals: Access How-To (ACC[6]), CDRom How-To (CDROM), Beowolf How-To (BEO), Ethernet How-To (ETH), an extract from an Internet Explorer user manual (WINHELP), a documentation for Panasonic TV (PAN) and Aiwa (AIW) products. The texts total about 55000 words, containing 653 pronouns, out of which 546 anaphoric.

4.4.4.2 French texts

The French texts selected for annotation are Latex Manuals (three files: TEX1, TEX2, TEX3) and Linux manuals (BEOFR, CDROMFR, ACCFR). Although all the technical manuals that the English corpus consists of were available in French as well, their quality was not always good enough to justify their inclusion in the corpus. The Linux manuals selected for annotation do not represent the exact translation of their English counterparts, therefore they cannot be considered parallel. The size of the resulting French corpus is about 36000 words, with 482 pronouns, out of which 367 anaphoric.

[6]The texts will be further identified by these acronyms.

4.4.4.3 Parallel corpus

For the purpose of cross-genre evaluation, a small subset of the BAF Canadian corpus[7] has been annotated, containing an extract from a scientific text[8] (CITI) and the first three chapters of a novel[9] (VERNE). The texts represent translations from French to English and are aligned at sentence level. Although for evaluation purposes it was not essential to use a parallel corpus, doing so enabled us to perform some experiments regarding the possibility of mutually enhancing the performance of the English and French anaphora resolvers (as described in Chapter 7). This corpus was not used in training, as it was not my intention to learn resolution rules from non-technical texts, but to assess how rules learnt on technical texts perform on other text genres.

Table 4.1 describes the corpus used for training and evaluation, in terms of number of files, number of words, number and types of pronouns.

4.4.5 Data analysis

4.4.5.1 Composition of the corpus

The analysis of the corpus reveals some interesting features of technical manuals with respect to the distribution of pronouns (results in table 4.2 [10]). The pronouns[11] represent about 1% of the total number of words, but there is a high number of non-anaphoric pronouns, amounting to almost 25% of the total number of pronouns. The second observation concerns the low incidence of masculine and feminine pronouns,

[7] BAF was developed at the RALI laboratory, University of Montreal

[8] Geoffroy, Catherine (1994). Les technologies de communication de l'information et les aînés. Rapport technique du CITI

[9] Verne, Jules. De la terre la lune.

[10] In the table, the columns display, in this order, the number of paragraphs, sentences, definite NPs, indefinite NPs, prepositional NPs, personal pronouns, possessive pronouns and reflexive pronouns.

[11] Only the pronouns of interest for the project are considered, i.e. personal, possessive and reflexive pronouns. Other types of pronouns are considered as normal words and are not included in the numbers.

	File	#words	#pronouns	#non-anaphoric	#anaphoric
English	ACC	9993	182	26	156
	CDR	10756	97	12	85
	WIN	2862	51	1	50
	ETH	20269	254	43	211
	AIW	6723	38	14	24
	PAN	4841	31	11	20
	Total	55444	653	107	546
French	TEX1	2684	24	6	8
	TEX2	1678	19	4	15
	TEX3	4264	49	13	36
	BEOF	6841	98	24	74
	CDRF	11028	136	38	98
	ACCF	10168	156	30	126
	Total	36663	482	115	367

Table 4.1: Corpus description

justified by the impersonal nature of the texts. The data analysis shows that more than 3/4 of the English pronouns are neuter (*it* or *they*). Similar analysis is not suitable for French, where all pronouns have masculine or feminine grammatical gender[12].

4.4.5.2 Complexity of the corpus

The complexity of a corpus only takes into account those features that could make it more or less difficult to process by an anaphora resolver.

[12]The generic *on* excluded.

	#pars	#sents	#NPs			#pronouns		
			#def	#indef	#prep	personal	possessive	reflexive
English	1277	2573	16826	3543	4005	466	68	11
French	883	1552	13247	2873	2601	376	89	17

Table 4.2: Composition of the corpus

The complexity of a corpus annotated for anaphoric relations is defined in terms of several measures: the first one is the average distance between the anaphors and their antecedents (computed in number of intervening noun phrases). This distance has to be computed separately for different types of anaphors, since they all have different referential power. A possessive pronoun, for example, has a more restricted referential span than a personal pronoun, while reflexives mostly refer within the same sentence. The intuition is that the greater the distance between an anaphor and its antecedent, the more difficult that anaphor will be to be solved. Again, this measure has to be computed separately for each language.

The second measure is similar, except that the distance between an anaphor and its antecedent is computed in number of sentences.

The third measure is the proportion of intra-sentential anaphors vs inter-sentential anaphors.

The figures presented in the following table should only be taken as approximate values, since they have been computed automatically, following the pre-processing of the texts with a shallow parser. Some errors in the pre-processing may have been introduced in the computation of these values.

		Avg NP distance	Avg sentence distance	#Intrasentential
English	ETH	5.1	0.7	35.07%
	ACC	4.24	0.61	60.2%
	CDR	4.5	0.5	53.6%
	WIN	3.95	0.17	80%
	PAN	2.4	0	100%
	AIW	2.26	0.13	54.1%
	Average	4.43	0.54	48.26%
French	TEX1	2.3	0.21	88.8%
	TEX2	3.7	0.43	73.3%
	TEX3	3.43	0.29	80.5%
	BEOFR	2.98	0.28	63.51%
	CDRFR	4.1	0.39	69.38%
	ACCFR	4.1	0.43	69.04
	Average	3.81	0.31	70.29%

Table 4.3: Complexity of the corpus

4.4.5.3 Preliminary observations

A shallow analysis of the composition of the corpus reveals a number of preliminary considerations for the design of the system. The first observation is that the number of intrasentential anaphors is very high, amounting to 2/3 of the total number of anaphoric pronouns, and the number of cases where the distance between the anaphor and the antecedent is greater than 1 is practically negligible. Intuitively, this leads to two conclusions: the searching space for antecedents can be set to a limited number of adjacent sentences and no special consideration has to be given to discourse structure. Of course, relaxing these heuristics will help solve the more difficult cases of anaphora.

The second observation relates to the number of reflexive pronouns in the corpus. The data analysed contained a very small number of reflexives, that could all be solved using syntactic rules. This leads to the conclusion that the resolution of reflexives can and should be performed separately from the resolution of the other pronouns, and machine learning is not necessary in this case.

The large number of non-anaphoric pronouns makes it necessary for a pleonastic pronoun identifier to be incorporated in the system.

4.5 Corpus annotation

4.5.1 Overview

The annotation described in this section refers to two types of marking, that serve different purposes. The first type is the manual marking of coreferential expressions, which is used in the gold corpus (see section 6.2.1 for a discussion on gold corpora) employed for the evaluation of the system, and also in the training of the machine learning module.

The second type of mark-up is generated automatically during the execution of the system. This type of mark-up is not only generated for the gold corpus, but for any text submitted for analysis to the system. As it will be elaborated later in Chapter 5, the anaphora resolution models interface with the pre-processing modules through an XML-encoded text. This XML annotation is produced by merging the manual annotation (for coreference) and the automatic annotation produced in the pre-processing stage. This section describes the annotation procedure employed for the production of both types of annotation.

4.5.2 Manual coreference annotation

Although this corpus was developed for use in automatic resolution of pronominal anaphora, the annotation is similar to those used in coreference resolution. Encoding the full coreferential chains is important in both evaluation (where a pronoun is considered correctly solved if it is resolved to any NP in its coreferential chain) and training (where positive examples are generated using a pronoun and all the coreferential NPs, while the negative examples are built from non-coreferential NPs). For this reason, corpora annotated for anaphoric links only (pairing a pronoun with its most recent antecedent) are not considered suitable for the task.

4.5.2.1 Annotation scheme

For annotating the corpora, the MUC7 annotation scheme for coreference [Hirschman1997] has been initially adopted . Various researchers have commented on the disadvantages of that scheme, pointing to the shortcomings inherent in the restriction of the coreference task to the relation of identity only. In addition, the MUC-7 annotation scheme assumes markable elements to be continuous strings. It is therefore impossible to capture relations between plural pronouns and discontinuous antecedents as in:

(4.1) *John* goes to college on Mondays, *Mary* goes on Tuesdays, and *they* both go on Wednesdays.

In this respect, it was probably not the best scheme to be adopted, but it did present several advantages. The most important was that some of the data that was selected for using in ARMAL had already been annotated using this scheme. The second one was the availability of CLinkA, an annotation tool that could only handle the MUC annotation scheme.

The following example shows how the MUC annotation scheme is applied on a French text:

(4.2) <COREF id="e131">L'expression oeuvre scientifique</COREF>, objet de notre étude ne <COREF id="e132" ref="e131">se</COREF> laisse pas facilement appréhender par le droit. On peut <COREF id="e133" ref="e131">lui</COREF> donner un sens très général et considérer que l'expression vise toute production intellectuelle de caractère scientifique.

However, at a later stage the decision was taken to switch to a variant of the ELRA scheme[13] that allows for a more comprehensive annotation. All discourse entities are marked by an *EXP* tag, which contains an unique identifier. Referential expressions are marked by an *PTR* tag, with an *SRC* attribute taking as value the identifier of the discourse entity referred to by the referential expression. In the case of anaphors with split antecedent, the *SRC* tag holds a list of IDs corresponding to each component of the antecedent. The *TYPE* attribute contained within a *PTR* tag holds the type of the anaphoric relation, and it can take one of the values: *coref*, *indirect*, *one*.

The example below shows the same text as in example (4.1) annotated with the ELRA scheme:

(4.3) <exp id="e131">L'expression oeuvre scientifique</exp>, objet de notre étude ne <exp id="e132"><ptr type="coref" src="e131"/>se laisse pas facilement appréhender par le droit. </s><s> On peut <exp id="e133"><ptr type="coref" src="e131"/>lui</exp> donner un sens très général et considérer que l'expression vise toute production intellectuelle de caractère scientifique.

[13] This variant was adopted during a collaborative project between the University of Wolverhampton and Université de Grenoble III dealing with the construction of a bilingual annotated corpus.

A text containing split antecedents will be annotated as:

(4.4) <exp id="e10">*John*</exp> goes to college on Mondays, <exp id="e11">*Mary*</exp> goes on Tuesdays, and <exp id="e12"><ptr type="coref" src="e10 e11">*they*</ptr></exp> both go on Wednesdays.

A larger annotated text extracted from our corpus can be found in Appendix A.

The adoption of the new scheme meant that the texts already annotated had to be converted. This was done with the aid of an automatic converting tool that transformed the MUC annotation scheme into the ELRA annotation scheme. The relations that have not been previously annotated due to the unsuitability of the MUC scheme have been added manually in the new annotation.

4.5.2.2 Markables

As a general rule, the annotation has to encompass all coreferential chains present in a document. Therefore all entities which could be potentially coreferential were annotated[14]:

- definite descriptions (pre-modified by either a definite article or by a demonstrative)
- referential indefinite descriptions (I have *a cat. It* is tortoise shell.)
- referential pronouns - including third person personal pronouns, reflexives, possessives
- possessive determiners
- referential indefinite pronouns (such as *one, another*)

[14]Even though this could mean annotating coreferential chains that contain only one element

The following expressions are not considered markables and therefore they are not annotated:

- non-referential personal pronouns such as pronouns of first and second person
- pleonastic pronouns

  (4.5) It rains.

  They have lessons in everything these days.
- pronouns referring to a verb phrase, adjectival or adverbial phrase
- non-anaphoric indefinite descriptions

  (4.6) Have you got *a cat*?[15]
- interrogative and relative pronouns: *who, what, which,that*
- noun phrases in wh- constructions: *Which car*

When annotating a noun phrase, all the modifiers will be subsumed by the annotation, including appositional phrases, relative clauses, prepositional phrases and any other type of pre- or post-modifiers.

4.5.2.3 Relations

The only anaphoric relation that presents interest for this project (and, therefore, the only one that was marked) is *identity of reference anaphora*, the type of anaphora where the proform and the antecedent refer to the same real world object. Hence, cases of indirect anaphora, one anaphora, or relations between numerical values associated to the same entity have not been marked. These are cases of *identity of*

[15]It is assumed here that the indefinite NP *cat* will not be further referred to. For example, in a situation like ”Have you got a cat? Yes, it is a blue persian.”, *a cat* will be considered markable. Such situations make it more difficult to identify markables incrementally.

sense anaphora, as in the following examples:

(4.7) I have *a cat*. My sister has got *one* too. (not the same cat)

(4.8) The temperature of the sea was *17 degrees Celsius* yesterday. Today, it is *20 degrees Celsius*. (both values of the temperature of the sea, but not coreferential)

4.5.2.4 Annotation procedure

The preliminary stage of the annotation process consisted in drawing clear annotation guidelines that specified the markable entities, the type of relations to be annotated and solutions for dealing with difficult or border cases.

The initial MUC-based annotation was done with the aid of CLinkA [Orăsan2000], a visual annotator that allows for the easy marking of entities and links and produces SGML output. The ELRA type of annotation was performed using PALinkA, a general visual annotator developed by Constantin Orăsan at the University of Wolverhampton, providing a much more extensive range of relations.

Some of the English texts were annotated by two annotators closely following the guidelines. This was meant to ensure better quality of the annotation. Unfortunately, such a process is both time consuming and expensive, therefore it could not be applied for the whole corpus. The French texts and the remaining English texts were annotated by only one annotator.

When performing manual annotation for corpora used in evaluation, a whole range of problems have to be addressed, mostly related to the quality of the corpus produced. The inter-annotator agreement for the texts annotated by more than one person, was computed in terms of precision and recall, as defined in Chapter 6. The precision obtained was of about 60%. Although these figures seem low, the results are deceptive,

since most cases of disagreement appeared in the identification of noun phrases and especially in their delimitation[16]. These cases were easily solved through mutual consent. The identification of the anaphoric relations was done with almost perfect accuracy.

4.5.3 Automatic annotation

This section addresses the type of markup added to the corpus by the anaphora resolution system in the pre-processing stages. The exact procedure employed and the pre-processing tools that add certain types of annotation will be described in Chapter 5. This subsection only deals with the description of the actual XML annotation.

Segmentation markers

Texts are marked for multi-paragraph segments, paragraphs and sentences. As a general observation, all segmentation units are delimited by a tag indicating their type. Each tag contains a compulsory **ID** attribute that holds the index of the respective unit within the larger segmentation unit. Table 4.4 describes the tags and the attributes used for marking the segmentation units.

Unit	Tag	Attribute	Value
Segment	SEG	ID	index of segment in text
Paragraph	P	ID	index of paragraph in segment
Sentence	S	ID	index of sentence in paragraph

Table 4.4: Marking of segmentation units

[16]This is consistent with the findings of [Hirschman1997]. To avoid such problems, in some annotation projects, the decision has been taken to provide the annotators with the noun phrases already identified in text. In this way, the only differences in annotation will appear at the level of the identification of the anaphoric relations.

Title marking

Titles, subtitles and section/subsection headings are marked using the **TITLE** tag. The **ID** attribute of each TITLE tag holds the index of the title in the document.

Lists marking

Constructions that appear in the form of lists are marked using HTML-like markup. The block of listed elements are delimited by **LIST** tags. Each element in a list is included in an **ITEM** tag. Each ITEM has an ID attribute which uniquely identifies the item within a list.

Marking non-parsable entities

In this category are included constructions that are not considered important for the anaphora resolution process: tables, examples and commands[17]. They are marked with a generic **NONP** tag. No other attributes are associated to such tags, as their only role is to indicate that they have to be omitted by the anaphora resolver.

Marking linguistic units

The linguistic units marked are individual words and noun phrase compounds. A noun phrase (NP) is a sequence of WORD elements that has as attributes morphological and syntactical features, such as number, gender, grammatical function, syntactic dependency. They are also identified through and ID element and their position within the sentence is stored in a POS attribute.

4.5.4 The DTD

The DTD presented below represent the XML tree with the root *doc*, including the elements described above and their attributes.

[17]The decision of what is not ”important” was based on the likelihood of an antecedent to be found within certain constructions. There is no guarantee that a pronoun will not refer to a noun phrase contained in a table, or an example, however the probability of such phenomena occurring is low. The difficulties introduced by including tables and examples in the processing supersede by far the potential advantages.

```
<!-- document structure>
<!ELEMENT doc (seg+)>
<!ELEMENT doc (p*)>
<!ELEMENT p (s*)>
<!ELEMENT s (np*|w*)>
<!ELEMENT np (w*)>
<!ELEMENT np (#PCDATA)>
<!ELEMENT title (np*|w*)>
<!ELEMENT nonp (#PCDATA*)>
<!ELEMENT list (item*)>
<!ELEMENT item (s*)>
<!ATTLIST seg id ID #IMPLIED>
<!ATTLIST p id ID #IMPLIED>
<!ATTLIST s id ID #IMPLIED>
<!ATTLIST np id ID #IMPLIED>
<!ATTLIST w id ID #IMPLIED>
<!ATTLIST list id ID #IMPLIED>
<!ATTLIST item id ID #IMPLIED>
<!ATTLIST title id ID #IMPLIED>


   <!-- coreferential annotation structure>
<!ELEMENT exp (ptr*, ptr-i*, (#PCDATA|exp)+>
<!ELEMENT ptr EMPTY>
<!ELEMENT ptr-i EMPTY>
<!ELEMENT seg (#PCDATA|exp)+>
<!ATTLIST exp id ID #IMPLIED
```

```
<!ATTLIST ptr type (coref)
    src IDREFS #REQUIRED>
```

4.6 Conclusions

This chapter has described the use of annotated corpora in anaphora resolution, and in particular in the ARMAL system. Far from being an easy task, the annotation of anaphoric links poses a number of problems to the human annotator, which inevitably reflect upon the quality of the corpus produced.

It has been shown that using technical manuals as the target of the anaphora resolver is a justifiable choice in the context of domain-specific NLP applications and the XML annotation scheme used in marking the corpus has been described.

The corpus described in this chapter represents on its own a by-product of the research, being a valuable resource for further experiments on anaphora resolution. A preliminary analysis of the corpus composition and the distribution of pronouns in the corpus revealed some possible directions to be taken in the implementation of the anaphora resolver.

An example of a text annotated with coreferential links, as well as with the special tags described in section 4.5.3 can be found in Appendix A.

Chapter 5

A system for bilingual anaphora resolution

5.1 Overview

This chapter describes the design and implementation of ARMAL, a system for multilingual pronoun resolution based on an original approach. The anaphora resolver combines high confidence handcrafted rules with machine learning techniques.

The decision of using machine learning for the task is justified by current trends in NLP and by characteristics of the problem. Two main types of supervised learning methods are described: lazy learning, represented by memory based learning, and eager learning, represented by decision trees. The same learning features are used for both methods, therefore the design of the system is not affected by switching to either of the learning engines. The analysis of these two methods shows that their disadvantages can be minimised by using an optimisation of memory based learning where the instances are stored as a decision tree. However, the main disadvantage of supervised learning remains the fact that they require for training a prohibitive

amount of annotated data. This has justified an attempt to design a less expensive unsupervised method. Section 5.4 presents a genetic algorithm (GA) as a cheaper and faster alternative to the main machine learning method. The evaluation of the GA in Chapter 6 will explain why it is not a suitable method for anaphora resolution, despite its apparent advantages.

While the first two sections of this chapter deal with the design of ARMAL, section 5.5 describes the actual implementation of the anaphora resolution system. The pre-processing stages are described in some detail, as are the technical implementation and the graphical interface.

5.2 Machine learning for anaphora resolution

Machine learning is a branch of Artificial Intelligence that deals with programs that learn from experience. As the main part of this approach to anaphora resolution is based on machine learning, this section provides a basic introduction in the field, with a deeper insight into instance-based and decision trees learning, which are used by the described system.

5.2.1 Brief outline of machine learning techniques

The main difference between symbolic and machine learning methods lies in the fact that while the former use handcrafted rules (sometimes built intuitively, other times deduced from previous experience), the latter predict the outcome of a problem by automatically generalising from previous examples, therefore being able to improve with experience. The difference between various machine learning methods appears in the way this generalisation is performed, in the storage and processing of previous examples and in the similarity methods used for classification.

Although different in methodology, statistical methods share with machine learning techniques the same basic philosophy of learning from examples. However, while statistical methods assume a certain distribution of the training data and employ explicit probabilistic measures, machine learning assumes them implicitly. A large number of classifiers have been built on statistical methods, including the Bayesian classifiers, Hidden Markov Models, Maximum Entropy Model and many others.

A wide range of problems have proved suitable for solving using machine learning, including all types of optimisation problems, financial, meteorological and medical predictions, information filtering, data mining, games playing, autonomous driving and many others.

5.2.1.1 General classification scheme

There are three main steps in the design of a general supervised learning algorithm: (a) defining the target function, (b) specifying the training space and technique and (c) deciding the classification procedure.

In designing the target function, one has to specify which is the concept to be learned and what kind of knowledge is necessary for learning it. The domain of definition of the target function will be a space of hypotheses and the result will be in a set of classes (i.e., the classes that the instances will be classified into). The target function itself will be a combination of attributes (extracted from the knowledge sources available) and weights (reflecting the relative importance of features).

In defining the training space and technique, one has to take into account the type of knowledge available for training and the representativeness of the training set, i.e., how close the distribution of the training examples is, compared to the expected distribution of unseen cases. If an anaphora resolver, for example, is trained only on instances that are generated from plural reflexive pronouns, it will most certainly fail to solve real

texts, where the frequency of such pronouns is very low.

5.2.1.2 Weighting measures

In both lazy and eager learning, a central problem is represented by the definition of measures of similarity between training and testing instances. In the most simplistic scenario, a similarity metric could take into account the number of differences in attributes over two instances. Such a measure is called *overlap metric* and it has the disadvantage that it considers all attributes to be equally important. This is however not satisfactory in most real learning problems, and especially in Natural Language Processing, where some linguistic phenomena have to be considered more important than others. Therefore, a number of feature weighting measures are used by different learning algorithms, including *chi-squared weighting*, *exemplar weighting*, *information gain* and *gain ratio*. The last two weighting measures will be briefly defined in the following, as they are used by the learning algorithms in the implementation of ARMAL.

In the definition of weighting measures, a basic concept is the *entropy* of a set of examples as a measure of the impurity of that set of examples. If all the examples are in the same class, the entropy will be 0, while if there is an equal number of positive and negative examples, the entropy will be 1. For any other distribution of positive and negative examples, the entropy will take a value between 0 and 1. Several weighting measures are based on evaluating the effect of the classification on the entropy of a collection of examples.

Information gain

The information gain of an attribute *A* relative to a collection of examples *S* is

calculated as:

$$Gain(S,A) = Entropy(S) - \sum_{v \in Values(A)} \frac{S_v}{S} * Entropy(S_v) \quad (5.1)$$

where:

- Entropy(S) is the entropy of the initial collection

- Values(A) is the set of possible values for the attribute A

- S_v is the subset of S for which the attribute A has value v

Intuitively, the equation above states that the information gain is the expected reduction in entropy, given that the attribute A is known. Therefore, the better A classifies the set of examples, the smaller the remaining entropy will be, because the number of classes decreases.

Gain ratio

The gain ratio weighting measure [Quinlan1993] is a normalised version of the information gain. The latter favours tests with many outcomes (corresponding to attributes that can take many values). Although the partition obtained on the training set is maximal, the predictive power of such a division is minimal. The gain ratio normalises the gain attributed to each test, thus rectifying the apparent gain of tests with many outcomes.

A function called *split information* is used, that measures the potential information generated by dividing the training set S into n subsets. Consequently, the gain ratio of an attribute A is defined as:

$$Gainratio(A) = \frac{gain(A)}{split(A)} \quad (5.2)$$

5.2.1.3 Memory based learning

Memory based learning[1] represents probably the most successful type of learning methods used in NLP. In memory based learning, a set of training instances are stored in memory, and unseen cases are classified by using similarity metrics computed on the stored examples. For classifying an unseen instance, this method requires a similarity measure for evaluating the importance of examples. The general name of *lazy learning* given to this class of algorithms reflects the fact that almost no computation is done on the training examples until a new instance is presented for classification. This allows for local approximations of the target function to be computed for each new instance. However, this also implies that the computational cost of employing such a learning method will be very high, since all computation is performed at classification time. This contrasts with the eager type of learning, such as decision tree learning, where the learning algorithm is committed to a single representation of the target function, computed over the whole set of training instances. Consequently, on the same classification problem lazy algorithms will be significantly slower than eager ones. The same considerations make the problem of *indexing* the training instances a major issue in memory based learning.

Most memory based learning algorithms are direct descendants of a widely-used algorithm, the *k-Nearest Neighbour*. In the k-Nearest Neighbour algorithm, an unseen instance is compared to the training instances in terms of the Euclidean distance and selects the k closest instances. The test instance is classified in the most common class among the selected training instances.

[1] Also known as *instance-based, example-based, exemplar-based, case-based or analogical learning*

5.2.1.4 Decision Trees

Decision tree learning is a method for approximating discrete-value target functions, in which the learned function is represented by a decision tree. This method comes into the category of *eager learning* methods, since it builds a concept description by generalising from the training data. This type of machine learning is suitable for problems that present certain features [Mitchell1997]: the instances can be described as (attribute, value) pairs, there are few possible values for an attribute, the response is a value in a discrete set of elements (usually *yes/no*).

A decision tree contains intermediary nodes and leaves (terminal nodes). Each intermediary node specifies a test of an attribute of the instance and each branch descending from that node corresponds to one of the possible values for this attribute.

An instance is classified by starting at the root node of the tree, testing the attribute specified by this node, then moving down on the branch corresponding to the answer to the test. This procedure is repeated recursively until the example is classified. Since a decision tree is a disjunction of conjunctions of predicates, it can be translated for easy reading in sequences of if-then statements. An important issue in learning decision trees is the order of the attributes according to their importance as classifiers. This is reflected in the tree, the best attribute appearing in the root node, the second best on the second level and so on. Several measures have been employed for evaluating the worth of an attribute, a good measure being the *information gain*. Another problem in building decision tree classifiers is dealing with the overfitting of the training data. This situation appears when a decision tree is built that partitions the training set correctly but has got low predictive power, performing poorly on unseen cases. The result is often a complex decision tree that infers more structure than is justified by the training cases. The simplification of these trees can be done by prunning, i.e.

removing those branches that do not contribute to the classification accuracy on unseen cases and replacing them with leaves. The branches that are dropped are selected according to a statistical measure (c2 test of statistical significance, information gain, error-reduction). The resulting classifier almost always performs worse on the training data but has higher accuracy on unseen cases.

Most of the decision tree-based algorithms are variations on a core algorithm. Quinlan's ID3, and its successor, C4.5 [Quinlan1993], are the best known decision tree classifiers, being successfully used in a large range of applications, from financial predictions to life expectancies in patients with certain medical conditions. The first improvement that C4.5 brings to the generic decision tree classifier is seen in the induction strategy, by substituting the information gain criterion with the gain ratio (see section 5.2.1.2).

A useful feature of the decision tree classifiers is that they can deal with unknown attributes. If an instance contains missing attributes, a number of problems arise: if this example occurs in the training data, how is it used in partitioning, since it cannot be associated a certain outcome? And if the example occurs in the evaluation set, how is it classified? C4.5 adopts a probabilistic approach to solving these problems. It mainly assumes that unknown test outcomes are distributed probabilistically according to the relative frequency of known outcomes. A case with an unknown test outcome is divided into fragments whose weights are proportional to these relative frequencies, with the result that a single case can follow multiple paths in the tree. This method is employed both in partitioning the training set and in classifying an unseen case.

Regarding the pruning method, C4.5 employs a pruning technique based on error-reduction, where error estimates for leaves and subtrees are computed assuming that they were used to classify a set of unseen cases of the same size as the training set. If replacement of a subtree with a leaf would lead to a decrease in prediction error,

pruning is performed on that subtree.

In anaphora resolution, decision trees have been the chosen learning method in [Aone and Bennett1995], [McCarthy and W.Lehnert1995] and [Meng et al.2001]

5.2.1.5 Combining lazy and eager learning

The previous sections showed that the main difference between lazy and eager learning is in the way and moment the learning hypotheses is formulated, this leading to an increase or decrease in speed as well. The advantages of each set of methods can be exploited in hybrid systems that use the better indexing technique of decision trees with the higher freedom of learning of memory based methods. Such a hybrid method is used in the implementation of ARMAL, as it will be shown in section 5.3.

5.2.1.6 Genetic algorithms

Genetic algorithms have been introduced by John Holland at the University of Michigan in the 1960s [Holland1975], and have developed since as a highly successful method of tackling optimisation problems.

Genetic algorithms draw on the general mechanisms of natural evolution. They represent hypotheses as strings whose interpretation depends on the specific application. The evolutionary process starts with a population of chromosomes (hypothesis) and in each generation some of the hypotheses are allowed to survive through a process of selection, combine through crossover and mutate. The decision of which hypotheses should survive is based on their quality as genetic material, assessed by means of a fitness function. The genetic operators ensure that the most valuable part of a chromosomes will be preserved in the next generation, and at the same time allow for new genetic information to appear in the population.

The main issues in building a genetic algorithm tailored to a certain problem are

the choice of a data encoding and of a fitness function. Related decisions are the choice of the genetic operators (or the design of new ones), of the population size and of the number of generations in the evolutionary process.

In natural language processing, genetic algorithms have not been as extensively used as other machine learning methods. The main applications of genetic algorithms have been in information retrieval, for improving queries, [Yang1993] and learning syntactic rules, as in [Losee1996]; morphology, for word categorisation [Lankhorst1994] and segmentation [Kazakov1997]. [Han2001] uses genetic algorithms for constructing bilingual dictionaries.

To our knowledge, there have been so far only two attempts of using genetic algorithms for anaphora resolution. [Orăsan et al.2000] used genetic algorithms for finding the best combination of weights for the indicators used within the implementation of Mitkov's knowledge-poor approach. Therefore, their method is not an independent anaphora resolver and, moreover, it makes use of coreferentially annotated data. [Byron and Allen1999] presents a genetic algorithm that calculates the best weight assignment for a set of "voters". A voter is a module that indicates a preference for a certain noun phrase to be chosen as antecedent for a pronoun. The genetic algorithm identifies the weights by running a pronoun resolution on a training corpus (extract from the Penn Treebank and annotated with coreferential links) and identifying the best possible combination of weights for all voters. Although the results reported are good (71% overall success rate), showing a slight improvement over Hobbs' naive algorithm on the same test set, it has to be noticed that only pronouns with simple nominal antecedents were included in the evaluation; therefore, pleonastic pronouns, those referring to verb phrases or clauses, plural pronouns with split antecedents were not solved.

5.2.2 Anaphora resolution as a classification task

Several authors [Roth2000, Cardie and Wagstaff1999] promote the view that machine learning techniques are especially suitable for solving ambiguity problems in natural language processing, which can be recast as classification tasks. Machine learning was so far extensively used in different fields of NLP, such as speech recognition, syntactic parsing, part of speech tagging, word sense disambiguation and named entity recognition.

What makes anaphora resolution a suitable task for using machine learning techniques is the possibility of formulating the problem as a classification task. Basically, identifying the antecedent of a pronoun means classifying its possible antecedents (a restricted set of the noun phrases in the text) into those that are true antecedents and those that are not. A different approach to the classification problem could be to partition the set of possible antecedents in several classes according to their likelihood of being antecedent for a specific pronoun.

A possible reason why rule-based methods do not seem to perform above an 80% threshold could be the fact that they do not deal efficiently with the exceptions. Most of the times, the exceptions are extremely varied, and dealing with special cases can be detrimental to the overall algorithm. Machine learning methods, on the other hand, learn from a large number of examples, therefore they can learn exceptions as well as common cases.

Another advantage of using machine learning is the fact that they provide a more effective and scientifically correct way of dealing with missing attributes, i.e, it is possible to infer from underspecified examples.

Chapter 3 has presented some of the best known anaphora resolvers that make use of machine learning. The most popular choice of supervised learning algorithm

seems to be the decision trees classifier, while unsupervised methods have not been extensively used due to their lower performance.

5.3 A hybrid anaphora resolver

5.3.1 General description

This section introduces ARMAL, a pronoun resolution system based on a hybrid approach[2]. It combines a set of filtering rules and a machine learning module. Experience shows that pronouns display different degrees of difficulty of solving. Certain cases of pronominal reference are easily resolved without employing specialised methods. CogNIAC, the system proposed in [Baldwin1997] employs a few high confidence rules for identifying antecedents with a high precision. These rules are consecutively applied to the set of possible antecedents of a pronoun; if one of the rules can be applied, that pronoun is solved, otherwise no resolution is attempted. This kind of resolution is done with low recall, given that most pronouns will not be attempted to be solved. Based on CogNIAC's results, it seems reasonable to believe that, by applying high confidence rules for solving certain classes of anaphors, and applying a machine learning method on the set of pronouns that couldn't be solved in the first stage, the performance of the system would improve. Furthermore, although a method based purely on machine learning could probably perform equally well in terms of accuracy of resolution, the computational cost involved would be higher.

[2] A preliminary design and evaluation of the system has been previously published in [Barbu2001].

5.3.2 Filtering rules

The rules included in the preliminary stage are designed to eliminate from automatic learning pronouns that have a limited number of possible antecedents, out of which only one has a much higher probability of being correct than the others. Some preliminary experiments have been performed with a set of rules in order to identify the most successful ones. In selecting the filtering rules, the focus was on the precision of resolution, and not on the recall.

A set of four rules were used for filtering antecedents in both English and French:

Rule 1: Unique antecedent: if there is an unique noun phrase in the discourse that can be antecedent for a pronoun, select it as antecedent.
This is an obvious, straightforward rule that will fail in a restricted number of cases, mainly when the pronoun is non-referential, exophoric or does not agree morphologically with its antecedent.

Rule 2: Unique subject: if the pronoun is a subject and there is only one subject in the previous portion of discourse, select that subject as antecedent.
This rule models the intuition that a pronoun subject is likely to refer to a previous subject NP. If the previous portion of discourse contains more than one subject, the problem becomes more complicated, and it is often necessary to use other types of information in order to disambiguate the pronoun.

Rule 3: Unique in current and previous sentence: if there is only one possible antecedent in the previous sentence and in the read-in portion of the current sentence, select that NP as antecedent.
This rule is based on the empirical observation that long distance anaphora is an

unfrequent phenomenon, and two sentences can be considered a reasonable searching space.

Rule 4: Reflexive pronouns: if the pronoun under processing is a reflexive, select as antecedent the subject of the verb that both pronoun and antecedent are arguments of.

The rationale of this rule is that reflexive pronouns obey strict syntactic rules that make their resolution easier than that of other types of pronouns. According to GB, a reflexive has to be bound in its governing category. As a general rule, reflexive pronouns are bound by the subject of their clause. The second reason for not subjecting the reflexives to resolution by machine learning is the fact that their frequency in real texts is quite low, therefore requiring a huge amount of texts for extracting a significant number of training examples.

Another two rules are specific to French only:

Rule 5: Subject Clitics

If a clitic is subject in an inverted interrogative construction, select as antecedent the subject of the same verb the clitic is an argument of. This rule is justified by the fact that in this situation, the clitic is redundant. The same applies for subject clitics in a clause introduced by a detached NP.

(5.1) *Le chat*, dort-*il*?

(5.2) *Ta soeur*, *elle* est merveilleuse.

Rule 6: Redundant clitics

Clitics serving as direct or indirect object to a pronominal verb always refer the subject of the verb (cf. section 2.2.1).

(5.3) *Il* doit *se* diriger vers l'école.

The rules 1 to 3 are applied in this order for English texts, while for French texts, the language-specific rules are applied before the rules 2 and 3. The order of applying the rules is based on their discrimination power.

5.3.3 Disjunction rules

While the filtering rules work on reducing the number of pronouns subjected to the automatic classifier, the role of the disjunction rules is to eliminate, for each pronoun, those noun phrases that cannot be antecedents for that pronoun.

- **Disjunction rule 1** A non-reflexive pronoun cannot corefer with a co-argument. In the example below, *John* and *him* cannot be coreferential.

  (5.4) *John* saw *him*.

- **Disjunction rule 2** A pronoun cannot corefer with a noun phrase that contains it. (*His* and *his father* cannot corefer)

  (5.5) John met *his father*.

- **Disjunction rule 3 (French only)** A clitic pronoun cannot co-refer with an NP in the same subordinate context. (*le bureau* and *il* cannot corefer)

  (5.6) Paribas s'appuiera pour cela sur *le bureau de représentation* qu'*il* a ouvert en 1966.

5.3.4 Machine learning module

5.3.4.1 Technique

Building an application based on machine learning consists of three important steps: design, training and learning.

In the first step, the problem to be solved has to be clearly identified, by defining the working hypothesis and the expected outcome. Secondly, the learning features have to be identified. The design stage is followed by the actual implementation, where the training set is defined, and the training examples are generated. Following the generation of training examples, a learning algorithm has to be selected and applied according to its rules.

The choice of a specific learning algorithm for using in ARMAL has followed some experiments with Bayesian classifiers, C4.5 decision trees and memory based learning. The memory based learning algorithms investigated were those included in the TIMBL 4.0 distribution.

TIMBL[3] [Daelemans et al.2000] is a package of programs developed at Antwerp University, containing several memory based learning algorithms. TIMBL provides a straight forward way of obtaining preliminary results and allows experiments with learning parameters for optimised results. Experiments with some of the classifiers included in the package (nearest neighbour search, information gain tree (IG-Tree)) showed that the best results (although not significantly better) are obtained when using TRIBLE, a hybrid generalisation of decision trees and k-nearest neighbour search. The experiments were performed using the default settings of the algorithms.

The decision trees learning was done using the free distribution of the classical C4.5 algorithm (revision 8)[4].

5.3.4.2 Engine

Learning hypothesis

The system tries to learn a function defined on the set of pairs (*pronoun*, *noun phrase*)

[3]The acronym stands for *TIlburg Memory Based Learning*.
[4]Available from http://www.cse.unsw.edu.au/quinlan/

that can take the values *yes* or *no*, depending on whether the noun phrase is the correct antecedent for the pronoun or not.

The function takes into account a set of fourteen core features and 6 language independent features that will be described below.

Training

The data used for training consists of 90% of the corpus described in Chapter 4, while the remaining 10% is used for evaluation. The size of the training corpus is an important factor that influences the performance of a machine learning system, since not enough data leads to overfitting and decreases its performance. The amount of training data depends on the number of features used in classification.

5.3.5 Training/learning features

5.3.5.1 Choosing the features

The second step in the system design was the selection of the features used for training. In order to do this, an experiment was performed on part of the bilingual corpus, by selecting the pronouns that are ”equivalent” in the two languages[5] and finding the characteristics that help identifying their antecedents. This experiment showed that some rules that have been successfully applied for English do not hold for French (for example, collocation patterns will have to be implemented differently). Nevertheless, the experiment helped identifying a set of 14 core features, that apply for both English and French.

It has to be mentioned that, although there are more factors that contribute to

[5]By *equivalent* it is meant a pronoun that is translated as the same type of pronoun in the other language. For example, if the French ”il” is translated in English as ”it”, they will be considered equivalent, whereas if translated as ”this”, they will not be considered equivalent.

solving anaphora, the goal is to develop a fully-automatic system, so only those features that can be computed automatically with a fair degree of accuracy are to be selected. For this reason, none of the features included needs advanced knowledge sources, and even the semantic class of a noun, which can be extracted from WordNet, has been avoided. Although some authors consider that selecting the first meaning of a word, as appearing in WordNet is sufficient, this can be inaccurate, especially for technical/scientific texts, where the correct meaning of the word may not be the one indicated as the most frequent by WordNet. The large number of words without a description in WordNet was another reason for avoiding using it.

5.3.5.2 Core features

The machine learning method uses for training and learning a set of 14 core features. The set of features contains both unary features (that apply to either the pronoun or the antecedent) and binary features (that involve both the anaphor and the antecedent).

Unary features

The unary features (summarised in table 5.1) apply to either the pronoun or the antecedent. They take into account lexical, syntactical and textual features of the entity described[6].

- **Lexical features**

  Five types of lexical features have been considered. The *type of pronoun or noun phrase* can take the values *personal pronoun*, *possessive*, *reflexive*, *common noun*, *personal noun*. For French, the list of possible values is extended with the values for the locative clitic ”y” (*locative*) and the genitive clitic ”en” (*genitive*).

[6] In table 5.1, a feature name is marked with the indicator of the entity it applies to. An indicator *i* means that the feature refers to the pronoun, while *j* means that it refers to the noun phrase.

The *definiteness* is a feature applied to nouns only, and can take the values *yes* or *no*. Indefinite NPs are considered those noun phrases that are either not preceded by any article or preceded by one of the articles *a / un, une* or *some / quelque, quelques*. All other NPs are considered definite.

A *prepositional* feature can take the values *yes* or *no* and captures the fact that a noun phrase may or may not be part of a prepositional phrase.

The *gender* feature simply captures the grammatical gender of pronouns or nouns - it can take one of the values *neuter*, *feminine masculine*. Plural pronouns and nouns are always considered neuter. This feature is particularly suitable for French, that features grammatical gender. Its use in English anaphora resolution is much more restricted, as most if not all non-animate entities will be considered neuter.

The *number* feature stores the number of the pronoun or noun and can take one of the values *singular*, *plural* and *both*, where the latter is attributed to entities that can be referred at both singular or plural.

- **Syntactic features**

  In this category comes the *grammatical function* of the entity, which can take the values *subject, direct object, indirect object, oblique* and *other*.

- **Textual features**

  This category of features relate to the context of occurrence of the entity in text. The *frequency of appearance* feature reflects the number of times the noun phrase has been repeated in the current paragraph. Intuitively, this value reflects the noun phrase's probability of being in focus and therefore more likely to be referred to by a pronoun.

  The *anaphoric richness* feature was introduced in order to account for repetitive

references to the same NP within a paragraph; the reason for that is the fact that repetitive pronominalisation indicates that a certain discourse entity is kept in focus, therefore it is more likely to be referred to again by a pronoun. In the following paragraph, for example, the noun phrase *The Ethernet How-To* will have the anaphoric richness value of 2.

(5.7) <COREF ID="15" TYPE="IDENT" REF="1"> The Ethernet-Howto </COREF> covers what cards you should and shouldn't buy [...] <COREF ID="21" TYPE="IDENT" REF="1"> It </COREF> contains <COREF ID="22"> detailed information on the current level of support for all of the most common Ethernet cards available. <COREF ID="25" TYPE="IDENT" REF="1">It </COREF> does not cover the software end of things...

The *member of set* feature indicates the fact that the noun phrase it applies to is or is not part of a conjunction or disjunction of noun phrases. The selection of this feature was justified by the fact that, apart from special cases, noun phrases found in coordinated construction are not normally referred to individually by a pronoun. In the example below, each of the emphasised noun phrases will have their *member of set* feature set to *yes*:

(5.8) You know *the Daily Mirror, and the Sun, and ITV, and the Unions*...

Binary features

The binary features (see table 5.2) describe the link established between the anaphor and the antecedent in terms of morphological and contextual relationships.

The two *referential distance* features (*sentence distance* and *NP distance* take numerical values representing the number of intervening sentences and, respectively, noun phrases between the pronoun and the antecedent.

Feature	Values	Significance
morphological role$_{i,j}$	personal pronoun/ possessive/ reflexive/ noun	morphological role of the head of the NP
grammatical role$_{i,j}$	Subj / Direct obj / Indirect obj / Oblique / Adverbial	grammatical function of the NP or pronoun
$definite_j$	Yes/No	the NP is a definite description
$prepositional_j$	Yes/No	the NP is part of a prepositional phrase
$embedded_j$	Yes/No	NP embedded in another NP
$member\ of\ set_j$	Yes/No	included in a conjunction / disjunction of NPs
$gender_i$	FM/Neuter	singular pronouns can be either neuter or feminine/masculine; plural pronouns are considered neuter
$number_{i,j}$	SG/PL/Both	number of pronoun or NP; *Both* denotes entities that can be referred at both singular and plural
$frequency_j$	0/1/2/more	number of times the head of the NP appears in the paragraph
anaphoric richness$_j$	0/1/2/more	number of times the NP has already been referred by a pronoun in the current paragraph

Table 5.1: Training/learning unary features

The *agreement* features capture the *morphological agreement* in gender and number between the pronoun and the noun phrase and the possible *syntactic parallelism* (the pronoun and the noun phrase have the same grammatical function).

The *CT* feature takes one of the values 1,2,3,4 according to whether the centering relation established if the noun phrase is the antecedent for the pronoun is a continuation, retain, smooth shift or abrupt shift. A sentence is considered a unit and a paragraph is considered a segment. Special cases where sentences have no Cb (for example, at the beginning of the segment) are treated as *continuation*. The value ”0” means that the pronoun and the noun phrase are in the same sentence.

The *correlated* feature can take the values *yes* or *no*, depending on whether the

pronoun and the noun phrase are co-arguments of the same verb.

There are two positional features. The *collocation* feature, captures the fact that the pronoun and the noun phrase may be arguments of different instances of similar verbs, as in the following example:

(5.9) Print *the file*... print *it*.
Mount *the CD*, then unmount *it*.

The *cataphora* feature takes the value *yes* if the pronoun precedes the noun phrase, and *no* otherwise.

Feature	**Values**	**Significance**
NP distance	1/more	number of intervening NPs between the anaphor and the current NP
gender agreement	Yes/No	Pronoun and NP agree in gender
correlated	Yes/No	pronoun and NP are arguments of the same verb
CT	0/1/2/3/4	CT score
collocation	Yes/No	pronoun and NP appear in a similar context
cataphora	Yes/No	pronoun precedes the NP

Table 5.2: Training/learning binary features

5.3.6 Building the classifier

Given the training corpus, the noun phrases are determined automatically by running a sequence of pre-processing tools (as described in section 2.1.1). The training examples are pairs <anaphor, noun phrase>, extracted from the annotated corpora. It is important to feed the system both positive examples (pairs for which the NP is a

correct antecedent for the anaphor) and negative ones (pairs where the NP cannot be antecedent for the pronoun).

Positive examples are selected by choosing all the pairs <pronoun, NP> for which the NP is correct antecedent for the pronoun and it is found in a text span of 3 sentences behind the pronoun and in the same sentence as the pronoun (both preceding and following the latter).

The negative examples are constructed by pairing pronouns with all the noun phrases found between a pronoun and its closest antecedent in the text (as marked in the annotated corpus). For each pair <pronoun, NP>, a feature vector containing the features described in the section above is constructed. All features are computed automatically both in the training and the learning stage, therefore allowing for the introduction of noise (the full execution flow of the pre-processing modules will be presented in more details in section 5.5.3). The resulting feature vectors are sent to the classifier.

If one considers for example the simple text ”John went to the market. He bought apples.”, with the additional knowledge that the correct antecedent for *He* is *John*, the feature vectors will be as in figure 5.1.

<He, John> : [Pers_pron, PN, Subj, Subj, yes, no, no, no, FM, sg, sg, 0, 0, yes, 1, 1, yes, no, 1, no, no]

<He, the market> : [Pers_pron, CN, Subj, Adv, yes, yes, no, neuter, sg, sg, 0, 0, no, 1, no, no, 2, no, no]

<He, apples> : [Pers_pron, CN, Subj, Dir_obj, yes, yes, no, neuter, sg, pl, 0, 0, no, 1, no, no, 0, no, no]

Figure 5.1: Example of feature vectors

Table 5.3 presents the distribution of the positive and negative training instances in the training corpus.

File	Positive	Negative	Total	Average examples per pronoun
English	1426	4204	5630	10.33
French	867	3243	4110	11.19

Table 5.3: Training examples

5.3.7 Applying the classifier

Given a new text, the task is to identify the antecedent of the pronouns using the classifier described previously. First, the same pre-processing tools as in the training stage are applied on the text in order to identify the noun phrases. Each pronoun in the text is paired with all noun phrases appearing before it in a text span of 3 sentences and following it in the same sentence. The pairs are associated with feature vectors as described in the previous section. All the features are computed automatically. The feature vectors obtained are ordered according to the position of the noun phrase, starting with the one that is the closest to the pronoun. Then the vectors are sent in sequence to the classifier. If the classifier responds yes, the noun phrase is marked as correct antecedent and the process continues with the next pronoun. Otherwise, the next feature vector in sequence is sent to the classifier, and the process is repeated until there are no more noun phrases in the text. There is no difference in working mode depending on the learning method employed. For both memory based and decision trees learning, the general working flow applies.

A special comment is due for the computation of the *anaphoric richness* feature. When calculating the number of times a noun phrase has been referred to by pronouns, the number of pronouns solved so far to that noun phrase are taken into account and this value is incremented every time a pronoun has been found to refer to that specific noun phrase.

5.4 A genetic algorithm for anaphora resolution

5.4.1 Introduction

Previously, a supervised method for pronoun resolution has been described. The main disadvantage of the method was the fact that the anaphora resolver needs to be trained on an important amount of annotated data, which is expensive to build. Furthermore, adding new features implies using more and more data for training. In the following, a cheaper alternative to supervised learning will be presented: a genetic algorithm. As opposed to the supervised learning method, the genetic algorithm does not make use of annotated data for training, but its accuracy of resolution is lower[7].

5.4.2 System description

The genetic algorithm sees anaphora resolution as a multiple optimisation problem. The first aim is to identify the set of links <pronoun, NP> that maximises a certain target function, according to a number of syntactic and morphological features of the entities involved. The second aim is to determine the combination of indicators that allows for the best set of links to be identified, i.e. what weight should be associated to each indicator. Evolving the weights of the indicators as well means that they will differ from segment to segment. This method differs from other approaches to using weighted rules for anaphora resolution. The assumption is that this weight selection method is consistent with the intuition that in different contexts, the importance of indicators varies. The method is also consistent with the findings of [Orăsan et al.2000], where the authors noticed that optimal results are obtained when using different sets of weights for different testing files.

[7]The design and development of the genetic algorithm described in this section has been previously published in [Barbu2002]

5.4.2.1 Data representation

A chromosome has to encode two types of information: the <pronoun, antecedent> assignment and the <indicator, value> assignment. If a segment has P pronouns and N noun phrases and one wants to consider I indicators, a chromosome will be represented as a string of length P+I, holding real numbers. Each gene from position 0 to P-1 will hold an integer value representing the position of the selected antecedent for a pronoun. Therefore, chromosome[i] =j means that the i-th pronoun in the segment has been assigned as antecedent the j-th noun phrase in the segment. The genes from position P to P+I-1 will contain real values representing the weight associated to each indicator. Note that this kind of representation confines the search for antecedents within the limit of a segment. Figure 5.2 shows the representation of a chromosome for a segment containing 7 pronouns, with 6 indicators used for evaluation. The first seven genes encode <pronoun, antecedent> pairs, while the last 6 encode the weights of the features.

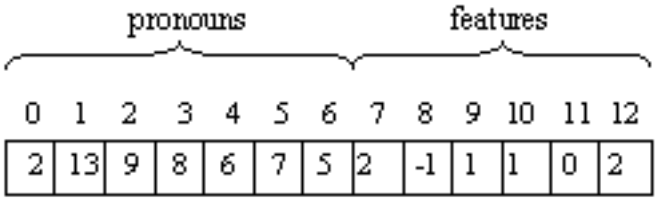

Figure 5.2: Representation of a chromosome

5.4.2.2 Initialisation

Fully random initialisation does not give the algorithm good starting points for searching, so the convergence could be very slow. Therefore, the random initialisation

is controlled in such a way that each pronoun (corresponding to a gene in the chromosome) is associated as antecedent a random noun phrase in the same or previous sentence. The second part of the chromosome, corresponding to the indicators, is filled with randomly generated real values in the [-2, 2] interval. This controlled initialisation ensures that the initial population will contain chromosomes with better fitness values than chromosomes in a random population. At the same time, the large number of noun phrases as compared to the number of pronouns ensures the diversity of the initial population.

5.4.2.3 Fitness function

The fitness function was designed to fulfil two main functions: it evaluates the quality of individual <pronoun, antecedent> assignments and it assesses the correctness of the assignments in rapport to each other within a chromosome. Each individual assignment is evaluated taking into account a number of 6 factors:

- morphological agreement between a pronoun and its associated NP antecedent (MA), possible values 0,1
- position of the pronoun in rapport to the noun phrase (POS), possible values 0,1
- number of intervening NPs between the pronouns and the associated NP (NPC), possible values 1, 2, 3, 4
- syntactic parallelism (SP), possible values 0,1
- pronoun embedded in the noun phrase(EMB), possible values 0,1
- definiteness of the noun phrase (DEF), possible values 0,1

The individual assignment of a chromosome (IC) is calculated as sum of the factors multiplied by their associated weight. With the notations previously introduced, the individual assignment function for a chromosome *c* is:

$$IA(c) = \sum_{i=0}^{P-1} (MA(i) * c(i+P) + POS(i) * c(i+P+1) + NPC(i) * c(i+P+2) + SP(i) * c(i+P+3) + EMB(i) * c(i+P+4) + DEF(i) * c(i+P+5))$$

Apparently, the only part of the chromosome that seems to be evaluated is the one corresponding to the antecedent assignment. However, the evaluation of the <indicator, weight> assignment is built into the fitness function. The overall consistency of a coreferential class is a function depending on two factors: morphological agreement between elements of the class (GMA), number of pronouns in the class (NPPRON). The first factor tries to capture cases where a morphological disagreement occurs at some point in the coreferential chain, for example a plural pronoun referring to a singular NP. This is a frequent phenomenon, however further references to the same NP cannot normally be done using singular pronouns. Therefore the number of morphological disruptions can be an indicator of the (in)consistency of a coreferential class. The second factor captures the intuition that a noun phrase that was referred to by a pronoun is likely to be referred to by a pronoun again. The overall consistency is a very simple measure that by no means describes a coreferential class exhaustively. For better results, a more comprehensive measure needs to be designed. The overall consistency measure can be summarised as follows:

$$OC(class) = \frac{\sum_{i \in class} MDIS(i,i+1)}{|class|} * \frac{NPRON(class)}{|class|}$$

For a chromosome, the overall consistency is calculated as the product of the overall consistency of all independent coreference classes, therefore the function looks like:

$$OC(c) = \prod_{class \in c} OC(class)$$

With the two measures defined above, the fitness function of a chromosome c is:

$$fitness(c) = IA(c) * OC(c)$$

5.4.3 Operators

The three operators that power the mechanism of a genetic algorithm are crossover, mutation and reproduction. The reproduction is performed using the traditional roulette wheel method, that replicates chromosomes according to their fitness value. The crossover operator used was a simple single crossover point operator. The potential parents are selected randomly from the pool of chromosomes generated in the reproduction phase, and they mate with a certain (constant) probability. The crossover point is generated randomly, and two offsprings are created from the two parents by swapping the genes around the crossover point. The offsprings replace the parents in the population. Mutation ensures that genetic material ignored by reproduction and crossover is given a chance of survival in a new population. This process is entirely random and is performed by altering the value of a gene in certain chromosomes. The position of the gene and the chromosomes altered are chosen randomly. The new value of the mutated gene depends on the position of the gene. If the gene corresponds to a pronoun (i.e, it has an index between 0 and P-1) , the new value will be a randomly generated number between 0 and the number of noun phrases in the text. If the gene corresponds to an indicator (has an index between P and P+I, a random number between -2 and 2 is generated, and it is added to the old value, provided that it does not come outside the [-2, 2] range. A series of experiments showed that the best results are obtained when using a 0.95 crossover rate and 0.1 mutation rate.

5.4.3.1 The evolutionary process

First, the text is divided into multi-paragraph segments, following [Hearst1994]. For each segment, the noun phrases and the referential pronouns targeted are extracted. The genetic algorithm is only applied to those segments containing at least 4 referential pronouns. The reason for discarding segments poor in pronouns will become evident from the data representation described below. Segments with few pronouns do not allow the construction of long chromosomes, therefore transforming the evolutionary search into a random search. The search process follows the steps of a simple genetic algorithm. The size of the population was set (after a number of experiments targeted at finding the best value) to 50 chromosomes. For each segment, the initial population is constructed in the manner described above. The chromosomes are evaluated according to the fitness function and they are selected for reproduction. The crossover and mutation operators are applied. At the end of each cycle, the chromosome with the best fitness value is kept in the new population, by replacing the chromosome with the worst fitness value. The process is repeated for a number of 100 generations.

5.4.3.2 Conclusion

The main advantage of the genetic algorithm is that it is an unsupervised method, that learns without the need of annotated data. Although the indicators used and their possible values are determined intuitively, their weights are calculated during the evolutionary process and can differ from segment to segment. The genetic algorithm also has a built-in check of the consistency of a coreference class. This is done without the need for expensive backtracking testing.

The idea that the GA was based upon was potentially interesting and valuable. However, the evaluation that will be presented in Chapter 6 will show that this

assumption was not confirmed experimentally.

5.5 The system

5.5.1 Overview

So far, the underlying machine learning method that the anaphora resolver (ARMAL) is based upon has been described. This section presents a more extended description of the system functionality, including the steps taken for the preparation of the text and the general execution flow.

This section refers in particular to ARMAL, although some of the pre-processing steps are used for the genetic algorithm as well.

Sections 5.5.2 and 5.5.3 discuss the pre-processing steps performed in order to provide the anaphora resolver with the relevant information about the text. The first category of pre-processing deals with the preparation of the text, rendering it in a format more suitable to automatic processing. The second category of pre-processing tools is applied to the text obtained in the previous step, and deal with the identification of morphological, syntactical and semantic information that the anaphora resolver and the learning module require.

5.5.2 Text preparation

In Chapter 4 there have been enumerated some of the peculiarities of technical manuals. Among those, some formatting characteristics are especially important for automatic processing. One aspect that has to be mentioned is that the aim is to perform the identification of these structures with very high precision, while recall is not essential. Overall, it is more important to have fewer special structures identified

as such, than marking normal text as special.

5.5.2.1 Identification of titles, section and subsection headings

Some basic rules are used for the identification of titles, section and subsection headings. Among these, a title is considered a sentence preceded by numbers (e.g. 1.1), a sentence without a verb, a sentence without a final punctuation mark and not included in a list. Titles are marked using a <TITLE> tag.

5.5.2.2 Identification of lists

Lists are identified using typographical markings (such as special characters usually used as bullet points). Each list is tagged with a <LIST> tag and each list item is tagged with an <ITEM> tag. Identification of lists is considered important for the determination of precedence relations between anaphoric elements. The distance between all items in a list and the sentence preceding the list is considered constant (for example, if a list contains 3 items, the distance between the third item and the sentence preceding the list will be considered 1). This is due to the fact that it is likely that the items in the list refer to entities that appeared before the list, and not within the list.

5.5.2.3 Removal of formulae, equations, commands, examples, tables

The information contained in tables, commands or examples is not considered essential for the interpretation of texts. Moreover, although there may be unfrequent cases where a pronoun refers to something contained in such a structure, the disadvantages of keeping these structures in the texts are higher. For example, they disrupt the sentence and paragraph splitting and introduce unnecessary text. The examples found in technical manuals also contain mostly unknown words (such as commands or names of applications) which may not be correctly parsed, if left in the text.

Tables and examples are identified using keywords found in the text and typographical symbols. The formulae and equations are identified by the presence of frequent mathematical symbols. These are only marked for removal when they appear outside a sentence (on a separate line, for example). All these are marked with a <NONP> tag, indicating the fact that they will have to be ignored by the parser and the anaphora resolver.

5.5.2.4 Paragraph and sentence splitter

The identification of paragraphs is a fairly straightforward task, which is only based on typographical marking. Special care is taken for avoiding the identification of examples and list items as paragraphs, although they appear to be so most of the times. This is the reason why the paragraph splitter is only applied after the identification of the special structures described above.

The sentence splitter is a basic rule based application, that takes into account the special features of the technical manuals. The most important special feature is that the sentence splitter has to provide a tighter control of the stop words. It is quite common to encounter full stop or exclamation marks inside a sentence (in email addresses, URLs, some product names, commands, and most commonly, in section numbering). Characters that are not normally considered stop words when processing free text can be stop words in technical documents (for example, colon when followed by a special structure: list, example, table). The second decision to be taken is the treatment of list items; although apparently sentences, they must not be identified as such.

5.5.3 Language specific pre-processing

5.5.3.1 Shallow parser

The morpho-syntactic information necessary for the anaphora resolver is provided by the FDG parser [Tapanainen and Järvinen1997] for the English texts and the Xelda parser from Xerox for the French texts.

The FDG parser has to be fed the text with the paragraphs marked by an empty line. The output is a list of words grouped in paragraphs and sentences. For each word, the shallow parser indicates its lemma, morphological role, type of syntactic dependency and the word that it is dependent of. The parser provides all the possible dependencies for a word; however, in ARMAL only the variant with the maximum reliability confidence is considered (the one that appears first in the list of choices). FDG does not provide solutions for incomplete parses, in this case the dependency type and word are simply omitted.

From this description it should be clear that FDG does not provide NP compounds - for identifying prepositional phrases and noun phrases the syntactic dependencies are used (as it will be described in the next section).

```
0
1       This            this            subj:>2         @SUBJ PRON SG
2       is              be              main:>0         @+FMAINV V
3       an              an              det:>4          @DN> DET SG
4       example         example         comp:>2         @PCOMPL-S NSG
5       of              of              mod:>4          @<NOM-OF PREP
6       the             the             det:>7          @DN> DET
7       functionality   functionality   pcomp:>5        @<P N SG
8       of              of              mod:>7          @<NOM-OF PREP
9       the             the             det:>10         @DN> DET
10      parser          parser          pcomp:>8        @<P N SG
$.
$<s>
```

Figure 5.3: Example of FDG shallow parser output

The parser used for French is part of the Xelda package of language processing tools produced by Xerox. It is an incremental robust parser that provides a structure of syntactic chunks (NP, PP, clauses, etc.) and a set of dependency relations (Subject, Object, infinitive control, noun arguments, etc.). The output is encoded in a bracketed format similar to the Penn Treebank style [Marcus et al.1994]. An advantage of the Xelda architecture is the availability of APIs for both C++ and Java that allow easier integration of the parser in the overall application. The figures 5.3 and 5.4 present the

```
--> Sentence: 0 <--
0 -> 10
SUBJ(This^This+PROP(0),is^be+VBPRES(1))
BEOBJ(is^be+VBPRES(1),example^example+NOUN_SG(3))
NNPREP(functionality^functionality+NOUN_SG(6),
of^of+PREP(7),parser^parser+NOUN_SG(9))
NNPREP(example^example+NOUN_SG(3),of^of+PREP(4),
functionality^functionality+NOUN_SG(6))
0:  (ROOT)
   1: SC(TAG <0 1>)
      2: NP(TAG <0 0>)
         3: This^This+PROP[0](WORD)
         3: SUBJ(FUNC)
      2:  :v(HEAD)
      2: is^be+VBPRES[1](WORD)
   1: NP(TAG <2 3>)
      2: an^an+DET_SG[2](WORD)
      2: example^example+NOUN_SG[3](WORD)
      2: OBJ(FUNC)
   1: PP(TAG <4 6>)
      2: of^of+PREP[4](WORD)
      2: the^the+DET[5](WORD)
      2: functionality^functionality+NOUN_SG[6](WORD)
   1: PP(TAG <7 9>)
      2: of^of+PREP[7](WORD)
      2: the^the+DET[8](WORD)
      2: parser^parser+NOUN_SG[9](WORD)
   1: .^.+SENT[10](WORD)
```

Figure 5.4: Example of Xelda parser output

output of the FDG and of the Xelda parser for a short text.

5.5.3.2 Noun phrase extractor

In the case of English, the FDG shallow parser does not explicitly identify the noun phrase compounds. However, it does provide the functional dependencies between words, which enabled the construction of a noun phrase extractor on top of the shallow parser. For processing the French texts, the Noun Phrase Extractor from the Xelda package has been used.

5.5.3.3 Named entity recogniser

A basic named entity recogniser has been built specifically for pre-processing tasks in ARMAL. Its role is to categorise entities into one of the following 2 classes: PROPER NAME and ORGANISATION. The PROPER NAME category is further divided into FEMININE PROPER NAME and MASCULINE PROPER NAME. Although, in general, named entity recognisers are designed to classify a larger range of entities, for the purpose of anaphora resolution the three categories mentioned above were considered sufficient. This is because the only type of information ARMAL requires from a named entity recogniser is the ability of a certain entity to be referred to by a masculine, feminine, neuter or plural pronoun[8].

The two components of the named entity recogniser are a gazetteer[9] containing a large number of proper names (over 5000 first names and about 90000 surnames) classified by gender and a grammar consisting of 6 regular expressions. The basic rules of the grammar are given in table 5.4[10].

[8] Named entities classified under the Organisation category will be able to be referred to by plural pronouns.

[9] Available from the US Census Bureau web page: http://www.census.gov/genealogy/names/

[10] In the table, CW is short for ”capitalised word” and refers to words that are not at the beginning of a sentence.

Apart from these rules, independent capitalised words that match entries in the gazetteer are also considered proper names. As it can be seen, the grammar is far from comprehensive, but initial evaluation showed that even these simple rules can improve the results of anaphora resolvers by at least 2%. This is mainly due to the fact that in the absence of a NE recognizer, proper names will not have any gender information attached, therefore masculine and feminine pronouns will almost always be solved incorrectly, while some neutral pronouns could potentially be solved (incorrectly) to masculine or feminine proper names. The information added by the named entity recogniser consists of gender for the human proper names and number (plural) for the names of organisations.

Rule	Recognises
$Mr.\|Mr\|Mister\|Sir + CW \rightarrow PROPER-NAME-MASC$	masculine proper names
$Ms.\|Miss\|Mrs. + CW \rightarrow PROPER-NAME-FEM$	feminine proper names
$(CW) * + co\|inc\|ltd \rightarrow COMPANY$	companies and organisations

Table 5.4: Named Entity Recogniser rules

5.5.3.4 Text pre-processing output

The pre-processing modules described above are applied in sequence. Figure 5.5 describes the execution flow of the pre-processing modules. The overall output of the pre-processing is an XML encoded file with the structure described in Chapter 4.

5.5.4 Language identification module

The first step in any multilingual processing system is the identification of the language of the input text. Automatic language identification avoids having to ask the user to

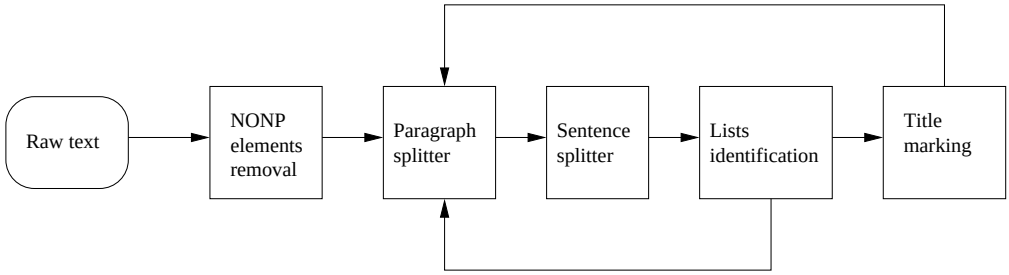

Figure 5.5: Text pre-processing execution flow

select the language before beginning to interrogate the system, therefore minimising the human intervention.

Language identification is a problem that can be solved with almost 100% accuracy. The two main methods that are traditionally used for this task involve either the use of short words (prepositions, articles) or that of short sequences of characters (n-grams). [Greffenstette1995] shows that about 600 short words or 2000 trigrams are sufficient to obtain almost 100% classification accuracy once the text to be identified contains 20 or more words.

The method used by the language identifier for ARMAL is based on collecting trigrams from a set of training data. Table 5.5 shows the most frequent 10 trigrams in English and French. These frequencies are used for identifying the language of the input text.[11]

5.5.5 Anaphora resolution module

The anaphora resolution module is composed of several processing modules and two resolution submodules.

[11] The language identification module can be found online at *http://clg.wlv.ac.uk/LIS/LisApplet.html"*

English	French
the 0.02188023	ent 0.009942032
and 0.005263968	les 0.005452083
ing 0.005169614	que 0.005413789
ent 0.003962874	ait 0.004389429
ion 0.003664914	ant 0.004279334
tio 0.002870353	ion 0.003628339
her 0.002726338	lle 0.003446444
hat 0.002696542	tre 0.003374643
his 0.002627018	men 0.003355496
for 0.002622052	des 0.002986918

Table 5.5: Most frequent trigrams in the training data

- Identification of possible antecedents

  This module deals with the selection of those noun phrases that can be antecedents for a pronoun.

- Filtering module

  The filtering module deals with the application of the filtering rules described in section 5.3.2. In this stage, the filtering rules are applied in the order described above and the pronouns that have been tackled are removed from the list of pronouns submitted to processing.

- Learning module

  The learning module deals with the generation of training instances. This process has been described in section 5.3.6.

- Classification module

  The classification module is a simple piece of code that calls the machine learning classifier.

- Evaluation module The evaluation module calls the evaluation procedure integrated in the Evaluation Workbench (see Chapter 6). This module is responsible for the selection of the evaluation set and the identification of correctly solved anaphors. The evaluation metrics are also built into this module.

5.5.6 Output

ARMAL can be run in both evaluation mode (when it is applied to previously annotated texts) and application mode (when it is applied to previously unseen texts).

Corresponding to the running mode, the output of the anaphora resolver consists of two types of data. The first one displays the results of the evaluation, if the anaphora resolver has been applied to previously annotated data. The second type of output is available for both evaluation mode and application mode and contains the marking of the outcome of the anaphora resolver. This second type of output is available under 3 forms: a graphical output, an output in XML format or a text-only format.

Graphical output

In its graphical form, the output consists of the list of pronouns in the input text, displayed within their context, and having their identified antecedent also displayed in its context. The pronoun and the antecedent are highlighted for easier reading.

XML-encoded output

Alternatively, the program can produce an XML file where the antecedent and the anaphor are marked with COREF tags, the pronoun having a REF attribute pointing to

its antecedent. This encoding scheme is similar with the MUC-based scheme used for manual annotation of corpora (see Chapter 4).

Text-only output

In this format, the input text is replicated in the output, with the only difference that the pronouns are followed by their identified antecedent, in brackets.

Appendix B presents the output of the system in the three formats for short passages of English texts.

5.5.7 Execution flow

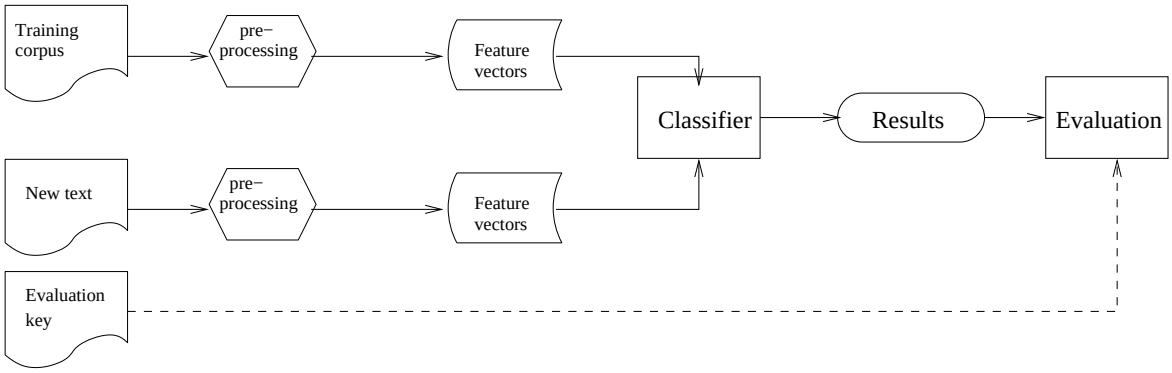

Figure 5.6: Execution flow

Figure 5.6 shows the general execution flow of the system. In training mode, the training corpus is passed to the pre-processing modules and the resulting annotated text is used for generating training instances. In testing mode, an unseen text is similarly pre-processed, feature vectors are extracted from this output and then passed over to the classifier. The results are then passed to the evaluation module, that produces the final output.

5.5.8 Technical implementation details

The anaphora resolver has been implemented in Java, thus being potentially platform independent. However, this relative freedom is much restricted by the pre-processing tools. The shallow parser for English runs under Solaris, while the one for French is a Windows NT application. The work around this problem was the integration of the anaphora resolver in a client-server application, thus allowing for the anaphora resolver to run on one computer, while the pre-processing tools run on their respective machines.

5.6 Conclusions

This chapter has presented a system for bilingual anaphora resolution. The main idea is to use filtering rules for solving easy cases of anaphora and to send the remaining anaphors to a decision tree-based classifier. A cheaper but less successful alternative is also presented in the form of a genetic algorithm that works on raw data.

Chapter 6

Evaluation

6.1 Overview

A major part in the development of an NLP system is addressing the problem of evaluating the system in a number of different areas. Evaluation not only allows for comparative assessment of different individual systems, but also gives important clues about the areas that can be improved in order to obtain higher success rates. Therefore, evaluating a system is important not only in the final stage of development, but at each step during the development; assessing the intermediary results suggests areas to be improved in the system. In fact, the problem of evaluation in NLP has reached such a high level of interest that it has become an area of research in its own right.

Broadly speaking, three main methods of evaluation are used in NLP. *Adequacy evaluation* concerns the functionality and usability of a product for a purpose. *Diagnostic evaluation* is used during different stages of the implementation for deciding whether the system is progressing in the right direction, whether it needs improving and in what areas. *Performance evaluation* represents the measurement of a system's performance, assessed either independently (*qualitative evaluation*) or

in rapport to other similar systems (*comparative evaluation*). The performance of a system can be also assessed in the context of a larger system (*task-oriented evaluation*).

This chapter deals with the evaluation of ARMAL, the anaphora resolution system presented in Chapter 5. Some general issues concerning evaluation in anaphora resolution are addressed, followed by a detailed evaluation of this particular system. The type of evaluation used is *performance evaluation*. Each of the components are assessed both individually and in context and the system is also compared to other state-of-the-art anaphora resolvers.

6.2 Evaluation issues in automatic anaphora resolution

6.2.1 Gold corpus

Anaphora resolution is one of the NLP areas that are suitable for automatic evaluation. A pre-requisite of this type of evaluation is the existence of a manually annotated set of data; this *gold corpus* is used as a standard and the results of an anaphora resolver are compared against manually marked anaphoric links. The main problem with this type of evaluation is the fact that the performance of the system cannot be separated from the reliability of the gold corpus. Errors in the manual annotation will inevitably reflect in the evaluation, making the accuracy gold corpus act as an upper limit for the accuracy of the system. Modern studies in discourse processing [Carletta1996, Carletta et al.1997, Walker and Moore1997] promote the view that linguistic theories have to be based on the judgement of several subjects. In the construction of a gold corpus, this is reflected in the necessity of having more than one annotator marking the text. Although this method reduces the risk of human errors in the gold corpus, it

introduces the additional problem of agreement: if the annotators mark anaphoric links differently, which version should be considered correct, if any?; and how should the reliability of the annotation be computed?

The Kappa Statistic [Siegel and Castellan1988], adopted by Carletta as a measure of agreement for discourse analysis [Carletta1996], is a test suitable for cases where several subjects have to assign items to one of a set of classes. The κ coefficient of agreement among coders depends on the number of coders, the number of items being classified, and the number of choices of classes to be ascribed to items; it also takes into account the possibility for the coders to agree by chance. The κ coefficient of agreement between annotators is defined as:

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

where P(A) is the proportion of times the annotators agree and P(E) is the proportion of times that the annotators are expected to agree by chance. Complete agreement between annotators is equivalent to κ=1, while complete disagreement implies κ=0.

[Krippendorff1980] offers an interpretation of the correlation coefficient, stating that, when computing the correlation of two variables, if the coefficient of agreement is less than 0.8 on one of the variables then strong relationships are likely to be missed even if they do exist. This implies that for such purposes, a value larger than 0.8 for κ is generally taken to indicate good reliability, whereas $0.68 < \kappa < 0.8$ allows for tentative conclusions to be drawn. A value less than 0.68 indicates unreliability.

Whilst it is a very indicative method for evaluating inter-annotator agreement, the kappa-statistic is not very useful in coreference annotation tasks. This is mainly due to the fact that the number of classes a certain anaphor can be assigned to is not known a priori. This could be rectified by presenting the annotators with the classes of coreference (indicated by the first element in chain) already marked. However, this

already means performing part of the annotation and, moreover, it violates the principle of incremental annotation. For this reason, an alternative method for computing inter-annotator agreement is calculating the *precision* and *recall* of the annotation. The precision specifies the percentage of cases on which the annotators agree out of the total number of identified cases. This was the method chosen in the annotation performed for ARMAL.

6.2.2 Qualitative evaluation

6.2.2.1 Evaluation measures

The main problem in qualitative evaluation is defining evaluation metrics suitable for the task. In anaphora resolution evaluation, two measures have mainly been used: precision and recall, although the way they have been defined slightly varies from a system to another. This can be seen in the following two definitions:

Definition 1: (Aone & Benett)

- Precision = $\frac{number\ of\ correctly\ resolved\ anaphors}{number\ of\ anaphors\ attempted\ to\ be\ resolved}$
- Recall = $\frac{number\ of\ correctly\ resolved\ anaphors}{number\ of\ all\ anaphors\ identified\ by\ the\ system}$

Definition 2: (Baldwin)

- Precision = $\frac{number\ of\ correctly\ resolved\ anaphors}{number\ of\ anaphors\ attempted\ to\ be\ resolved}$
- Recall = $\frac{number\ of\ correctly\ resolved\ anaphors}{total\ number\ of\ anaphors}$

Conceptually, *precision* is a measure of the accuracy of the system, while *recall* is a measure of the coverage of the system. While the definition of *precision* is pretty much standard, some differences appear in the definition of *recall*. These differences are mainly due to disagreements regarding the definition of the *total number of anaphors* parameter.

The *F-measure* is a measure of overall accuracy that combines the precision (P) and the recall (R) through their harmonic mean:

- F = $\frac{2*P*R}{P+R}$

Additionally, [Mitkov et al.1999] introduces a simpler metric that combines precision and recall.

- Success rate = $\frac{number\ of\ correctly\ resolved\ anaphors}{number\ of\ all\ anaphors}$

The main disadvantage of the latter metric is the fact that it does not penalise a system for trying to solve non-anaphoric pronouns (such as pleonastic or exophoric pronouns). However, this measure can be successfully used in situations where such errors are irrelevant. *Success rate* is closer in its definition to the *recall*.

6.2.2.2 Key matching

Assuming that a list of (pronoun, antecedent) pairs are returned as a result by an anaphora resolver, these assignments will have to be compared against the correspondent pairs in the gold corpus in order to assess the correctness of the resolution process. There are two main issues related to the assessment of a pronoun resolution as correct. The first one regards the decision of whether a pronoun has been correctly solved or not. Given that coreference is a relation of equivalence, a pronoun can be considered correctly solved if its associated antecedent has been found anywhere in the coreference chain. This means that the proposed antecedent does not necessarily have to coincide with the closest antecedent for the pronoun, but can be found anywhere in the coreferential chain. The same does not hold for cases of indirect anaphora, for example.

The second issue refers to the degree of matching required between a noun phrase identified as antecedent by an automatic anaphora resolver and the correct antecedent as marked in the gold corpus. The guidelines designed within the MUC coreference annotation task specify that a noun phrase antecedent found by a system is considered correctly identified if it shares a minimum text span with the antecedent identified by the human annotator. This minimum span is marked in the gold corpus with a MIN attribute, and is most often the head of the noun phrase, or whichever part of the noun phrase the annotator decides that can be used to better identify the noun phrase.

However, if full identity between the found antecedent and the one marked by the annotator is required, a drop in precision is expected. This is due to pre-processing error at the stage of part-of-speech tagging and noun phrase extraction. Let us consider the sentence below as an example and assume that the system has identified *the I/O setting* as antecedent for the pronoun *it*.

(6.1) Even if <COREF ID="1250" MIN="the I/O setting">the I/O setting of your card</COREF> is not in the list of probed addresses, you can supply <COREF ID="1253" TYPE="IDENT" REF="1250">it</COREF> at boot.

Although the correct antecedent is *the I/O setting of your card*, the proposed antecedent shares the same head with it, and therefore, according to the MUC specifications, the resolution is considered correct. The reason why the NPs do not fully match lies in the inability of the underlying tools (parser, NP chunker) to correctly identify the NPs. Such an evaluation rule ensures the fact that errors for which the anaphora resolver itself is not responsible are not reflected in the results.

6.2.3 Comparative evaluation

It is a known fact that the value of a new method lies not only in its individual performance but also in the improvement that it brings on similar methods. Traditionally, anaphora resolution methods have only benefited from indirect comparison, since access to the code of previous systems and to the same data set used have not been available. Some systems (such as [Hobbs1976],[Hobbs1978], [Walker1989]) were only evaluated manually, a procedure that necessarily restricts the amount of data used for evaluation, allows for the introduction of human errors and is not affected by pre-processing.

In the late years, however, the necessity of providing a more consistent and fair evaluation has been emphasised [Mitkov2002, Byron2001]. Comparing the results reported by different anaphora resolution systems is not enough. Some of the current issues in evaluation have been raised in [Mitkov2002], which states that a fair way of comparing automatic anaphora resolver should take into account the data used for evaluation, the evaluation measures and the running mode of the systems.

First of all, not all anaphora resolution systems employ the same processing mode. The vast majority of approaches rely on some kind of pre-editing of the text which is fed to the anaphora resolution algorithm, while other methods have been only manually simulated. In [Dagan and Itai1990], [Dagan and Itai1991], [Aone and Bennett1995] and [Kennedy and Boguraev1996] pleonastic pronouns are removed manually , whereas in [Mitkov1998][Mitkov1998][1] and [Ferrandez et al. 1998] the outputs of the PoS tagger and the NP extractor/partial parser are post-edited. Ge at al's [1998] and Tetrault's systems [Tetreault1999] make use of annotated corpora and thus do not perform any pre-processing. If the results of the aforementioned

[1]This refers to the original system described in [Mitkov1998]. Mitkov's knowledge poor approach has been later implemented in the MARS system [Mitkov et al.2002].

systems are compared with those of the systems in the MUC competition, which are fully automatic, an important difference in precision (of up to 20%) can be noticed; this difference can be partly explained by difficulties in fully automatic processing. Errors are inevitably introduced at each pre-processing step, and these errors are reflected in the overall success of the system. Similar conclusions have been reached during work on MARS, a fully automatic anaphora resolver based on Mitkov's knowledge poor approach [Mitkov et al.2002]. Therefore, it would not be fair to compare the success rate of an approach that operates on texts which are perfectly or partially analysed by humans, with the success rate of an anaphora resolution system that has to process the text at different levels before activating its anaphora resolution algorithm. Secondly, the evaluation methodology may differ from a system to another, in terms of key matching strategy and evaluation measures. The third problem in evaluation is that of the testing/training data used. Different texts (or even types of texts) feature anaphoric links with different degrees of difficulty of solving. It is expected for a scientific text/technical manual to display less difficult anaphoric phenomena than a narrative text, for example. Moreover, comparing different systems on the same evaluation set is not always fair, since most anaphora resolvers are tailored for tackling a certain text genre, and may not perform well if a change of registers occurs.

In a different line of investigation, [Byron2001] makes an attempt at standardising evaluation results reporting. She calls for a standard, extended description of the coverage of the anaphora resolvers and for a comprehensive analysis of the evaluation set, indicating a set of minimal parameters that have to be specified. Although such reporting of evaluation results is beneficial because it offers a clear picture of the anaphoric phenomena tackled by an individual system, it is by no means complete, since the results reported by two different systems will still have to be correlated.

To summarise, in order to perform fair, reliable and consistent evaluation one has

to take into account for each system involved in the comparison at least the following elements:

- *Coverage*: which anaphoric expressions are tackled by the system
- *The evaluation metrics* used by each system. Differences in evaluation measures mean differences in results. Ideally, these should be the same for all systems involved in the comparison, otherwise an analysis of the way they differ should be in place.
- *Genre*. It is obvious that different text genres differ in anaphoric devices and some anaphora resolvers are specially designed for handling specific text genres. Ideally, systems involved in comparison should be designed to perform on the same or similar text genres.
- *Testing data*. The ideal comparison should be performed on the same testing data; however, since this is not always possible, the corpora used for testing should be at least similar in terms of complexity, coverage and size.
- *Running mode*. Only systems that perform in the same fashion should be compared - for example, systems that benefit from some kind of human intervention are expected to perform better than fully automatic ones, however the results would not reflect the true difference in performance.

6.2.4 Evaluating machine learning methods

Although in the case of machine learning methods the evaluation is always performed on unseen data (separate from the data used for training), this separation is not always genuine. It is common and good practice to perform evaluation during the development of a system and to use the intermediary results for improving the system. Therefore,

classes of errors noticed in the execution of the program can be tackled by modifying features or weights in the machine learning program. However, since these errors are noticed in the testing data, it is likely that with each set of modifications, the machine learning program will learn the features of the training set, which will thus become unsuitable for further evaluation. [Manning and Schütze1999] propose a different method of evaluation to be used for machine learning system. In their approach, the data has to be divided in three parts: the first one to be used in training, the second one in intermediary evaluation and the third one in final evaluation. The downside (or rather the difficulty) of this type of evaluation is the higher cost necessary for providing larger amounts of data for use in training, testing and final evaluation.

6.2.5 Reliability of the evaluation

Considering an anaphora resolution system that has been evaluated both independently and against similar systems, the problem that still remains is: what do the results tell us about the performance of the system?

It is intuitively clear that a 80% precision is simply irrelevant if the testing data consisted of 10 pronouns; it is also clear that a system achieving 90% precision on the same test set is not statistically better than the one achieving 80%, since this translates into only one extra pronoun that was correctly solved. This section deals with two major issues that appear when designing an evaluation testbed: the correct selection of the testing set and the computation of significance measures.

The role of a significance test is to determine if the difference in performance between two algorithms is statistically significant. There are several significance tests for two dependant samples, like McNemar's, Marginal Homogeneity, Sign, and Wilcoxon Tests [Siegel1956, Siegel and Castellan1988]. The choice of a particular test depends on a number of factors, like the size of the testing set. In coreference

and anaphora resolution, the most used test is McNemar's [McNemar1947], which has been employed, among others, by [Tetreault2001] and [Cardie and Wagstaff1999].

6.3 Evaluation methodology for ARMAL

6.3.1 Aims of the evaluation

The evaluation was carried out according to the following aims:

- to assess the overall accuracy of the system
- to assess the suitability of the machine learning module
- to compare ARMAL with state of the art anaphora resolvers
- to identify causes of error in the system
- to assess the ability of the system to cope with different text genres
- to assess the time performance of the system

6.3.2 General issues

Apart from the overall accuracy of the system (expressed in terms of precision, recall, success rate and F-measure), the aim is to provide more specialised types of results as well. The individual resolution of each type of pronoun will be evaluated, in order to determine the difficulty of solving it. The errors due to the pre-processing tools will also be identified. The evaluation methodology will follow the glass-box model, where the performance of each of the components is assessed individually. The evaluation will target the preliminary processing stages, the initial filtering rules and the machine learning program.

Concerning the resolution correctness issue, a pronoun is considered correctly solved if the antecedent found by the system belongs to the coreferential chain of the pronoun and is a full noun phrase (not another pronoun). For practical reasons, no complete matching between the proposed antecedent and the correct antecedent is required. The only requirement is that the two noun phrases should share the same head.

6.3.3 Comparative evaluation

6.3.3.1 Evaluation workbench

Some of the current evaluation issues outlined in 6.2.3 have been addressed in a practical attempt to develop an environment for anaphora resolution evaluation. The *evaluation workbench* [Barbu and Mitkov2001] is a system that incorporates a few anaphora resolution methods which share the same philosophy (i.e., they are rule-based approaches) and allows for comparative evaluation to be performed. Three knowledge-poor approaches that have been extensively cited in the literature have been selected for comparative evaluation: Kennedy and Boguraev's parser-free version of Lappin & Leass' RAP [Kennedy and Boguraev1996], Baldwin's pronoun resolution method [Baldwin1997] and Mitkov's knowledge-poor pronoun resolution approach [Mitkov1998]. All three algorithms share a similar pre-processing methodology: they do not rely on a parser to process the input and use instead POS taggers and NP extractors; none of the methods make use of semantic or real-world knowledge. Additionally, two syntax-based methods have also been incorporated: Hobbs' [Hobbs1978] and Lappin & Leass' [Lappin and Leass1994]. All the aforementioned algorithms have been reimplemented based on their original description. A set of three baselines are also included: in the first one, the most

recent candidate is selected as antecedent; the second one always selects the most recent subject, while the third one selects the antecedents randomly from the list of candidates.

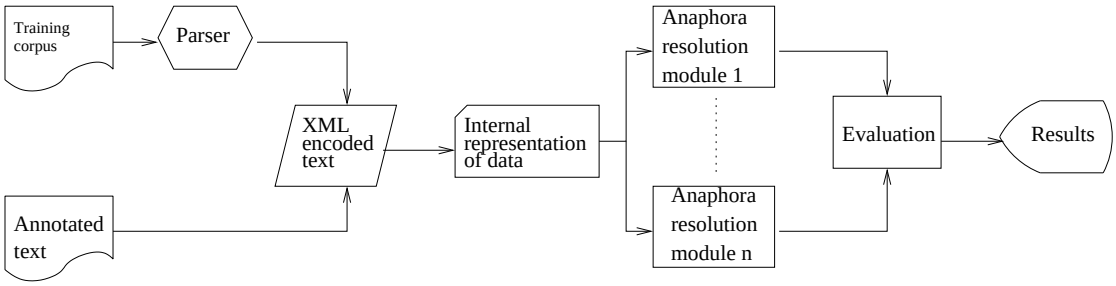

Figure 6.1: Evaluation workbench: execution flow

The main advantage of the workbench is the fact that it provides fair ground for comparative evaluation. All the methods incorporated are fed the same data, use the same pre-processing tools in the same manner and they are fully automatic. Furthermore, the same evaluation measures are used (*precision* and *recall* as in Definition 1, *F-measure* and *success rate*), and whenever a new evaluation measure is introduced, it is applied on all methods in the same way.

Apart from evaluating the performance of the anaphora resolvers, the workbench also provides measures of the complexity of the evaluation set and facilities for data sampling[2].

6.3.4 Qualitative evaluation

6.3.4.1 Methodology

In an influential work in the evaluation of NLP applications, [Sparck-Jones and Galliers1996] summarise their vision of evaluation in two steps:

[2] A detailed description of the implementation and features of the evaluation workbench and the way it is used for evaluation of anaphora resolution systems is available in [Barbu and Mitkov2001]

> 1. Unpack the evaluation by working systematically through the series of relevant questions ... pushing into the necessary detail and decomposing the evaluation subject in corresponding detail for proper answers;
> 2. Envisage the evaluation from the start as a programme of coherentlyrelated component evaluations based on a performance factor grid.

The qualitative evaluation of ARMAL will follow this basic philosophy of starting from the overall application and moving into smaller components.

6.4 Results

6.4.1 Filtering rules

As mentioned previously, the role of the filtering rules is to eliminate from further processing those anaphors that do not require large amounts of computation for their resolution. The more pronouns are filtered out by the rules, the quicker the resolution process for the entire text, however the accuracy of the resolution is likely to decrease. Hence, the main problem is to keep a balance between the number of pronouns filtered out and the accuracy of the system. The performance of the system was considered far more important than the speed of computation; therefore the filtering rules need to achieve very high precision, while their discrimination power, expressed by the recall metric, plays only a secondary role.

The evaluation of the filtering rules was performed on the entire corpus described in Chapter 4, as this corpus has not been used in the design of the rules.

The results, presented in table 6.1 show that the filtering rules only apply for about 5% of the total number of pronouns, but their accuracy is very high.

	#pronouns	applied	precision
English	545	29	96.55%
French	432	23	100%

Table 6.1: Accuracy of the filtering rules

6.4.2 Learning module

In assessing the performance of the learning module, the values of the following measures have been computed:

- The **predictive power** of the learning module represents the classification performance on the training instances. This measure gives an upper bound to the performance on unseen instances.

- The second measure is the **classification accuracy on test instances**, calculated in a 10-fold cross-validation experiment.

- The **average typicality of misclassified test instances** measures the proportion of misclassified instances that are ”atypical”. The typicality of a test instance *i* is computed as ”*i*’s average similarity to instances with the same classification *c*, divided by its average similarity to instances with another classification than *c*” [Zhang1992].

	Predictive power	Classification accuracy	Avg typicality
French	80.42	59.78	1.2
English	84.36	62.96	1.4

Table 6.2: Performance of the learning module

The first observation that can be made over these results concerns the predictive power of the classifier. This, as it can be seen, is relatively low, forcing a low upper bound for the performance of the classifier. The main reason for this is the small size of the training test. An experiment performed using 70% and 80% of the data for training shows that the predictive power increases with the amount of training data available.

The second observation concerns the value of the average typicality. The higher the average typicality, the less atypical instances are misclassified (relative to typical test instances). As it can be noticed, for the English texts the system misclassifies a smaller percent of the atypical instances than for the French texts.

The performance of the learning module has been evaluated independently of the filtering rules on unseen data. This has been done by selecting 10% of the corpus for evaluation, while the remaining 90% was used in training. The testing set was selected virtually randomly, by extracting a contiguous portion of text from the corpus, containing the required number of pronouns. It was considered necessary for the testing data to be used on contiguous text in order to take advantage of the *anaphoric richness* feature used by the learning algorithm. This feature takes into account how many times an NP has been referred to by a pronoun in the same paragraph, therefore this information has to be available to the learning module. For the same reason, pieces of text included in the testing set had to contain whole paragraphs. The contiguity rule was only allowed to be violated when the end of a file has been reached (i.e, it is possible to select the last paragraph of a file and to continue selecting data from a different file).

With these specifications of the test set, a number of 64 pronouns from the English corpus and 48 pronouns from the French corpus have been selected for testing. The filtering rules eliminated 2 out of 64 English pronouns and 1 out of 48 French pronouns. The remaining 62 and respectively 47 pronouns were passed on to the

learning algorithm for solving.

For the English texts, the learning module solved correctly 28 pronouns, while 13 were non-anaphoric. This translates into a precision and recall of 45.1% and a success rate of 57.14%.

In the case of the French texts, the learning module solved correctly 20 pronouns, while 7 were non-anaphoric. The precision and recall of the French resolver were 42.5%, and the success rate was 50%.

6.4.3 Overall evaluation

The overall evaluation concerns the assessment of the performance of the system as a whole, combining the filtering rules and the automatic classifier.

6.4.3.1 General results

Table 6.3 presents the evaluation results obtained by ARMAL on the English and French texts. These results were obtained using the same testing set as in the previous section. The machine learning module was trained on the remaining pronouns from the two corpora.

	#pron	#anaphoric	precision	recall	success rate
English	64	51	46.87%	46.87%	58.82%
French	48	41	43.75%	43.75%	51.21%

Table 6.3: Results for ARMAL on the English and French testing texts

6.4.3.2 Pronoun-driven evaluation

In this section, the evaluation results are broken down into categories, according to the type of pronouns and the morphological features of the pronouns. The main problem of the results that are going to be presented in the following is the fact that the test set was very small (64 pronouns for English and 48 for French), which makes the numbers for individual classes of pronouns even smaller. Therefore, the results may be deceptive and unlikely to be statistically significant.

Table 6.4 displays the evaluation results obtained on the testing data in the English and French corpus, when taking into account personal pronouns and possessive pronouns only. In the case of possessive pronouns, all the three measures used for evaluation, i.e., precision, recall and success rate will have the same value, since all possessive pronouns in the corpus were anaphoric. In the case of personal pronouns, precision and recall will be different from success rate since there is an important number of non-anaphoric personal pronouns. There were no reflexive pronouns in the English test set, and just one reflexive pronoun in the French test set (it was not included in the results table).

	personal			possessive	
	#	prec=rec	succ rate	#	prec=rec=succ rate
English	48	37.5%	51.42%	16	62.5%
French	31	38.7%	50%	16	75%

Table 6.4: Resolution of personal and possessive pronouns in the testing corpus

Table 6.5 presents the evaluation results for the resolution of pronouns grouped by number. Since plural pronouns have a more hectic behaviour than singular pronouns, it was expected that such an evaluation will provide interesting means of comparison.

File		#pronouns	precision	recall	success rate
English	sg	55	40.0%	40.0%	52.38%
	pl	9	66.66%	66.66%	66.66%
French	sg	40	35%	35%	42.42%
	pl	8	87.5%	87.5%	87.5%

Table 6.5: Resolution of pronouns according to their number

Table 6.6 presents the evaluation results for the resolution of English neuter pronouns (*it* and *they*). All French pronouns are marked for masculine or feminine gender, so a similar evaluation for French is not possible.

#pronouns	precision	recall	success rate
59	44.06%	44.06%	56.52%

Table 6.6: Resolution of neuter English pronouns

6.4.4 Discussion

Three main conclusions can be drawn from this qualitative evaluation. Firstly, the filtering rules are only applied in a few cases (about 4% of the total number of pronouns), however they work for a very high percent of the selected pronouns. Secondly, the performance of the learning module is quite low, but, as this is largely due to the reduced dataset, it has the potential to improve.

The performance of the system as a whole is not impressive, but the results are not very low, considering the fact that the system performs consistently for both English and French.

The analysis of the evaluation results for different types of pronouns shows that

for both English and French there seem to be similar proportions of wrong resolutions in all pronoun types. The size of the testing data, however, is too small to allow for significance testing.

6.4.5 Cross-genre evaluation

The cross-genre evaluation was performed on the extract from the BAF corpus described in 4.4.4.3. It contains an extract from a narrative text (Jules Verne. *From the Earth to the Moon*) and a scientific text (Geoffroy, Catherine. *Les technologies de communication de l'information et les aînés*. Rapport technique du CITI).

The scientific text in this corpus is similar to technical manuals with respect to its structure, but differs in the frequency and distribution of pronouns. The narrative text is the furthest away from the technical manuals in both structure and distribution of pronouns.

The results of the evaluation are presented in table 6.7. Apart from the scientific and narrative texts, the table also displays, for easier comparison, the results obtained when evaluating ARMAL on technical manuals.

For training purposes, the whole corpus of technical manuals has been used.

Surprisingly, the scientific text was proven to be the most difficult to tackle for the English system, while the narrative text scored closer to the technical texts. At a closer look, this seems to happen because of the large number of references to people that appear in the narrative texts. In these cases, the resolution of pronouns is assisted by the gender and animacy agreement, in most cases rendering the pronouns unambiguous. This conclusion is supported by the fact that for the French text, the anaphora resolver scores lower on the narrative text, since in this case the gender agreement has a lower discrimination power.

File		#pronouns	#anaphoric	precision	recall	success rate
English	TECH	64	51	46.87%	46.87%	58.82%
	SCI	96	73	41.66%	41.66%	54.79%
	NARR	151	130	43.04%	43.04%	50%
French	TECH	48	41	43.79%	43.79%	51.21%
	SCI	108	79	40.7%	40.7%	55.69%
	NARR	166	145	38.55%	38.55%	44.13%

Table 6.7: Cross-genre evaluation results

6.4.6 Comparative evaluation

As previously mentioned, the comparative evaluation was performed on English texts only, due to the unavailability of similar French anaphora resolvers. This section discusses the differences in resolution between the anaphora resolvers included in the evaluation workbench and the machine learning methods presented in Chapter 5.

The main problem faced while attempting comparative evaluation was the lack of sufficient testing data. Since 90% of the corpus is necessary (or even, most likely, insufficient) for training of the machine learning system, only 10% of the corpus is available for evaluation. This translates into a number of 63 English pronouns and 52 French pronouns. Although appropriate for qualitative evaluation, a test set of this size does not allow for significance tests to be performed reliably.

For the sole purpose of comparative evaluation, one of the files in the corpus was used (CDR), consisting of 97 pronouns, out of which 85 were anaphoric. Although this allows for better comparison, the increase in the evaluation set is done at the expense of the data available for training. This explains the differences between the overall results of ARMAL presented in the previous section and those that will follow.

Method	precision	recall	success rate
ARMAL	50.51%	50.51%	57.6%
K&B	56.70%	56.70%	64.71%
Mitkov	57.73%	57.73%	65.88%
CogNIAC	32.99%	32.99%	37.65%
Hobbs	58.76%	58.76%	67.05%
RAP	60.82%	60.82%	69.41%

Table 6.8: Comparative evaluation results

6.4.7 Evaluation of the genetic algorithm

For reasons that will become apparent in the following, I have left aside so far the evaluation of the genetic algorithm presented in Chapter 5. Although the idea of using a genetic algorithm for anaphora resolution seemed a good alternative to supervised methods, the results did not confirm the initial assumption.

The qualitative evaluation of the model envisaged its performance (measured as precision and recall) according to a number of factors: number of pronouns in the segment, different crossover and mutation rates, different population size, variable number of iterations. Apart from the model described above, the aim of the evaluation was to see how the performance changes when the genetic algorithm is applied not on multi-paragraph but on single paragraph segments. Reducing the size of the searching space can theoretically have a beneficial effect on the results, however it is accompanied by a reduction in the length of the chromosomes. In order to have a realistic view of the system's performance, it was tested against state-of-the-art rule-based anaphora resolution methods. The comparative evaluation strategy involved using the evaluation workbench and also the two supervised learning methods

presented in Chapter 5.

As mentioned before, the genetic algorithm is only applied on segments with a sufficient number of pronouns. For the comparative evaluation, a robust implementation has been used[3], where the pronouns in the omitted segments are resolved to the most recent compatible noun phrase.

6.4.7.1 Testing corpus

The genetic algorithm was tested on the same corpus described in Chapter 4. Table 6.9 displays for each file, in this order, the number of words, segments, paragraphs, pronouns, and anaphoric pronouns. The last two columns represent the data processed by the genetic algorithm: the number of segments with more than 4 pronouns and the number of pronouns contained in these segments.

File	#words	#seg	#par	#pron	#anaph	#seg GA	#pron GA
ACC	9617	77	180	182	160	20	54
BEO	6392	42	192	92	70	11	57
WIN	2773	11	79	51	47	4	37
CDR	9490	99	236	97	85	9	36
ETH	7296	22	88	104	56	10	51
Total	35568	251	775	526	418	39	235

Table 6.9: Evaluation corpus

As a first observation on the distribution of pronouns, one can see that only a small part of the segments contain more than four pronouns, which is the minimum limit for the genetic algorithm. However, out of the segments with less than four

[3]The term *robust* in this context refers to anaphora resolvers that try to solve all pronouns.

pronouns, almost a half contained no pronouns, therefore can be entirely omitted without consequences on the resolution rate. The pronouns considered by the genetic algorithm were distributed evenly in segments with 5 to 7 pronouns, while segments containing 4 pronouns represented the majority of the cases. A second observation is that technical manuals present a highly structured architecture, being split in many small length sections and subsections. As the section marking does not necessarily correspond to the logical structuring of discourse, identification of segments is not always accurate with respect to the meaning of the text.

6.4.7.2 Experiments

Four sets of experiments have been performed. The first one envisaged the values of the genetic algorithm parameters: population size, number of iterations, crossover rate and mutation rate. After a number of executions with different parameters, the best results were obtained when using a population with 50 chromosomes, evolved during 100 iterations, with a crossover rate of 0.95 and a mutation rate of 0.1. This is the set of parameters that has been further used in evaluation. For evaluation purposes, the genetic algorithm was run 10 times and the results displayed in the tables represent the most consistent outcome of the 10 runs. The second experiment was designed to assess the performance of the genetic algorithm independently on segments containing more than 7, 7, 6, 5 or 4 pronouns. The results are displayed in table 2 . The overall precision of the genetic algorithm on all segments considered was 62.7%. Figure 6.2 shows how the performance of the genetic algorithm is affected by the number of pronouns per segment. The third experiment envisaged the comparative performance of the genetic algorithm and of the three rule-based anaphora resolvers on segments containing more than 4 pronouns (results in figure 6.2) and on the whole texts (results in Table 3). The last experiment tried to evaluate the performance of the genetic algorithm when

paragraphs and not segments are used as searching space. Unfortunately, this last experiment did not produce any significant results, since the number of paragraphs containing more than 4 pronouns was very small.

6.4.7.3 Comparative evaluation results

Table 6.10 displays the results obtained when running the robust version of the genetic algorithm and the three methods included in the evaluation workbench on the same testing set. The results are not fully comparable, since the precision of the genetic algorithm is much diminished by the poor resolution of pronouns in the omitted segments. Figure 6.2 displays a more accurate comparison of the genetic algorithm and the other methods. The evaluation is performed on segments rich in pronouns alone.

File	>7		7		6		5		4	
	#pron	Prec	#pron	Prec	#pron	Prec	#pron	Prec	#pron	Prec
BEO	17	94.12%	7	81.0%	12	66.7%	5	46.7%	16	43.8%
ACC	16	93.8%	7	57.1%	6	66.7%	5	60.0%	20	60.0%
CDR	23	78.26%	7	42.9%	6	66.7%	0	N/A	0	N/A
WIN	0	N/A	0	N/A	0	N/A	5	60%	32	53.1%
ETH	21	85.7%	7	71.4%	6	50%	5	40%	12	41.7%
Total	77	87.0%	42	69.0%	30	63.3%	30	50%	100	45%

Table 6.10: Comparative evaluation of the GA

As it can be seen, the genetic algorithm outperforms the other anaphora resolvers on segments rich in pronouns, but its performance drops on segments with less than 5 pronouns.

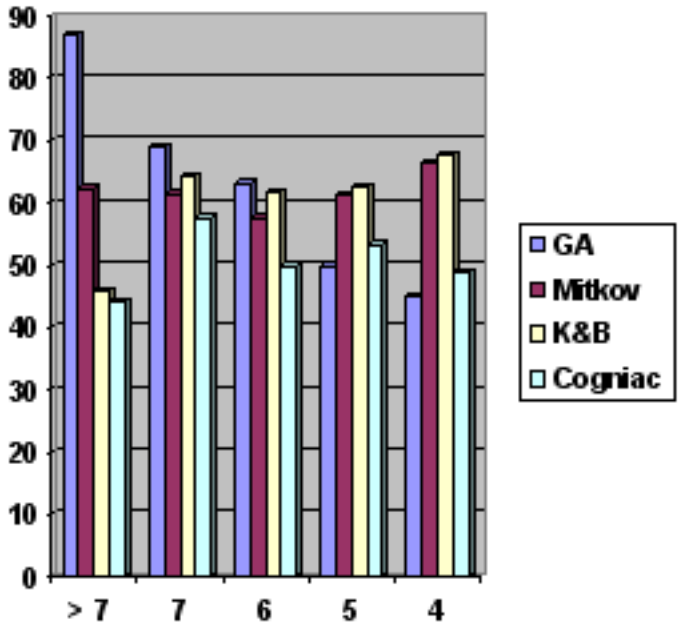

Figure 6.2: Comparative evaluation results in rapport with the number of pronouns per segment

6.4.7.4 Discussion

The results obtained by the genetic algorithm vary significantly with the number of pronouns in the segment. This characteristic is a direct consequence of the data encoding. By analysing the errors produced by the genetic algorithm, it was noticed that only two cases represented inter-segment anaphora, which was not tackled at all. The main advantage of the genetic algorithm is that it is an unsupervised method, that learns without the need of annotated data. Although the indicators used and their possible values are determined intuitively, their weights are calculated during the evolutionary process and can differ from segment to segment. The genetic algorithm also has a built-in check of the consistency of a coreference class. This is done without the need for expensive backtracking testing. The main disadvantage of the genetic

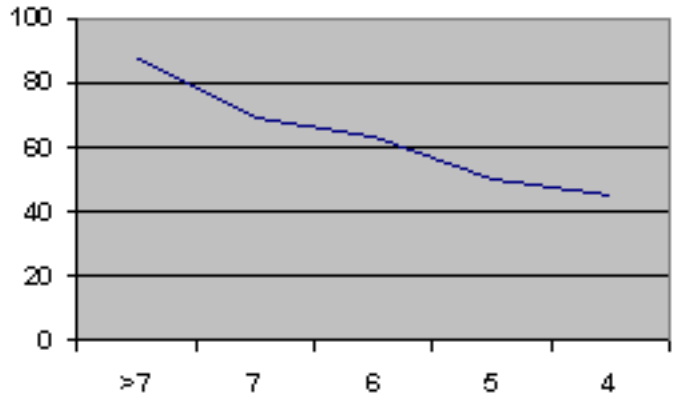

Figure 6.3: Evolution of precision according to the number of pronouns per segment

algorithm is that it does not perform well on segments poor in referential pronouns. The search in such cases is close to a random walk among possible antecedents. Experiments showed that in real technical texts, most segments have an average of 4 pronouns, which is below the minimum necessary for the genetic algorithm to perform reasonably well. As it only works within text segments, the genetic algorithm does not solve any anaphoric links that span more than one segment. However, this is only a minimal drawback, as experience shows that inter-segment anaphora is a rare phenomenon, that can be ignored without a significant drop in performance. The running time is also sensibly longer than for the other knowledge poor methods, but this can be improved by a more efficient implementation of the selection technique and by a reduction in the number of iterations.

The main conclusion that can be drawn from the results obtained using the genetic algorithm is that, in its present form, the GA method does not challenge traditional anaphora resolvers, but it does present some interesting features that could be further exploited in a better system. Special consideration has to be given to the data representation and the fitness function. A better encoding may allow for the pronoun-

poor segments to be processed as well.

6.5 Mutual improvement

6.5.1 Overview

During the stage of corpus analysis for the acquisition of features to be used in ARMAL, as well as during the evaluation of the system, it became apparent that language specific features have a role to play in the performance of an anaphora resolver. The most obvious such feature is the grammatical gender distinction in French, a feature that, if present in English, could bring an important contribution to anaphora resolution. This and other similar considerations 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[4]. In addition to exploiting 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. Both the English and the French modules are based on Mitkov's [1998] knowledge-poor approach. The choice of this particular method was based on its proven multilingual potential and its ease of integration with the enhancement mechanism. For several reasons, ARMAL was not considered appropriate for replicating the experiment. The most important reason was the fact

[4]The design and development of the system is collaborative work with Ruslan Mitkov, description and results have previously been published in [Mitkov and Barbu2000]

that the machine learning method does not allow interactions (i.e, the weights of the features are computed automatically, and it is a violation of the learning algorithm to allow these weights to be modified by an external application). Another reason is the fact that the machine learning algorithm is not based on preferences, i.e. it does not return a list of candidates ordered by a certain measure of confidence.

6.5.2 Brief outline of the bilingual corpora

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. The composition of the individual files in number of words, pronouns, type of pronouns, has been presented in Chapter 4. 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 6.11 presents a summary of the different ways in which English pronouns were translated into French and the cases giving rise to new French pronouns.

6.5.3 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

File	Direct translations	English pron to French NP	English NP to French pron	New French pron	English pron 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

Table 6.11: Pronoun translation correspondences

more reliably, whereas in others the English module is likely to propose the correct antecedent with higher probability.

6.5.3.1 Cases where the French version can help

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, and these cases have been discussed in Chapter 2. Since gender agreement works for most cases in French, whenever the antecedent in French is resolved directly after gender agreement, its equivalent in English is adopted as the antecedent. 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 (2 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) , then 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 4, and the difference between the two best English candidates is only 1, then the proposed antecedent of the French module will be preferred.

6.5.3.2 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 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 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. The algorithm implemented for this project incorporates the following syntax filters as used in Kennedy&Boguraev [Kennedy and Boguraev1996]:

- A pronoun cannot refer with a co-argument.
- A pronoun cannot co-refer with a non-pronominal constituent which it both commands and precedes.
- A pronoun cannot co-refer with a constituent which contains it.

These constraints are a modification of the syntax constraints reported in Lappin and Leass [Lappin and Leass1994] 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 [Mitkov2001] which is indicative of the confidence of the proposed antecedent. 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) or 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 4 taken again as a threshold (see above and also the following section).

6.5.4 Selection strategy

The selection strategy of the implemented mutual enhancement algorithm is based on favouring cases where one of the systems is expected to perform better, as described in section 3, and addresses pronouns that cannot be resolved directly in either of the languages. This strategy benefits from the outputs of Mitkov's algorithm 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 2 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 section 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 be-tween 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), 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,.., N; N is the number of all candidates) and its equivalent French candidate $French\_C_i$ (i = 1,.., N), the weighted score 0.6 x $English\_C_k$ + 0.4 x $French\_C_i$ 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 higher than 0.8 is employed and if the highest decision power of the indicators activated in the other language is lower than 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.

6.5.5 Evaluation

The evaluation was based on parallel texts aligned for NPs featuring 25499 words (281 of which were pronouns) in English and 28037 words (390 pronouns) in French. 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*. 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.

File	#pron	before enhancement		after enhancement		Improvement
		#correctly solved pron	success rate	#correctly solved pron	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 6.12: Resolution for English before and after enhancement

File	#pron	before enhancement		after enhancement		Improvement
		#correctly solved pron	success rate	#correctly solved pron	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

Table 6.13: Pronoun resolution for French before and after enhancement

Tables 6.12 and 6.13 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 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.

6.5.6 Discussion

This study shows that bilingual corpora for English and French can improve anaphora resolution in both languages. The implemented mutual enhancement algorithm delivers an average improvement of 4.27% for English and 3.59% for French. Whereas this project has been restricted to French and English and the testing set was limited by the unavailability of data, the results confirm more generally that resolving anaphora in a language that features grammatical gender as French can help anaphora resolution in a morphologically poor language as English.

The question that the experiment raised is whether such an approach is appropriate for real applications. First of all, whether the slight increase in performance can justify a higher increase in computation. Secondly, whether high level pre-processing

tools will be available in the foreseeable future, especially good quality word aligners. Thirdly, if bilingual corpora will be easily available for such applications.

Concerning the first issue, we believe that, although speed of processing is important, accuracy of resolution is essential, and a decrease in speed is a price worth paying for an improvement in accuracy. Furthermore, it is not intended to use the enhancement mechanism artificially, for improving the performance of an anaphora resolver in a single language; the intended use is in a bi- or multilingual environment, where running anaphora resolutions in all languages is a requirement. In these cases, the cost of the enhancement mechanism alone is expected to be negligible.

Bilingual corpora, especially in the technical domain, are quite easily accessible, and will be even more so in the following years.

The problem that remains is the availability of powerful word aligners, which for the moment is an open issue. In these circumstances, multilingual anaphora resolution by mutual enhancement is rather a potential solution that depends on external factors.

6.6 Conclusions

This chapter has addressed the problem of evaluation of anaphora resolution systems. The first part of the chapter was dedicated to general issues in automatic evaluation, followed by a discussion of the evaluation strategy for ARMAL. The qualitative evaluation showed a performance around 55% for French and slightly higher for English, while the comparative evaluation situated ARMAL at an average level in rapport to state-of-the-art anaphora resolvers.

In a different line of investigation, the genetic algorithm for anaphora resolution described in Chapter 5 is evaluated as well. Its evaluation shows that although a promising idea, the GA does not prove a valuable alternative to the supervised learning

method used in ARMAL.

The last part of the chapter describes an approach to mutual enhancement of anaphora resolvers for English and French by way of parallel corpora.

Chapter 7

Difficult issues in pronoun resolution

7.1 Overview

In Chapter 3, some of the most cited and successful pronoun resolution system have been outlined. Subsequently, an original approach based on machine learning techniques has been presented in Chapter 5. A glance at the literature on automatic anaphora resolution and at the comparative evaluation results in Chapter 6 shows that the resolution rate of pronouns lies at around 80%, [Hobbs1978, Ge et al.1998, Mitkov1998] but the results decrease significantly when fully automatic methods are employed. Despite many years of research in the field, and several original approaches tried, pronominal anaphora resolution is not significantly more successful now than it was twenty years ago.

The relatively low success rate of anaphora resolution methods can be intuitively attributed to a number of problems that the task arises: the impossibility of using world knowledge (due to the prohibitive expense needed for encoding this kind of knowledge), the lack of semantic information, errors in the pre-processing stage and, not in the last, the inherent ambiguity of the natural language.

Although this problem has been acknowledged by different authors [Mitkov2000], to our knowledge there has been no study of the way different factors induce a decrease in the performance of anaphora resolvers. Such a study could enable researchers working in the field to identify areas that can be improved and to deal with phenomena less studied and that could increase the performance of anaphora resolution. As it is not possible as yet able to make use of extensive semantic and world knowledge information, it is nonetheless possible to make better use of easily acquired knowledge sources, as morphological or syntactical information.

This chapter discusses some of the difficulties in multilingual anaphora resolution. Section 1 presents a general overview of the most common and toughest problems in pronoun resolution. The following sections address in depth more specific problems.

First, the main types of general errors are identified and a corpus based analysis of their frequency is performed. Secondly, the problem of morphological disagreement is studied in depth and a large scale corpus analysis is presented, showing the distribution and frequency of the phenomena across different text genres. Thirdly, errors specific to each language investigated are identified.

7.2 Tough problems in anaphora resolution

7.3 Error analysis

This section describes a series of experiments targeting the analysis of errors reported by four pronominal anaphora resolvers, including ARMAL. The main idea was to identify classes of errors that appear frequently and that are common to more than one system. Section 7.3.3 presents a first direction of investigation that consisted in identifying classes of pre-processing errors and in assessing their influence in the

overall result. Section 7.3.4 describes a different thread of investigation, that followed the differences in resolution of certain types of anaphoric expressions. Statistics were generated with regard to the resolution potential of each of the methods on each of the classes of anaphors. This analysis aimed at discovering if there are statistically significant differences in the resolution of certain types of anaphors. The investigation led to a pilot implementation of a probabilistic system (presented in Section 7.3.5) that incorporates the best features of each of the methods analysed in order to take advantage of their ability to deal with certain classes of anaphors.

7.3.1 Methodology

The investigation is based on the analysis of the results provided by four anaphora resolvers: three rule-based approaches (Mitkov's knowledge-poor method [Mitkov1998], Kennedy&Boguraev's parser-free method [Kennedy and Boguraev1996], the robust version of Baldwin's CogNIAC [Baldwin1997] and ARMAL. The choice of these particular approaches was based on the fact that they all make use of (more or less) the same set of indicators, although combined in different ways: CogNIAC applies a set of rules in the order of their confidence, Kennedy&Boguraev compute scores associated to coreference classes and classify a pronoun in the highest ranked class, Mitkov associates scores to possible antecedents and link a pronoun to the noun phrase with the highest score, while ARMAL builds decision trees using a set of features computed over a training corpus. Moreover, all methods have been implemented to run in a fully automatic mode, using the same pre-processing tools and the same testing data[1].

These common features make it possible to assess the influence of both the errors in the pre-processing stage for computing the indicators, and the errors due to

[1]The anaphora resolvers used were implemented within the evaluation workbench (see chapter 6)

			ACC	WIN	BEO	CDR	**Total**
#words			9617	2773	6392	9490	28272
#pronouns			182	51	92	97	422
#anaphoric pronouns	personal	it	103	30	31	51	215
		he, she	5	0	4	0	9
		they	32	6	17	22	77
	possessives		18	11	18	10	60
	reflexives		3	0	0	2	5
	Total anaphoric		161	47	70	85	366

Table 7.1: Distribution of pronouns in the training corpus

malfunctioning of the methods themselves.

The performance of the systems has been computed in terms of four evaluation measures: precision and recall (as defined in [Aone and Bennett1995]), F-measure, that combines the two, and success rate (as defined in Chapter 6). For the sake of the clarity, in some of the following reports the performance of the resolution is only expressed in terms of success rate.

7.3.2 Corpus

The corpus used for these experiments is a subset of the English corpus described in Chapter 4. As these experiments contain an important part of human intervention, the analysis has been necessarily restricted to a smaller corpus.

7.3.3 Analysis of the influence of the pre-processing tools

Several works and experiments have demonstrated that fully automatic anaphora resolution is more difficult than previous work has suggested [Mitkov2002, Barbu and Mitkov2001]. Errors are inevitably introduced at each pre-processing step, and these errors are reflected in the overall success of the system. However, it has not yet been investigated what percentage of the errors in the resolution of pronouns are due to the pre-processing, and, more precisely, which aspects of pre-processing are responsible for failures in pronoun resolutions. Although this work tries to give an answer to some of these problems, it still has to be mentioned that the results presented here cannot be indiscriminately applied to other anaphora resolution systems. This is due to the fact that the results were obtained and on a certain type of texts (technical manuals) using a certain set of pre-processing tools. It has to be made clear that these experiments do not envisage the performance of anaphora resolution algorithms, but of fully automatic systems.

The influence of pre-processing increases exponentially with the number and complexity of the pre-processing tools used. All the aforementioned systems require a limited amount of easily computed morphological, syntactical and textual knowledge, acquired by means of a shallow parser (the FDG shallow parser [Tapanainen and Järvinen1997] in the current implementation) and a noun phrase extractor (built on top of the output of the shallow parser).

The methodology used for assessing the influence of the pre-processing tools consisted in manually post-editing the results returned by the shallow parser and re-running the systems on the perfect input. Due to the expenses involved in this operation, it was not possible to fully post-edit the results, thus we have only decided to correct those errors that were intuitively considered more likely to influence the

		ACC	WIN	BEO	CDR	Average
Success rate	Mitkov	52.7%	55.3%	48.5%	71.7%	56.2%
	CogNIAC	45.9%	44.6%	42.8%	67.0%	49.7%
	K&B	55.0%	63.8%	55.7%	74.1%	61.6%
	ARMAL	55.9%	63.8%	52.8%	72.9%	59.8%
Precision	Mitkov	42.8%	50.9%	36.9%	62.8%	48.8%
	CogNIAC	37.1%	41.1%	32.6%	58.7%	42.6%
	K&B	48.3%	58.8%	42.3%	64.9%	52.8%
	ML	49.4%	58.8%	40.21%	63.9%	51.2%

Table 7.2: Initial evaluation results

performance of anaphora resolvers. Accordingly, there have been corrected the delimitation of sentences, the prepositional phrase attachment, the identification and attachment of articles, the composition of noun phrases, the attachment of noun phrases to verbs and the grammatical function of noun phrases; we have ignored the attachment of adverbials, the composition of verb-phrases, the mal-recognition of adverbials and of other non-anaphoric entities, the grammatical function and morphological features of non-referential entities, the features of anaphoric or non-anaphoric pronouns which are not tackled by the system (demonstratives, relatives, interrogatives, personal and possessive pronouns of first and second person).

The first step was to run the individual systems over the uncorrected input. The success rate ranged from 49.7% for CogNIAC to 61.6% precision for Kennedy&Boguraev's method. The full results are displayed in Table 7.2.

The second step was to re-run the systems on the corrected input. As a result, the performance of all systems improved considerably, by up to 9% on one of the

files, with an average improvement of 6.5% in precision (full results are shown in Table 7.3). Discovering that pre-processing influences significantly and consistently the performance of the anaphora resolvers, the second step was to break-down the cause of errors.

By analysing the common indicators that all anaphora resolution systems used, we made the assumption that three main types of pre-processing errors could account for failures: mal-identification of noun phrases, errors in verb attachment and errors in the identification of the syntactic function of noun phrases. Several other types of errors (such as wrong delimitation of sentences) were considered important, but not frequent enough to allow space for investigation, therefore they were ignored.

The analysis of the individual influence of the selected types pre-processing errors has proved difficult, especially because they were strongly inter-connected: errors in verb attachment led to wrong identification of the syntactic function of noun phrases, just as the wrong identification of noun phrases did.

7.3.3.1 Misidentification of noun phrases

The analysis of the errors introduced by misidentification of noun phrases was done by matching the noun phrases in the un-corrected parser results with those in the post-edited results. This experiment showed that, although the noun phrase extractor did not eliminate any of the correct noun phrases, it introduced additional ones that made the search space for antecedents on average 12% larger. This obviously has an influence not only on the final accuracy of the anaphora resolvers, but in their time efficiency as well.

The second type of misidentification of noun phrases that introduced errors in the anaphora resolvers was the wrong delimitation of noun phrases. The main consequence of this type of error was that some to noun phrases are wrongly identified

as embedded. As all anaphora resolvers penalise embedded noun phrases, there have been cases where the correct antecedent was eliminated as a result. The example below shows this kind of misidentification, where the correct antecedent *the Beowulf HOWTO* was rejected in the favour of the wrongly identified noun phrase *a year*:

(7.1) Over <NP>a year <NP>*the Beowulf HOWTO*</NP> </NP> grew into a large document, and in August 1998 *it* was split into three.

Wrong delimitation of noun phrases was the most common type of error produced by the noun phrase extractor; nevertheless its influence could not be fully assessed due to the evaluation method employed. All methods considered the resolution of a pronoun correct if the antecedent found spanned a substring of the correct antecedent, including the head noun; if the type of evaluation took into account perfect matchings only, the influence of errors in the noun phrase extractor could be far more extensive.

The assessment of each individual error in the output of the noun phrase extractor has proved too time consuming, for this reason only a global assessment has been performed. This was done by using the post-edited results of the NP extractor and the initial, uncorrected output of the shallow parser[2]. It was noticed that all methods have approximately equally improved (about 10%), with a slightly higher improvement for CogNIAC and slightly lower for the machine learning method.

7.3.3.2 Verb phrase attachment

The second step was to assess the importance of correct VP attachment. VP attachment can influence pronoun resolution due to several rules that make

[2]Of course, as the noun phrase extractor was built on top of the output of the shallow parser, this uncorrected input conflicted with the corrected output of the noun phrase extractor; all conflicts were ignored, meaning that the corrected output was preferred to the initial one. The same observation applies to all subsequent experiments.

use of this information: collocation patterns (Mitkov), existential constructions (Kennedy&Boguraev), resolution of reflexives. By leaving the noun phrases as identified by the noun phrase extractor and with the un-corrected syntactical function and using the corrected results of the shallow parser, it was noticed that the method least sensitive to errors in the VP attachment was CogNIAC (1.9% improvement), while the most sensitive was Mitkov's (4.4% improvement).

7.3.3.3 Syntactic function

All anaphora resolution methods make use of information about syntactic function, in rules such as: preference for a subject antecedent, syntactic parallelism, resolution of reflexives to the subject, Mitkov's collocations pattern rule. An experiment involving the syntactic function, similar to the ones described before showed that CogNIAC was the most sensitive to pre-processing errors, while again the machine learning method was the least influenced. This experiment also showed that the identification of the syntactic function of noun phrases was relatively reliable, being the least important cause of errors in our implementation (approximately 1.2% improvement).

It has to be mentioned that none of these experiments capture the (unlikely) situation where a pronoun is correctly resolved due to errors in the pre-processing stage. Although theoretically possible, intuitively this possibility is too remote to benefit from a special treatment; we are nevertheless aware that such cases may occur and account for a small percent of the resolution performance.

Table 7.3 summarises the improvement in success rate obtained when using the input that was selectively corrected. The results are global, for all the evaluation files. These results do not fully reflect the influence of pre-processing on pronoun resolution, but rather give an estimate, due to the fact that the input was not entirely correct, and the errors are not independent.

	Mitkov	K&B	CogNIAC	ARMAL
Initial results	56.2%	61.6%	49.7%	59.8%
NP identification	67.7%	71.0%	61.7%	67.7%
Syntactical function	58.1%	62.8%	52.3%	61.7%
VP attachment	60.6%	63.9%	51.7%	62.5%
Perfect input	69.9%	75.2%	65.8%	74.9%

Table 7.3: Improvement of the success rate when using corrected input

7.3.3.4 Other types of errors

During the analysis of the data, it became apparent that some of the errors appearing in the pre-processing stage were not due to the malfunctioning of the parser, but to the composition of the text itself. In this category enter spelling mistakes (one of the most repetitive was the employment of *it's* as a possessive determiner instead of *its*), wrong verb agreement (*The drivers is...*), inconsistencies in using references to gender underspecified individuals (*The user* sometimes referred to by *they*, and later by *he*), missing punctuation marks (e.g, full stop at the end of sentence). All these errors directly reflect on the performance of the parser and propagate towards the anaphora resolvers. It was not considered necessary to correct any of the spelling or style mistakes, in order to preserve the character of the file; although the input is not the best quality, one has to take into account that this is the kind of texts usually found on the Internet, so any automatic natural language processing system should find ways of dealing with malformed input.

7.3.4 Reliability assessment

A drawback of existent anaphora resolution algorithms designed for English is that, to our knowledge, none of them applies a specialised treatment to different classes of pronouns[3]. This is even more surprising considering the fact that it has been theoretically acknowledged the fact that different pronouns have characteristic anaphoric properties.

Subsequently, an attempt was made to analyse the reliability of each of the methods in the identification of the types of pronouns resolved. All three methods targeted the same types of pronouns (personal-third person only, possessives and reflexives), which made the comparison reliable. It has to be mentioned that none of the methods apply specific resolution rules according to the type of pronoun processed (apart from reflexives, which will be discussed later). However, differences in resolution rates may result from the application of other rules, apparently not related to the type of pronoun (for example, verb attachment and grammatical function can indirectly distinguish between a personal pronoun and a possessive determiner). The success rate of all methods in the resolution of three categories of pronouns was calculated.

Firstly, a category based on the morphological type of the pronouns has been constructed: neuter singular pronoun (*it*), masculine and feminine singular pronouns (*he* and *she*), plural pronouns (*they*), possessives (*his*, *her*, *their*, *its*) and reflexives (*himself, herself, itself, themselves*). Table 7.4 describes the resolution accuracy of each method for each type of pronoun[4]. As it can be noticed, results are not included for the resolution of reflexives. This is due to the fact that there has been no significant difference between the methods with respect to the resolution of reflexives (only a

[3]The only type of pronouns that benefit of a special analysis in most anaphora resolvers are reflexives. Some restricted treatment of possessives is performed in Kennedy&Boguraev's method, but it is combined with more general rules applying to all pronouns.

[4]However, it has to be noted that Mitkov's method was initially designed for the resolution of the pronoun *it* only.

very small number appeared in the testing corpus); this can be explained by the fact that all methods use the same constraints drawn from the Government and Binding Theory, which are never violated in the occurrences found in our corpus. Therefore, the reflexives have been omitted from all the subsequent results reporting.

	Mitkov	CogNIAC	K&B	ARMAL
it	63.2%	43.2%	70.2%	59.5%
he, she	30%	22.2%	66.6%	55.5%
they	50.6%	38.9%	54.5%	45.4%
possessives	40%	30%	73.3%	78.3%

Table 7.4: Success rate according to the morphological category of pronouns

The second classification was based on the syntactic function of the pronouns: subject, direct object, indirect object, attributive and others.

	Mitkov	K&B	CogNIAC	ARMAL
Subj	40.7%	48.2%	38.9%	50.4%
Dir obj	55.1%	34.8%	40.6%	56.5%
Ind obj	33.3%	66.7%	33.3%	66.7%
Attributive	76.5%	86.8%	63.2%	51.5%
Other	50.0%	50.0%	50.0%	100.0%

Table 7.5: Success rate according to the syntactic function of pronouns

The third category was based on the distance (in number of intervening noun phrases and sentences) between the anaphor and the real antecedent: one or two noun phrases (same sentence), same sentence more than two intervening noun phrases, previous sentence, distance greater than one sentence. Statistics were collected from

the corpus for the success rate of all the methods for each category of pronouns.

	Mitkov	CogNIAC	K&B	ARMAL
1 or 2 NPs	80.3%	83.6%	95.1%	60.7%
more than 2 NPs	36.9%	35.1%	66.7%	58.6%
previous sentence	59.7%	66.0%	23.3%	61.6%
more than one sentence	65.6%	81.3%	40.6%	59.4%

Table 7.6: Success rate according to the distance between anaphor and antecedent

7.3.5 A probabilistic anaphora resolver

This section presents a hybrid system for anaphora resolution that uses a probability model to calculate the likelihood of a method to correctly solve a certain type of pronoun.

Unlike Ge&Charniak's statistical model [Ge et al.1998], this is a much simpler probabilistic system, that only chooses between candidates already proposed by other anaphora resolvers. It is therefore not an independent system, does not perform a search for an antecedent in the space of possible antecedents, does not incorporate any new knowledge sources and does not aim at achieving ground-breaking accuracy rates. The only goal of this model was to show that it is still possible to improve the performance of anaphora resolvers by simply using the same knowledge sources in different ways, and by taking the best out of classical ideas.

7.3.5.1 Description

Intuitively, given a pronoun and the antecedents identified by the four systems, the method tries to estimate which is probability for a certain system to have found the correct antecedent, given that:

- the pronoun had the morphological function *m* (m in the set {*it*/*he*/*she, they*/*them, possessive*})
- the pronoun had the syntactic function *s* (s in the set {*subject, direct object, indirect object, attributive*})
- the distance between the pronoun and the antecedent was *d* (*d* in the set {*1*/*2 NPs in the same sentence, same sentence and more than 2 NPs, previous sentence, more than 1 sentence*})
- the antecedent had the syntactic function *as* (*as* in the same set as *s*)

On the basis of the statistics collected from the training corpus, the probabilities for each pair <pronoun, system> were calculated and the candidate found by the system that maximised the probability was selected as correct antecedent.

As mentioned before, there was no significant difference between the systems in the resolution of reflexives, thus it cannot be assumed that a certain system is more likely to solve a reflexive than an other. In this case, the antecedent returned by Kennedy&Boguraev's method was always selected. This was only done for the purpose of consistent comparative evaluation.

7.3.5.2 Evaluation

The qualitative and comparative evaluation envisaged comparison with the individual methods and with a combined method.

Testing corpus

In order to assess the performance of the new system, it has been evaluated it on unseen data, independent of the observation corpus described in section 7.3.2. The testing corpus consisted of 3 technical manuals, containing 113 pronouns, out of which 86 anaphoric.

A simple voting system

In order to evaluate the new system, we had at the same time in mind the time efficiency, so important in applications where anaphora resolution is only a component. For that reason, an attempt was made to show that the results obtained could not be surpassed or equaled by using a much less time-consuming system that combines the three methods using a simple voting procedure. Hence, a voting system was implemented that considers as correct a result reported by the majority of the systems; in case of a tie, the correct result was the one returned by Kennedy&Boguraev's system, as the one that outperforms systematically the others. The voting system was used as a baseline. By evaluating the baseline against the three systems, an improvement of up to 4% in success rate was noticed, compared to the best results on one testing file; the average improvement for all testing files was about 2%. However, this improvement was not consistent across all files used for testing. In some cases, the best results of the systems were better than the results of the baseline, therefore the voting system has decreased the performance of the best independent system.

Results

Table 7.7 displays the results obtained when running the hybrid system, as compared to the results of the individual systems and of the combined baseline. It can be easily seen that the increase in performance is significant, up to 7.5% over the best individual system on one of the files; the average improvement for all testing files over the best system was 4.6%. More important, the improvement is consistent over all testing files

	Success rate					
	Mitkov	CogNIAC	K&B	ARMAL	Baseline	Hybrid
Observation corpus	56.2%	49.7%	61.6%	59.8%	63.3%	73.0%
Unseen data	58.1%	52.3%	62.8%	62.7%	63.9%	67.3%

Table 7.7: Final evaluation results

and the hybrid system always outperforms the baseline.

We are aware of the fact that the small amount of training data does not allows us to draw a definite conclusion as to the resolution power of the individual systems, the differences in resolution not being statistically significant. Therefore, the probabilities do not fully express the likelihood of a certain method to be preferred over an other when resolving a certain type of pronoun. Nevertheless, the results show that the improvement in performance is consistent and significant.

7.4 Morphological agreement in anaphora resolution

In anaphora resolution, morphological agreement is often considered a hard constraint, i.e., noun phrases have to obey number and gender agreement with the anaphor if they are to be considered as possible antecedents. Some previous works [Barlow1998] warned about the danger of restricting too much the searching space for antecedents by considering morphological agreement as a hard constraint, thus introducing errors in the resolution process. The problem of gender agreement has been practically addressed more extensively [Hale and Charniak1998, Orăsan and Evans2001], but, although the problem of some cases of number disagreement have been previously identified and discussed [Barlow1998, Denber1998, Mitkov2002], no large scale

corpus-based investigation has been conducted to show the extent to which plural pronouns can refer to other constituents than simple plural noun phrases. Identifying the possible causes of number disagreement allows us to design methods for tackling these pronouns in an automatic anaphora resolution system.

A glance at the state-of-the-art literature in automatic anaphora resolution shows that most methods assume the necessity of number and gender agreement between the anaphor and the antecedent. This strong constraint is based on the intuition that anaphorically linked elements must share similar morphological and semantic features. Morphological agreement is used as a mechanism for reducing ambiguity by eliminating those noun phrases that do not agree with a pronoun from the list of possible antecedents of that pronoun.

This is the approach used by Lappin&Leass's RAP [Lappin and Leass1994], by Kennedy&Boguraev's parser-free resolution method [Kennedy and Boguraev1996] and by Mitkov's knowldege-poor algorithm [Mitkov1998]. To our knowledge, there has been no study indicating what percent of the errors in the resolution of pronouns is due to the elimination of the correct antecedent as a result of gender and number disagreement.

The importance of morphological agreement has been proved by several psycholinguistic studies, that put in evidence the fact that it influences the speed of interpretation in human anaphora resolution. Chapter 3 discusses some of the psycholinguistic studies that investigate the role of gender and number cues in the identification of antecedents.

In this work, the focus was only on the number disagreement phenomenon. This decision was based on the fact that gender (dis)agreement is not a real problem in automatic pronoun resolution for English. As English does not feature grammatical gender, the gender agreement problem is reduced to identifying animate and

inanimate references, therefore entities that can be referred to by both neutral (*it*) and animated, masculine or feminine (*he/she*) pronouns. This information can be used for improving the resolution of singular pronouns and it has been tackled in [Hale and Charniak1998], [Orăsan and Evans2001].

7.4.1 Corpus-based investigation

The corpus investigation had three main goals: to identify which are the factors that lay behind number disagreement, to assess how frequent this phenomena is in real texts and to assess the distribution of different cases of disagreement across several text genres. As a direct application of the corpus results, it was attempted to see which of these cases can be treated automatically and which can be categorised as more complex cases.

7.4.1.1 Corpus

The investigation was based on four types of text: technical manuals (the corpus descried in Chapter 4), narrative texts, health promotion and medical information leaflets and diverse newspapers articles (the last three being extracted from the British National Corpus [Burnard1995]). The narrative texts were extracts from 4 novels, totalling approximately 38000 words, the health promotion material was extracted from three documents describing the activity of the ACET AIDS organisation, while the newspapers extracts contained politics and sports information. The technical texts were Linux manuals and were included in the analysis in spite of the low number of plurals, due to the fact that they were annotated for coreference, therefore allowing us to perform an automatic analysis of the plural disagreement. Extracts from multiple texts were preferred to contiguous files in order to avoid certain usages of plurals that hold to the style of individual authors. The total number of plural pronouns was about

2500.

Table 7.8 describes the content of the corpus and the proportion of plural pronouns.

	Technical	narrative	medical	news
#files	5	3	3	2
#words	30000	38000	39500	36600
#plural pronouns	102	994	724	730

Table 7.8: Corpus

7.4.2 Methodology

The analysis consisted of two stages: identifying the situations of number disagreement and analysing their distribution in the corpus. The initial assumption was that the basic case of number agreement is represented by a plural pronoun referring to a plural noun phrase; everything else was considered as an exception. Following this basic definition of an ”exception”, approximately 300 pairs of plural pronouns and antecedents were collected and further classified into finer categories. The classification was done with respect to the automatic treatment of plurals. For example, there is no semantic difference between an antecedent consisting in a sequence of coordinated noun phrases and a split antecedent (consisting in a number of noun phrases further apart in the text). However, the automatic identification of a split antecedent requires a far greater amount of computation, therefore the two cases were classified in different categories. After the initial classification, ten cases of number disagreement were obtained. Nevertheless, the distinction between some of the categories was sometimes too fine; this resulted in problems of classification

and, from a more pragmatic point of view, did not bring any contribution to the automatic treatment of plural pronouns. Therefore, some of the less frequent and more similar categories were collapsed, resulting in the final classification described below. After agreeing on the classification, the remaining corpus has been analysed and the instances of number mismatches have been classified.

When analysing the texts, only those cases that displayed genuine disagreement had to be considered. Consequently, two assumptions have been made:

- first, the number of the pronoun has to be compared with the number of the last full noun phrase in its coreferential chain
- second, if a coreferential chain contained more than one occurrence of number mismatch, only the first occurrence was taken into account.

The first assumption was made the view of the efficiency of the analysis in mind. As the corpus was not previously annotated with coreferential links, and all the analysis was performed manually, it seemed extremely time consuming to try and identify the head of the coreferential chain containing a case of pronoun disagreement.

The second assumptions envisaged cases such as the sentence below, where the plural pronoun ”they” appears twice, referring to the indefinite pronoun ”everybody”:

(7.2) She had soon learned that almost *everybody* has something *they* want to hide, and something *they*'re eager to share.

However, only the first instance of the pronoun was taken into account as a genuine case of disagreement. This is consistent with the intuition that subsequent references can be interpreted with respect to the last element in the chain, independently of the head of the chain. The transitivity of the coreference relation ensures the fact that the mental representation constructed for an element in the chain gathers the semantic information of all the previous references.

7.4.3 Cases of apparent number disagreement

The final classification consisted of the following eight categories:

1. The antecedent is a **conjunction/disjunction of plural or singular NPs**:

(7.3) You know *the Daily Mirror, and the Sun, and ITV, and the Unions*, what are *they* telling people to do?

2. The pronoun has a **split antecedent**

(7.4) Only when *they* hang up did *Jay* realise that *she* hadn't given her a date.

3. **Collective nouns**

This category includes both collective noun such as *police*, *government*, *army*, singular nouns denoting more than one person, such as *a group of people*, *a number of people*, names of companies or associations.

(7.5) It belongs to *the Customs and Excise mob*.*They're* not using it any more.

4. **Class representation**

Singular noun phrases that stand for a class of entities can be referred to by plural pronouns in English, as in:

(7.6) I like cats, but I would much rather get *a dog*, because *they* are such clever animals.

5. **Gender underspecification**

In a context where the antecedent is a person, but does not feature grammatical gender, and the speaker is not aware of the gender of the person, it can be referred to by a plural pronoun, as an alternative to "he or she":

(7.7) You were called on the 30th of April at 21:38 hours. *The caller* withheld *their* number.

6. Plural pronouns that refer to a **quantified noun or indefinite pronoun** (”someone”, ”every person”)

(7.8) *Someone* will remember to wake me up early in the morning, won't *they*?

7. **Generic plurals**

Sometimes plural pronouns are used with an impersonal sense, therefore there is no antecedent in the text and it cannot be inferred from any other entity in the text, as in:

(7.9) *They* have lessons in everything these days, don't *they*?

8. **Indirect anaphora**

This is the case of plural pronouns whose interpretation is triggered by another entity in the text (possibly of singular number), as in:

(7.10) My sister's wedding was beautiful. *They* were the happiest couple.

where ”they” is interpreted as ”my sister and her husband”; this relation is established through a chain of mental inferences that links the textual elements ”wedding” and ”couple” to the inferred antecedent.

This last case is not a genuine case of disagreement, as one cannot talk about morphological agreement unless in the context of coreference. However, acknowledging and identifying such cases helps filtering unlikely candidates.

7.4.4 Results

Table 7.9 shows the distribution of the eight types of number mismatches across the four genres of texts considered. For each class and each type of text, the number of occurrences and the percentage reported to the total number of exceptions in that file are provided. The last column displays the percentage of each type of exception with respect to the total number of exceptions in all documents.

	technical		narrative		medical		newspapers		Total	%
Class 1	6	17.6%	56	21.2%	55	27.7%	28	24.3%	145	23.7%
Class 2	2	5.8%	49	18.6%	26	13.3%	14	12.1%	91	14.8%
Class 3	13	38.2%	44	16.7%	41	20.7%	22	19.1%	120	19.6%
Class 4	5	14.7%	15	5.70%	38	19.1%	10	8.69%	68	11.1%
Class 5	7	20.5%	11	4.18%	19	9.5%	19	16.5%	56	9.1%
Class 6	5	14.7%	12	4.56%	16	8.08%	17	14.7%	50	8.1%
Class 7	0	0%	54	20.5%	1	0.5%	1	0.8%	56	9.1%
Class 8	0	0%	22	8.36%	2	0.1%	4	3.4%	28	4.5%
Total	34		263		198		115		612	

Table 7.9: Distribution of plural pronouns

7.4.5 Interpretation

A first look at the results shows that out of the 2500 pronouns inspected, 612 were exceptions, meaning that almost a quarter of the plural pronouns could not be interpreted as referring to a plural noun phrase. The largest percentage of exceptions was displayed by the technical manuals, with 36 out of 102 plurals (35.29%) constituting exceptions. The widest variety of disagreement cases was displayed by the narrative texts, which contained a significant number of pronouns in each category. Overall, the most common case of disagreement seems to be represented by references to coordinated noun phrases, while the least common was the indirect anaphora. In fact, only narrative texts contained a significant number of indirect anaphora cases, in the other types of text the frequency of these cases being far below the frequency of the other cases of disagreement. The percentage of plurals referring to quantified nouns

was, as predicted, proportional with the frequency of quantified nouns in the different types of text.

Among the disagreements present in technical manuals, the largest category was represented by references to collective nouns. A closer look showed that most of them were names of companies. The second most important category were references to a class representative, while generic plurals and indirect anaphora never appeared in our texts. Pronouns with split antecedents were also poorly represented.

The medical information documents contained approximately the same percentage of pronouns referring to collective nouns as pronouns referring to a class representative. Most collective nouns were names of organisations, companies, hospitals, as well as nouns expressing the idea of collectivity (*a number, folk, mob*). Narrative texts contained an unexpectedly high number of generic plurals, which was only exceeded by the number of plurals with coordinated antecedent. It was also remarkable the low number of disagreement cases due to gender under-specification.

We do not believe that the data analysed is sufficient for us to comment on the linguistic and stylistic reasons and implications of the distribution of pronouns, and this was beyond the aim of our work. However, this investigation gives a good starting point for identifying those cases that are frequent enough to deserve a separate computational treatment in an automatic anaphora resolver.

7.4.6 A practical approach

7.4.6.1 Experiments

Table 7.10 presents the results reported by three anaphora resolvers in the resolution of plural pronouns. The evaluation was performed on the same set of technical manuals, which were manually annotated for coreferential links. The evaluation measure used

was the *precision*. We are presenting here the results obtained by Mitkov's knowledge-poor approach[5], Kennedy&Boguraev's parser-free method and Baldwin's CogNIAC as implemented in the evaluation workbench.

	#plurals	Mitkov	K&B	CogNIAC
WIN	6	0%	50.0%	50.0%
ACC	39	48.7%	48.7%	43.5%
CDR	24	29.1%	30.0%	29.1%
BEO	17	35.2%	41.17%	35.2%
Total	86	37.2%	43.0%	38.37%

Table 7.10: Resolution rates of plural pronouns in technical manuals

Two things can be noticed in the results presented in Table 7.10: first, the low success rate in the resolution of plurals, and second, the fact that the results obtained with the three methods are very close to each other. The first observation should be made in the context of the evaluation of the resolution on all types of pronouns. Evaluating the aforementioned methods on the same texts, [Barbu and Mitkov2001] report a precision ranging from 51% for CogNIAC up to 71% for Kennedy&Boguraev, therefore sensibly higher than for the interpretation of plurals alone. The second observation leads us to believe that the failures are not entirely due to the malfunctioning of the methods.

The small number of plurals in these texts allowed for a manual experiment to be performed, that was aimed at finding the upper limit of the resolution rate. In order to do this, the number of the antecedents of the plurals were manually corrected, so that there was no disagreement. For split antecedents, the number indication from each

[5]As mentioned previously, Mitkov's method was only designed for the resolution of the singular third person pronoun *it*, and does not tackle plural pronouns in a specialised way.

of the components was removed, and a resolution was considered correct if one of the components was identified as antecedent. No other features have been corrected from the output of the parser. By running the three systems on the corrected input, the results described in Table 7.11 have been obtained . Of course, the resolution of plurals is not only dependent on the correct identification of the number of the antecedent, many other types of pre-processing errors being responsible for failures in pronoun resolution. These pre-processing errors further add to those induced by the malfunctioning of the anaphora resolvers themselves.

	#plurals	Mitkov	K&B	CogNIAC
WIN	6	83.3%%	100.0%	75.0%
ACC	39	76.9%	82.0%	61.5%
CDR	24	66.6%	70.8%	58.3%
BEO	17	64.7%	76.4%	52.9%
Total	86	72.0%	79.0%	59.3%

Table 7.11: Accuracy rates on corrected input

7.4.7 Tackling difficult cases

The corpus investigation showed that the most frequent types of number disagreement belonged to categories 1 (coordinated antecedents), 4 (class representation) and 3 (collective nouns). This gives us an indication as to the areas that can be improved in anaphora resolution for achieving better resolution accuracy.

We have tried to identify which of the cases described above can be easily tackled in an automatic anaphora resolver, and which can be at least identified as difficult. The three categories of disagreement that could be solved more easily are:

reference to coordinated NPs, reference to collective nouns and references to indefinite pronouns and quantified nouns. Some basic rules for identifying references to a class representative could also be attempted, while solving indirect anaphora definitely requires the most amount of knowledge and the most complicated inferential process.

Coordinated NPs

This case is easily tackled automatically, since coordinated NPs can be identified with a high degree of accuracy using a small number or rules. The resulting noun phrase will be allowed to function as an antecedent candidate for a plural pronoun. At the same time, none of the constituents of such a composed noun phrase should be allowed as antecedent for a plural pronoun. For example, in ”Tom, John and Mary went to the cinema. They saw a comedy”, it is impossible for *they* to refer to any of the groups {Tom, John} and {John and Mary}.

Collective nouns

Collective nouns such as ”government” or ”police” constitute a restricted set and therefore can be singled out using a basic lexicon. The same applies to noun phrases pre-modified by a collective noun (*a number of, a group of*). A more difficult problem is posed by names of companies, which have to be identified as such by a named entity recogniser. A simple grammar consisting of a small number of rules has been implemented that identifies as a company name a capitalised string followed by one of the suffixes:*inc., co., ltd., lab., corp.* (or the full denomination *Corporation, Limited, Laboratories, Incorporated*). Experiments performed on technical manuals showed that when using this restricted grammar, the resolution of plural pronouns referring to companies improved by about 55%. In general, named entity recognisers perform extremely well, with an accuracy of classification approaching 97-98%.

Quantified noun/Indefinite pronoun antecedent

In most cases, such noun phrases are referred to by plural pronouns, therefore easy rules can be implemented that allow them as antecedents for plural pronouns. Moreover, singular pronouns should normally not be allowed to refer to quantified nouns or indefinite pronouns, unless in special circumstances.

Class representation

Identifying singular noun phrases that stand for a class of entities is not a trivial task. The only simple solution would be to allow non-definite noun phrases to function as antecedents for plural pronouns. This could help solving cases where there is no plural noun phrase in the text that can be antecedent. However, the method is likely to introduce errors in the resolution of plural pronouns, by making the search space too wide.

7.5 World knowledge in anaphora resolution

There has been stressed repeatedly the importance of world knowledge in the interpretation of anaphora. Sometimes, the argument went so far as to ascertain that anaphora resolution is not possible altogether without world knowledge. Although this is somewhat an exaggerated point of view, it is nevertheless clear that world knowledge plays an important role in resolving anaphoric links. This section presents a somewhat restricted investigation of this role.

(7.11) As of the 1.1.38 kernel, the sbpcd driver ejects the CD when *it* is unmounted.

In this example, even human readers would have problems understanding that the pronoun *it* refers to the CD and not to the driver. Knowledge about the fact that in

Linux the *mount* operation is performed on a file system (in this case, CD), and not on a driver, is necessary. However, it is possible to solve automatically the anaphoric relation using collocation patterns: the phrases *mount/unmount a/the CD* and *the CD is mounted/unmounted* appear several times in the same document. So the problem is: is this a case where world or domain knowledge is absolutely necessary or is it a case where the type of information necessary can be acquired automatically?

Other cases appear to be even simpler:

(7.12) `Put the CD in the drive and mount it.`

In this case, *the CD* can be identified as the antecedent only on syntactic grounds, due to the syntactic parallelism preference. However, the problem for human readers remains the same as in the previous example.

The purpose of the investigation that will be described in the following is to determine to which extend it is possible to perform pronoun resolution without world knowledge; in fact, in how many cases world knowledge is an essential factor in solving pronouns.

7.5.1 What is world knowledge?

The first question that needs clarifying is defining the categories of information that can be considered *world knowledge*.

- Semantic restrictions
- Domain knowledge (in our case, technical knowledge about the Linux environment)
- Cultural and localisation knowledge.

7.5.2 Methodology

Performing a large scale investigation of this sort is likely to be a time consuming task. Therefore, some sort of filtering of pronouns was considered necessary. The method chosen was to select for investigation only those pronouns that have been solved incorrectly by at least 3 of 4 anaphora resolvers. The 4 anaphora resolvers selected have been Hobbs', Kennedy&Boguraev's, Mitkov's (as implemented in the evaluation workbench) and ARMAL using memory based learning. This helps reducing the risk for a pronoun to have been correctly solved by chance. A further 10% of the remaining pronouns (that have been solved correctly by more than 2 of the anaphora resolvers) are also randomly selected in order to account for further risks of the anaphora resolvers agreeing by chance.

However, there will still be cases where pronouns are solved correctly due to morpho-syntactic factors (as in example 7.2).

7.5.3 Results and discussion

Out of 545 anaphoric pronouns, 92 were solved incorrectly by at least 3 of the methods. A further 54 pronouns selected randomly out of the remaining pronouns were added to the investigation, thus resulting in a total number of 146 pronouns.

Each pronoun has been subjected to an analysis that envisaged the sources of knowledge necessary for correctly solving it. More than a half of the pronouns (81) did not require world knowledge at all, therefore their misinterpretation was most likely due to other causes, like errors in the pre-processing or in the anaphora resolution methods. In analysing the remaining 65 pronouns, the main problem was the identification of cases where world knowledge was essential. Approximately one half of the pronouns (34) were subject to semantic restrictions and could, at least

theoretically, have been solved if such information was available. However, most of the pronouns could have been solved anyway if collocation information was used as a meaning of acquiring semantic restrictions. In a few of the remaining cases, the pronouns required some kind of world knowledge, but they have been solved correctly on the basis of syntactic information alone. In similar cases, some misinterpreted pronouns required world knowledge, but could have been solved correctly on the basis of syntactic information; errors in the pre-processing prevented this from happening, however.

7.6 Conclusions

This chapter has investigated some of the difficult issues in pronoun resolution, problems that are the main cause of errors in automatic resolution of pronouns.

The series of experiments presented envisaged pre-processing errors (morpho-syntactic errors in particular, since this is the major knowledge source used in knowledge-poor approaches), morphological disagreement in the resolution of pronouns and the lack of world and semantic knowledge. The investigation was backed by relatively extensive corpus analysis.

Chapter 8

Conclusions

8.1 General conclusions

This thesis has presented research in the area of bilingual pronoun resolution. The main aims of the research were to investigate the possibility of using machine learning for multilingual anaphora resolution and to design and implement a system for anaphora resolution that uses these methods.

These aims were fulfilled through the implementation of ARMAL, an extendable system for anaphora resolution based on a combination of rule based and memory based techniques. The system was designed for addressing anaphora resolution for English and French, but, since it is largely based on language-independent features, has the potential of being extended to other languages with similar pronominal systems.

Apart from these two main aims, some other spin-off threads of research have developed into individual research topics on their own right.

Chapter 5 presents an alternative to ARMAL based on genetic algorithms. Although the results obtained by this algorithm are lower than those obtained by ARMAL, the method has some clear advantages: it does not require a large amount

of annotated data for training and it is easily expandable to accommodate more knowledge sources as they become available.

The development of ARMAL has been supported by the development of a corpus of English and French texts annotated for coreference. This is a reusable resource that is an integral part of the efforts of the University of Wolverhampton Computational Linguistics Group to develop annotated corpora for coreference. Such resources are scarce, hence every freely available annotated corpus is an important contribution to the research community.

An important part of the research has been the evaluation of anaphora resolution algorithms in general and of ARMAL in particular. The evaluation workbench described in Chapter 6 has proved a valuable resource for comparative evaluation. Apart from being used in the evaluation of ARMAL, it has facilitated the analysis of the output of other anaphora resolvers, allowing for the comparison of errors and types of errors. The evaluation workbench has been used in every piece of research that involved evaluation, presented in this thesis.

Also related to evaluation issues have been a series of experiments relating to sources of errors in anaphora resolution. These experiments, described in Chapter 7, envisaged three main types of problems: the treatment of plural pronouns, morpho-syntactic errors and the necessity of world knowledge in anaphora resolution. The experiments have been performed manually or semi-automatically on a significant number of pronouns and have highlighted a series of important issues with impact in both anaphora resolution and evaluation of anaphora resolution systems.

8.2 Aims and objectives revisited

The introductory chapter presented the aims and objectives of this research. At this point, it can be summarised the way these aims have been achieved in the thesis.

- 1.1 - overview of the work done so far in the field

  This aim was fulfilled in Chapter 3, that presents the most relevant approaches to anaphora resolution and in particular to pronoun resolution.

- 1.2 - identify the drawbacks and the achievements of existing anaphora resolvers and identify areas that can be improved

  Most of the work conducted in the field is based on handcrafted rules, while some approaches use machine learning. The main drawback of these approaches is that they do not provide customised treatment of different types of anaphors, and most systems are tailored for a specific language, thus not suitable for multilingual applications. A detailed description of these advantages and drawbacks is present in Chapter 3.

- 2.1 - identify the common and distinct features of pronominal systems in English and French

  The investigation of differences between the English and the French pronominal systems is presented in Chapter 2.

- 2.2 - identify the largest set of common rules for English and French

  As a direct consequence of the aforementioned investigation, the common features of English and French pronouns have been extracted and this set of features was presented in Chapter 5.

- 3.1 - identify features to be used in automatic anaphora resolution

  The set of features to be used by the machine learning system consists of

language-specific features and language-independent ones. They are presented in detail in Chapter 5.

- 4.1 - decide the text genre and identify suitable texts

  The target text genre is technical manuals and documentation. The choice of the genre and of suitable text is explained and detailed in Chapter 4.

- 4.2 - collect and annotate data for use in training and evaluation

  A corpus of about 50000 words for each language has been collected and annotated for coreferential links. The corpus is described in Chapter 4.

- 5.1 - assess the suitability of machine learning techniques in the resolution of anaphora

  Previous machine learning approaches to anaphora resolution are described in Chapter 3 and outlined again in Chapter 5. The reason why machine learning is considered suitable for anaphora resolution is detailed in Chapter 5.

- 5.2 - design a system for bilingual anaphora resolution

  The original system, named ARMAL, is a hybrid method that combines handcrafted rules and automatic learning. It is described in detail in Chapter 5.

- 5.3 - select the most suitable learning strategy for the system

  Whilst designing the system, several machine learning techniques have been experimented with. The outcome of the experiments showed that memory-based learning performs best for the task. The experiments are discussed in Chapter 5.

- 6.1 - design an evaluation strategy

  Designing the evaluation strategy involved choosing the evaluation measure and

the evaluation experiments. A detailed presentation of the evaluation strategy is done in Chapter 6.

- 6.2 - develop an evaluation workbench for pronominal anaphora resolution systems

  The evaluation workbench (described in Chapter 6) allows for an anaphora resolver to be compared against some of the best known rule-based anaphora resolver.

- 6.3 - perform extensive quantitative and qualitative evaluation of the aforementioned system

  The performance of the ARMAL system was evaluated according to the evaluation strategy and the results of the evaluation are presented in Chapter 6.

- 7.1 - investigate the possibility of mutual enhancement within a multilingual anaphora resolver This aim is fulfilled through the implementation of a mutual enhancing anaphora resolver. The system is described in Chapter 7.

- 8.1 - identify the main difficulties in automatic anaphora resolution

  This issue is addressed in Chapter 7. Three main types of difficulties have been identified and discussed.

- 8.2 - perform a corpus based analysis of the most common errors

  This investigation is described in Chapter 7.

8.3 General overview of the thesis

In the following, each of the chapters of the thesis will be revisited, and the extent to which they fulfil the aims of the investigation will be discussed.

Chapter 1

Chapter 1 was an introduction into the research topic. The chapter defines the main aim of the research (investigation into multilingual pronoun resolution) and provides a breakdown of this aim into smaller problems that will be addressed in the thesis. The research topic is justified by means of its usability in the larger context of multilingual NLP applications. This chapter also defines the scope of the research (languages to be used - English and French, types of pronouns to be tackled) and the basic methodology used in the investigation (machine learning techniques).

Chapter 2

The aim of this chapter was to describe the possible anaphoric and non-anaphoric uses of personal, possessive and demonstrative pronouns in both English and French. This investigation led to some interesting observations that may help in the two stages of automatic pronoun resolution: filtering (rejecting elements that cannot possibly be referred to by a pronoun) and identification of antecedents. There have also been noticed some cases where the same type of pronoun presents different degrees of accessibility to discourse elements in English as compared to French. This implies that different language specific rules have to be applied in some cases. However, the majority of anaphoric expressions, or at least the most frequently used, have the same behaviour in English and in French, which makes it possible to use the same devices for their resolution. This chapter fulfils aim 2.1, which is the identification of common and distinct features of the English and French pronominal systems.

Chapter 3

This chapter was an introduction into the problem of automatic anaphora resolution. There have been three different issues addressed in this chapter:

- *Knowledge sources* used in automatic anaphora resolution, such as morphological, syntactical and discourse constraints

- *Pre-processing tools* used for accessing the knowledge sources and providing the information necessary to anaphora resolvers.

- *Types of anaphora resolvers*, ranging from traditional rule-based methods to newly emerging machine learning methods. Within this section the problem of automatic multilingual anaphora resolution has also been addressed, the main conclusion being that the few works carried out in this direction showed promising results but also showed there is still area for improvement.

This chapter has partially fulfilled aim 5.1 by presenting some of the existent machine learning approaches to anaphora resolution.

Chapter 4

This chapter has described the use of annotated corpora in anaphora resolution, and in particular in the ARMAL system. Far from being an easy task, the annotation of anaphoric links poses a number of problems to the human annotator, which inevitably reflect upon the quality of the corpus produced.

It has been shown that using technical manuals as the target of the anaphora resolver is a justifiable choice in the context of domain-specific NLP applications.

The corpus developed for this project consisted of approximately 55000 words (654 pronouns) for English and 36000 words (482 pronouns) for French. The corpus represents on its own a by-product of the research, being a valuable resource for further experiments on anaphora resolution. A preliminary analysis of the corpus composition and the distribution of pronouns in the corpus revealed some possible directions to be taken in the implementation of the anaphora resolver.

The choice of the corpus genre fulfilled aim 4.1, while the collection and annotation of the data fulfilled aim 4.2.

Chapter 5

This chapter has presented the main practical outcome of this project, ARMAL, a system for bilingual anaphora resolution based on a hybrid approach. The main idea is to use filtering rules for solving easy cases of anaphora and to send the remaining anaphors to a decision tree-based classifier.

Machine learning has been selected as the preferred approach to the problem of bilingual anaphora resolution due to a number of factors that have been identified in this chapter. The assessment of the suitability of machine learning for the present task has fulfilled the aim 5.1.

During the design of the machine learning approach, a set of 20 features have been identified, most of them being language-independent. The identification of the features followed a corpus analysis of the factors that influence pronoun resolution in English and French. Determining the training features has fulfilled the aim 3.1.

Choosing a precise learning method from a range of machine learning strategies has been done experimentally. The final option was a combination of memory based learning and decision trees. The choice of this method has fulfilled aim 5.3.

A cheaper but less successful alternative is also presented in the form of a genetic algorithm that works on raw data. The development and analysis of this method proved that it is less suitable for anaphora resolution than the memory based approach used in ARMAL. This investigation has partially fulfilled aim 5.3.

Chapter 6

This chapter has addressed the problem of evaluation of anaphora resolution systems. The first part of the chapter was dedicated to general issues in automatic

evaluation, followed by a discussion of the evaluation strategy for ARMAL. The qualitative evaluation showed a performance around 44% precision (51% success rate) for French and slightly higher for English(47% precision, 59% success rate), while the comparative evaluation situated ARMAL at an average level in rapport to state-of-the-art anaphora resolvers.

In a different line of investigation, the genetic algorithm for anaphora resolution described in Chapter 5 is evaluated as well. Its evaluation shows that although a promising idea, the GA does not prove a valuable alternative to the supervised learning method used in ARMAL.

The last part of the chapter describes an approach to mutual enhancement of anaphora resolvers for English and French by way of parallel corpora.

Through the design of an evaluation methodology, aim 6.1 was fulfilled. The development of the evaluation workbench fulfilled aim 6.2. This chapter also fulfilled aims 6.3 and 6.4, which refer to the evaluation of ARMAL and of the genetic algorithm. The design of the mutual enhancement strategy fulfilled aim 7.1.

Chapter 7

This chapter discusses some of the difficulties in anaphora resolution. Section 1 presents a general overview of the most common and toughest problems in pronoun resolution. The following sections address in depth more specific problems.

First, the main types of general errors are identified and a corpus based analysis of their frequency is performed. Secondly, the problem of morphological disagreement is studied in depth and a large scale corpus analysis is presented, showing the distribution and frequency of the phenomena across different text genres. Thirdly, errors specific to each language investigated are identified.

This chapter fulfilled the aims 7.2 and 7.3, which refer to the investigation of the

causes of errors in anaphora resolution.

8.4 Further work

During the research and development of ARMAL, a series of problems have remained open, either due to the unavailability of resources or to time restrictions.

The potential of ARMAL to be used as a bilingual system for English and French has been proved through extensive evaluation. The problem that remains open is assessing the feasibility of extending ARMAL for other languages. From a theoretical point of view, it is possible to use ARMAL for another language with a minimum effort of implementation, however this was not proved experimentally. This problem was not tackled due to the fact that such an extension requires pre-processing tools for other languages that were unavailable to us at the time of the development and a fair amount of annotated corpora for other languages, which was unavailable as well.

A second line of research that could be pursued having as starting point ARMAL would be the introduction of knowledge-rich learning features, to complement the simpler morpho-syntactical and topological features that ARMAL currently uses. Such features could be, for example, selectional restrictions or collocation patterns. The idea of using semantic information has been experimented with during the implementation of ARMAL, and an early method for acquiring selectional restrictions via Internet searching has been implemented. However, although promising, this direction of research has been abandoned due to the prohibitive time cost involved in performing the search operations. It is nevertheless clear that the use of semantic information is the way forward towards achieving higher performance in anaphora resolution, and the acquisition of semantic information using the Web as a corpus should be pursued further.

Another thread of investigation worth following up is the improvement of the genetic algorithm described in Chapter 5. The idea that the genetic algorithm is based upon has got the potential to advance anaphora resolution, due to certain advantages that it presents and that have already been enumerated. The low performance of the genetic algorithm could be rectified by using a better fitness function and a better encoding.

The investigation of classes of errors described in Chapter 7 opens a few directions of research as well. The treatment of plural pronouns, for example, is still an open problem, and the development of a method to tackle the cases of number disagreement would certainly prove beneficial to anaphora resolution.

Appendix A

Extract from the annotated corpus

A.1 MUC annotation

(A.1) **Raw text:**

Linux Access HOWTO Michael De La Rue, access-howto@ed.ac.uk v2.11, 28 March 1997

The Linux Access HOWTO covers the use of adaptive technology with Linux, in particular, using adaptive technology to make Linux accessible to those who could not use it otherwise. It also covers areas where Linux can be used within more general adaptive technology solutions.

1. Introduction

The aim of this document is to serve as an introduction to the technologies which are available to make Linux usable by people who, through some disability would otherwise have problems with it. In other words the target groups of the technologies are the blind, the partially sighted, deaf and the physically disabled. As any other technologies or pieces of information are discovered they will be added."

(A.2) **Annotated text:**

<TITLE><COREF ID="0"><COREF ID="1"><COREF ID="2">Linux</COREF>
Access</COREF> HOWTO</COREF> <COREF ID="3">Michael De La
Rue</COREF>, access-howto@ed.ac.uk v2.11, 28 March 1997 </TITLE>
<COREF ID="4" TYPE="IDENT" REF="0">The Linux Access HOWTO
</COREF> covers <COREF ID="5">the use of <COREF ID="6">adaptive
technology </COREF> </COREF> with <COREF ID="7" TYPE="IDENT"
REF="2">Linux</COREF>, in particular, using <COREF ID="8"
TYPE="IDENT" REF="6">adaptive technology</COREF> to make <COREF
ID="9" TYPE="IDENT" REF="2">Linux</COREF> accessible to <COREF
ID="10">those who could not use <COREF ID="11" TYPE="IDENT"
REF="2">it</COREF> </COREF> otherwise. <COREF ID="12" TYPE="IDENT"
REF="0">It</COREF> also covers <COREF ID="13">areas</COREF> where
<COREF ID="14" TYPE="IDENT" REF="2">Linux</COREF> can be used
within <COREF ID="15">more general <COREF ID="16" TYPE="IDENT"
REF="6">adaptive technology</COREF> solutions</COREF>.
<TITLE>1. <COREF ID="17">Introduction</COREF></TITLE>
<COREF ID="18">The aim of <COREF ID="19" TYPE="IDENT" REF="0">this
document</COREF> </COREF> is to serve as <COREF ID="20">an
introduction to <COREF ID="21">the technologies which are available
to make <COREF ID="22" TYPE="IDENT" REF="2">Linux</COREF> usable
by people who, through <COREF ID="24">some disability</COREF>
would otherwise have problems with <COREF ID="23" TYPE="IDENT"
REF="2">it</COREF> </COREF> </COREF>. In other words <COREF
ID="25">the target groups of <COREF ID="26" TYPE="IDENT"
REF="21">the technologies</COREF> </COREF> are <COREF ID="27">
<COREF ID="36">the blind</COREF>, <COREF ID="28">the partially
sighted</COREF>, <COREF ID="29">deaf</COREF> and <COREF ID="30">the
physically disabled</COREF></COREF>. As <COREF ID="31">any other
technologies or pieces of information</COREF> are discovered <COREF
ID="34" TYPE="IDENT" REF="31">they</COREF> will be added.

A.2 ELDA-based annotation

(A.3) **Raw text:**

FROM THE EARTH TO THE MOON

CHAPTER I

THE GUN CLUB

During the War of the Rebellion, a new and influential club was established in the city of Baltimore in the State of Maryland. It is well known with what energy the taste for military matters became developed among that nation of ship-owners, shopkeepers, and mechanics. Simple tradesmen jumped their counters to become extemporized captains, colonels, and generals, without having ever passed the School of Instruction at West Point; nevertheless; they quickly rivaled their compeers of the old continent, and, like them, carried off victories by dint of lavish expenditure in ammunition, money, and men."

(A.4) **Annotated text:**

<TITLE><P> <W ID="0">FROM</W> <EXP ID="0"> <W ID="0">THE</W>
<W ID="0">EARTH</W> </EXP> <W ID="0">TO</W> <EXP ID="1"> <W
ID="0">THE</W> <W ID="0">MOON</W> </EXP> </P></TITLE>
<P><TITLE><W ID="0">CHAPTER</W> <W ID="0">I</W><TITLE></P>
<P> <EXP ID="2"> <W ID="0">THE</W> <W ID="0">GUN</W> <W
ID="0">CLUB</W> </EXP> </P> <P> <W ID="0">During</W> <EXP
ID="3"> <W ID="0">the</W> <W ID="0">War</W> <W ID="0">of</W> <W
ID="0">the</W> <W ID="0">Rebellion</W> </EXP> <W ID="0">,</W> <EXP
ID="4"> <W ID="0">a</W> <W ID="0">new</W> <W ID="0">and</W> <W
ID="0">influential</W> <W ID="0">club</W> </EXP> <W ID="0">was</W>
<W ID="0">established</W> <W ID="0">in</W> <EXP ID="5">
<W ID="0">the</W> <W ID="0">city</W> <W ID="0">of</W> <W

ID="0">Baltimore</W> </EXP> <W ID="0">in</W> <EXP ID="6">
<W ID="0">the</W> <W ID="0">State</W> <W ID="0">of</W> <W
ID="0">Maryland</W> </EXP> <W ID="0">.</W> <W ID="0">It</W>
<W ID="0">is</W> <W ID="0">well</W> <W ID="0">known</W> <W
ID="0">with</W> <W ID="0">what</W> <W ID="0">energy</W> <EXP
ID="7"> <W ID="0">the</W> <W ID="0">taste</W> <W ID="0">for</W> <W
ID="0">military</W> <W ID="0">matters</W> </EXP> <W ID="0">became</W>
<W ID="0">developed</W> <W ID="0">among</W> <EXP ID="9">
<W ID="0">that</W> <W ID="0">nation</W> <W ID="0">of</W>
<W ID="0">ship</W> <W ID="0">-</W> <W ID="0">owners</W> <W
ID="0">,</W> <W ID="0">shopkeepers</W> <W ID="0">,</W> <W
ID="0">and</W> <W ID="0">mechanics</W> </EXP> <W ID="0">.</W>
<EXP ID="10"> <W ID="0">Simple</W> <W ID="0">tradesmen</W>
</EXP> <W ID="0">jumped</W> <EXP ID="12"> <EXP ID="11"> <REF
SRC="10"/> <W ID="0">their</W> </EXP> <W ID="0">counters</W>
</EXP> <W ID="0">to</W> <W ID="0">become</W> <EXP ID="13">
<W ID="0">extemporized</W> <W ID="0">captains</W> </EXP> <W
ID="0">,</W> <EXP ID="14"> <W ID="0">colonels</W> </EXP> <W
ID="0">,</W> <W ID="0">and</W> <EXP ID="15"> <W ID="0">generals</W>
</EXP> <W ID="0">,</W> <W ID="0">without</W> <W ID="0">having</W>
<W ID="0">ever</W> <W ID="0">passed</W> <EXP ID="16"> <W
ID="0">the</W> <W ID="0">School</W> <W ID="0">of</W> <W
ID="0">Instruction</W> <W ID="0">at</W> <W ID="0">West</W> <W
ID="0">Point</W> </EXP> <W ID="0">;</W> <W ID="0">nevertheless</W>
<W ID="0">;</W> <EXP ID="17"> <REF SRC="10"/> <W ID="0">they</W>
</EXP> <W ID="0">quickly</W> <W ID="0">rivaled</W> <EXP ID="20">

<EXP ID="21"> <REF SRC="10"/> <W ID="0">their</W> </EXP> <W ID="0">compeers</W> </EXP> <W ID="0">of</W> <EXP ID="19"> <W ID="0">the</W> <W ID="0">old</W> <W ID="0">continent</W> </EXP> <W ID="0">,</W> <W ID="0">and</W> <W ID="0">,</W> <W ID="0">like</W> <EXP ID="22"> <REF SRC="20"/> <W ID="0">them</W> </EXP> <W ID="0">,</W> <W ID="0">carried</W> <W ID="0">off</W> <EXP ID="23"> <W ID="0">victories</W> </EXP> <W ID="0">by</W> <W ID="0">dint</W> <W ID="0">of</W> <EXP ID="24"> <W ID="0">lavish</W> <W ID="0">expenditure</W> <W ID="0">in</W> <W ID="0">ammunition</W> <W ID="0">,</W> <W ID="0">money</W> <W ID="0">,</W> <W ID="0">and</W> <W ID="0">men</W> </EXP> <W ID="0">.</W> </P>

Appendix B

Output of the system

As mentioned in Chapter 5, ARMAL can display its results in three ways: text-only, XML-encoded text and graphical output.

B.1 Text-only output

In this format, the input text is replicated in the output, with each pronoun followed by its antecedent, in brackets. B.1 is an example of input text that has been processed by ARMAL and has produced the output in B.2 (the mark-up introduced by ARMAL is highlighted).

(B.1) *Input*

```
Linux has the vast advantage over Windows that most of its software
is command line oriented .  This is now changing and almost
everything is now available with a graphical front end .  However,
because it is in origin a programmers operating system, line
oriented programs are still being written covering almost all new
areas of interest .  For the physically disabled, this means that
it is easy to build custom programs to suit their needs.  For the
visually impaired, this should make use with a speech synthesiser
or Braille terminal easy and useful for the foreseeable future .
```

Linux’s multiple virtual consoles system make it practical to use as a multi-tasking operating system by a visually impaired person working directly through Braille .

(B.2) *Output*

Linux has the vast advantage over Windows that most of its **(Linux)** software is command line oriented. This is now changing and almost everything is now available with a graphical front end. However, because it **(Linux)**is in origin a programmers operating system, line oriented programs are still being written covering almost all new areas of interest . For the physically disabled, this means that it **(Linux)**is easy to build custom programs to suit their needs . For the visually impaired, this should make use with a speech synthesiser or Braille terminal easy and useful for the foreseeable future.

Linux’s multiple virtual consoles system make it **(Linux’s multiple virtual consoles system)**practical to use as a multi-tasking operating system by a visually impaired person working directly through Braille .

B.2 XML-encoded output

In this format, the antecedents of the pronouns are marked using the ELDA scheme described in Chapter 4. The input text in B.3 is processed to produce the output in B.4.

(B.3) *Input*

During the War of the Rebellion, a new and influential club was established in the city of Baltimore in the State of Maryland. It is well known with what energy the taste for military matters became developed among that nation of ship-owners, shopkeepers, and mechanics. Simple tradesmen jumped their counters to become extemporized captains, colonels, and generals, without having ever passed the School of Instruction at West Point; nevertheless; they quickly rivaled their compeers of the old continent, and, like them, carried off victories by dint of lavish expenditure in ammunition, money, and men.

(B.4) *Output*

During the War of the Rebellion, a new and influential club was established in the city of Baltimore in the State of Maryland. It is well known with what energy the taste for military matters became developed among that nation of ship-owners, shopkeepers, and mechanics. <EXP ID="38">Simple tradesmen</EXP> jumped **<EXP ID="39"><REF SRC="38"/>their</EXP>** counters to become extemporized captains, colonels, and generals, without having ever passed the School of Instruction at West Point; nevertheless; **<EXP ID="40"><REF SRC="38"/>they</EXP>** quickly rivaled **<EXP ID="41"><REF SRC="38"/>their</EXP>** compeers of the old continent, and, like **<EXP ID="42"><REF SRC="38"/>them</EXP>**, carried off victories by dint of lavish expenditure in ammunition, money, and men.

B.3 Graphical interface

ARMAL can also make use of the graphical interface of the evaluation workbench (see Chapter 6), thus displaying its results as show in figure B.1. Each pronoun and each antecedent are displayed in a context of 5 words. If ARMAL was run on an evaluation text (marked for coreference), the interface also displays information on whether the pronouns have been correctly solved or not. The values of the evaluation measures are also presented.

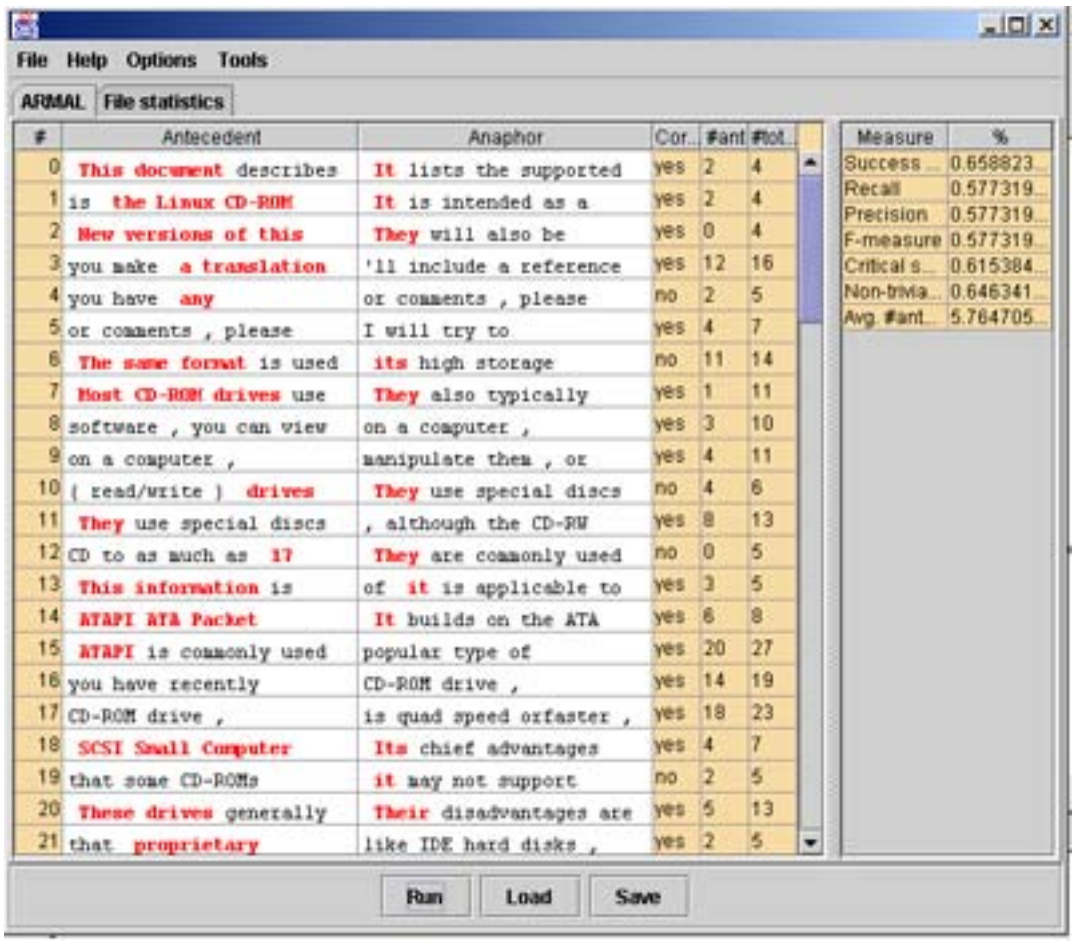

Figure B.1: Results in graphical interface

Appendix C

Previously published work

Some of the work described in this thesis has been previously published at different stages in proceedings of peer-reviewed conferences. In the following, a short description of these papers will be provided, along with a reference to the part of the thesis where they have been used or cited.

- Ruslan Mitkov and **Catalina Barbu** (2000). ”Improving pronoun resolution in two languages by means of bilingual corpora”. In Proceedings of the Discourse Anaphora and Reference Resolution Conference (DAARC2000), Lancaster, UK

  In this paper, the mutual enhancement algorithm also described in Chapter 6 is presented. The implementation of the algorithm has suffered no modifications since the publishing of this paper.

- **Catalina Barbu** and Ruslan Mitkov (2001). ”Evaluation tool for rule-based anaphora resolution methods”. In Proceedings of ACL’01, Toulouse, France

  This paper describes an early implementation of the evaluation workbench that was used in the evaluation of ARMAL. In this version, only 3 knowledge-poor approaches are used for comparative evaluation and there is no possibility

of selecting the testing set. Further improvements led to the version of the workbench described in Chapter 6.

- **Catalina Barbu** (2001). ”Automatic Learning and Resolution of Anaphora”. In Proceedings of RANLP’01, Tzigov Chark, Bulgaria

  This paper presents a preliminary implementation of ARMAL, the machine learning system described in Chapter 5. The system is only designed and evaluated for English.

- **Catalina Barbu** (2002): ”Error analysis in anaphora resolution”. In Proceedings of LREC’02, Las Palmas, Spain

  This paper discusses sources of errors in automatic anaphora resolution. The experiments and results are also presented in Chapter 7.

- **Catalina Barbu**, Richard Evans and Ruslan Mitkov (2002). ”A corpus based analysis of morphological disagreement in anaphora resolution”. In Proceedings of LREC’02, Las Palmas, Spain

  This paper investigates the types of morphological disagreement between pronouns and antecedents. The analysis is performed on a set of 2000 examples extracted mainly from the BNC. The experiments and results are presented in Chapter 7.

- **Catalina Barbu** (2002). ”Genetic algorithms in anaphora resolution”. In Proceedings of the Discourse Anaphora and Reference Resolution Conference (DAARC2002), Lisbon, Portugal.

  This paper describes the genetic algorithm that was developed as an alternative to the supervised learning method for anaphora resolution. A slightly modified

and improved version of the algorithm was described in Chapter 5 and evaluated in Chapter 6.

Bibliography

[Aone and Bennett1994] C. Aone and S.W. Bennett. 1994. Discourse tagging tool and discourse-tagged multilingual corpora. In *Proceedings of the International Workshop on Sharable Natural Language Resources*, pages 71–77, Nara, Japan.

[Aone and Bennett1995] Chinatsu Aone and Scot W. Bennett. 1995. Evaluating automated and manual acquisition of anaphora resolution rules. In *Proceedings of the 33th Annual Meeting of the Association for Computational Linguistics (ACL '95)*, pages 122–129.

[Azzam et al.1998] S. Azzam, K. Humphreys, and R. Gaizauskas. 1998. Coreference resolution in a mulitlingual information extraction system. In *Proceedings of the Linguistic Coreference Workshop, LREC'98,* pages 12–16, Granada, Spain.

[Baldwin1997] Breck Baldwin. 1997. CogNIAC: High precision coreference with limited knowledge and linguistic resources. In R. Mitkov and B. Boguraev, editors, *Operational factors in practical, robust anaphora resolution for unrestricted texts,* pages 38 – 45.

[Barbu and Mitkov2001] Catalina Barbu and Ruslan Mitkov. 2001. Evaluation tool for rule-based anaphora resolution methods. In *Proceedings of ACL'01,* pages 34–41, Toulouse, France.

[Barbu et al.2002] C. Barbu, R. Evans, and R. Mitkov. 2002. A corpus based investigation of morphological disagreement in anaphoric relations. *Proceedings of LREC 2002*, pages 1995–1999.

[Barbu2001] C. Barbu. 2001. Automatic Learning and Resolution of Anaphora. In *Proceedings of RANLP'01*, pages 22–27, Tzigov Chark, Bulgaria.

[Barbu2002] C. Barbu. 2002. Genetic algorithms for anaphora resolution. *Proceedings of DAARC 2002*, pages 7–12.

[Barlow1998] Michael Barlow. 1998. Feature mismatches in anaphora resolution. In *Proceedings of DAARC'98*, pages 34–42, Lancaster, UK.

[Biber et al.1998] Duglas Biber, Susan Conrad, and Randi Rippen. 1998. *Corpus Linguistics: Investigating Language Structure and Use*. Cambridge Approaches to Linguistics. Cambdidge University Press.

[Bolinger1977] D. Bolinger. 1977. *Meaning and form*, volume 11 of *English Language Series*. Longman, London.

[Botley1999] S. Botley. 1999. *Corpora and discourse anaphora: using corpus evidence to test theoretical claims*. Ph.D. thesis, University of Lancaster.

[Brennan et al.1987] S.E. Brennan, M.W. Friedmann, and C.J. Pollard. 1987. A Centering approach to pronouns. In *Proceedings of the 25th Annual Meeting of the ACL*, pages 155–162, Stanford.

[Bruneseaux and Romary1997] F. Bruneseaux and L. Romary. 1997. Codage des références et coréférences dans les dialogues homme-machine. In *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), pages 15–17, Ontario, Canada.

[Burnard1995] Lou Burnard. 1995. *Users Reference Guide British National Corpus Version 1.0*. Oxford University Computing Services, UK.

[Byron and Allen1999] D. Byron and J. Allen. 1999. Applying genetic algorithms to pronoun resolution. In *Proceedings of the Sixteenth National Conference on Artificial Intelligence*, pages 957–963.

[Byron2001] D. Byron. 2001. The Uncommon Denominator: A Proposal for Consistent Reporting of Pronoun Resolution Results. *Computational Linguistics*, 27(4):569–578, December.

[Cadiot1988] Pierre Cadiot. 1988. De quoi ça parle? A propos de la référence de *ça*, pronom sujet. *Le français moderne*, 56(3).

[Canning et al.2000] Y. Canning, J. Tait, J. Archibald, and R. Crawley. 2000. Replacing anaphora for readers with acquired dyslexia. In *Proceedings of DAARC 2000*, pages 83–88, Lancaster, UK.

[Carbonell and Brown1988] J.G. Carbonell and R.D. Brown. 1988. Anaphora resoluion: a multistrategy approach. In *Proceedings of the 12th International Conference in Computational Linguistics (Coling'88)*, volume I, pages 96–101, Lancaster, UK.

[Cardie and Wagstaff1999] Claire Cardie and Kiri Wagstaff. 1999. Noun phrase coreference as clustering. In *Proceedings of EMNLP/VLP*, pages 82–89, Maryland, USA.

[Carletta et al.1997] Jean Carletta, A. Isard, S. Isard, J. Kowtko, G. Doherty-Sneddon, and A. Anderson. 1997. The reliability of a dialogue structure coding scheme. *Computational Linguistics*, 23(1):13–32.

[Carletta1996] Jean Carletta. 1996. Assessing Agreement on Classification Tasks: The Kappa Statistic. *Computational Linguistics*, 22(2):249–254.

[Carreiras1997] M. Carreiras. 1997. Plural pronouns and the representation of their antecedents. *European Journal of Cognitive Psychology*, (9):53–87.

[Carter1987] D. M. Carter. 1987. *Interpreting Anaphors in Natural Language Texts*. Ellis Horwood, Chichester, UK.

[Chafe1970] Wallace L. Chafe. 1970. *Meaning and the Structure of Language*. University of Chicago Press, Chicago and London.

[Chomsky1980] Noam Chomsky. 1980. On binding. *Linguistic Inquiry*, 11(1):1–46.

[Chomsky1981] Noam Chomsky. 1981. *Lectures on government and binding*. Foris, Dordrecht.

[Chomsky1982] Noam Chomsky. 1982. *Some concepts and consequences of the theory of government*. MIT Press, Cambridge, Massachussets.

[Clifton and F.Ferreira1987] C. Clifton and F.Ferreira. 1987. Discourse structure and anaphora: some experimental results. In M. Coltheart, editor, *Attention and performance XII: The psychology of reading*, pages 635–654. Lawrence Erlbaum Associates Ltd.

[Connolly et al.1997] David Connolly, J.D. Burger, and D.S. Day. 1997. A machine learning approach to anaphoric reference. In Daniel Jones and Harold Somers, editors, *New Methods in Language Processing*, pages 255–261. UCS Press.

[Corbett1991] Greville Corbett. 1991. *Gender*. Cambridge University Press, Cambridge.

[Corblin1995] Francis Corblin. 1995. *Les formes de reprise dans le discours. Anaphores et chaines de référence*. Presses Universitaire de Rennes.

[Cristea et al.1998] D. Cristea, N. Ide, and L. Romary. 1998. Veins Theory: A Model of Global Discourse Cohesion and Coherence. In *Proceedings of ACL/Coling'98*, pages 281–285, Montreal, Canada.

[Cristea et al.1999] D. Cristea, N. Ide, D. Marcu, and V. Tablan. 1999. Discourse structure and co-reference: An empirical study. In *Proceedings of the ACL99 Workshop on the Relation between Discourse/Dialogue Structure and Reference*, pages 46–53, College Park, Maryland, USA.

[Daelemans et al.2000] Walter Daelemans, J. Zavrel, K. van der Slot, and Antal van den Bosch. 2000. TiMBL: Tilburg Memory Based Learner, version 3.0 Reference Guide. *ILK Technical Report - ILK 00-01*.

[Dagan and Itai1990] Ido Dagan and Alon Itai. 1990. Automatic processing of large corpora for the resolution of anaphora references. In *Proceedings of the 13th International Conference on Computational Linguistics (COLING'90)*, volume III, pages 1–3, Helsinki, Finland.

[Dagan and Itai1991] Ido Dagan and Alon Itai. 1991. A statistical filter for resolving pronoun references. In Y.A. Feldman and A. Bruckstein, editors, *Artificial Intelligence and Computer Vision*, pages 125 – 135. Elsevier Science Publishers B.V.

[Davies et al.1998] S. Davies, M. Poesio, F. Bruneseaux, and L. Romary. 1998. Annotating coreference in dialogues: proposal for a scheme for MATE. First draft. Technical report, http://www.hcrc.ed.ac.uk/ poesio/MATE/anno_manual.html.

[Day et al.1998] David Day, John Aberdeen, Sasha Caskey, Lynette Hirschman, Patricia Robinson, and Marc Vilain. 1998. Alembic Workbench Corpus Development Tool. In *Proceedings of the First International Conference on Language Resources and Evaluation*, pages 1021 – 1028, May.

[de Rocha1997] M. de Rocha. 1997. Supporting anaphor resolution with a corpus-based probabilistic model. In *Proceedings of the ACL'97/EACL'97 workshop on Operational factors in practical, robust anaphora resolution*, pages 54–61, Madrid, Spain.

[DeCristofaro et al.1999] J. DeCristofaro, M. Strube, and K. McCoy. 1999. Building a tool for annotating reference in discourse. In *Proceedings of the ACL99 Workshop on the Relation of Discourse/Dialogue Structure and Reference*, pages 54–62, College Park, Maryland, USA.

[Denber1998] M. Denber. 1998. Automatic Resolution of Anaphora in English. Technical report, Eastman Kodak Co, Imaging Science Division.

[Ehrlish and Rayner1983] K. Ehrlish and K. Rayner. 1983. Pronoun assignment and semantic integration during reading: Eye movements and immediacy of processing. *Journal of Verbal Learning and Verbal Behavior*, (22):75–87.

[Evans2001] R. Evans. 2001. Applying Machine Learning Toward an Automatic Classification of It. *Journal of Literary and Linguistic Computing*, 16(1):45 – 57.

[Fligelstone1992] Steve Fligelstone. 1992. Developing a scheme for annotating text to show anaphoric relations. Topics in English Linguistics, pages 152 – 170. Mouton de Gruyter.

[Gaizauskas and Humphreys2000a] R. Gaizauskas and K. Humphreys. 2000a. Quantitative evaluation of coreference algorithms in an information extraction system. In S P Botley and A M McEnery, editors, *Corpus-Based and Computational Approaches to Discourse Anaphora*, pages 143–167, Amsterdam. John Benjamins.

[Gaizauskas and Humphreys2000b] Robert Gaizauskas and Kevin Humphreys. 2000b. Quantitative evaluation of coreference algorithms in an information extraction system. In Simon Botley and Antony Mark McEnery, editors, *Corpus-based and Computational Approaches to Discourse Anaphora*, Studies in Corpus Linguistics, chapter 8, pages 145 – 169. John Benjamins Publishing Company.

[Garnham1992] A. Garnham. 1992. Minimalism versus constructionism: A false dichotomy in theories of inference during reading. *Psycoloquy*, 3(63):Reading Inference, 1.1.

[Garrod and Sanford1982] S.C. Garrod and A.J. Sanford. 1982. The mental representation of discourse in a focused memory system: Implications for the interpretation of anaphoric noun phrases. *Journal of Semantics*, 1(1):21–41.

[Ge et al.1998] Niyu Ge, J. Hale, and E. Charniak. 1998. A Statistical Approach to Anaphora Resolution. In *Proceedings of the Sixth Workshop on Very Large Corpora, COLING-ACL '98*, pages 161 – 170, Montreal, Canada.

[Gernsbacher1990] M.A. Gernsbacher. 1990. *Language comprehension as structure building*. Hillsdale, NJ: Lawrence Erlbaum Associates Ltd.

[Greffenstette1995] Gregory Greffenstette. 1995. Comparing two language identification schemes. In *Proceedings of the 3rd International Conference on the Statistical Analysis of Textual Data, JADT'95*, pages 263–268, Rome.

[Grevisse1969] M. Grevisse. 1969. *Le Bon Usage. Grammaire française avec des remarque sur la langue française d'aujourd'hui*. J.Duculot, S.A., Gembloux.

[Grosz et al.1995] B.J. Grosz, A.K. Joshi, and S. Weinstein. 1995. Centering: A framework for modelling the local coherence of discourse. *Computational Linguistics*, 21(2):203–225.

[Hale and Charniak1998] John Hale and Eugene Charniak. 1998. Getting Useful Gender Statistics from English Text. Technical Report CS-98-06.

[Halliday and Hasan1976] M. A. K. Halliday and R. Hasan. 1976. *Cohesion in English*. English Language Series. Longman Group Ltd.

[Han2001] B. Han. 2001. Building a Bilingual Dictionary with Scarce Resources: A Genetic Algorithm Approach. In *Student Research Workshop, the Second Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL-2001)*, Pittsburgh, June.

[Harabagiu and Maiorano2000] S.M. Harabagiu and S.J. Maiorano. 2000. Multilingual coreference resolution. In *Proceedings of the Language Technology Joint Conference on Applied Natural Language Processing and the North American Chapter of the Association for Computational Linguistics ANLP-NAACL2000*, pages 142–149.

[Hearst1994] Marti Hearst. 1994. Multi-paragraph segmentation of expository text. In *32nd. Annual Meeting of the Association for Computational Linguistics*, pages 9 – 16, New Mexico State University, Las Cruces, New Mexico.

[Heim1982] I. Heim. 1982. *The Semantics of Definite and Indefinite Noun Phrases*. Ph.D. thesis, University of Massachusetts at Amherst.

[Hirschman1997] L. Hirschman. 1997. MUC-7 Coreference Task Definition. http://www.muc.saic.com/proceedings/co_task.pdf.

[Hobbs1976] Jerry Hobbs. 1976. Pronoun resolution. Research report 76-1, City College, City University of New York.

[Hobbs1978] Jerry Hobbs. 1978. Pronoun resolution. *Lingua*, 44:339–352.

[Holland1975] J.H. Holland. 1975. *Adaptation in Natural and Artificial Systems*. University of Michigan Press.

[Kayne1972] R. Kayne. 1972. Subject Inversion in French Interrogatives. In J. Casagrande and B.Saciuk, editors, *Generative Studies in Romance Languages*, pages 70–126. Newbury House.

[Kazakov1997] Dimiter Kazakov. 1997. Unsupervised learning of morphology with genetic algorithms. In W. Daelemans, A. van den Bosch, and A. Weijters, editors, *Workshop Notes of the ECML/MLnet Workshop on Empirical Learning of Natural Language Processing Tasks*, pages 105–112, Prague, April.

[Kennedy and Boguraev1996] Christopher Kennedy and Branimir Boguraev. 1996. Anaphora for everyone: pronominal anaphora resolution without a parser. In *Proceedings of the 16th International Conference on Computational Linguistics (COLING'96)*, pages 113–118, Copenhagen, Denmark.

[Kibble and van Deemter2000] Rodger Kibble and Kees van Deemter. 2000. Coreference Annotation: Whither? In *Proceedings of the Second International Conference on Language Resources and Evaluation*, volume III, pages 1281 – 1286, Athens, Greece, 31 May – 2 June. ELRA.

[Krippendorff1980] K. Krippendorff. 1980. *Content Analysis: An introduction to its methodology*. Beverly Hills London: Sage Publications.

[Lamiroy1990] B. Lamiroy. 1990. Des aspects de la théorie syntaxique à la nouvelle théorie chomskyenne: rupture ou continuité. *Cahiers de l'Institut de Linguistique (U.C.L.)*, 15(1-2):88–110.

[Lamiroy1991a] B. Lamiroy. 1991a. Binding properties of the French pronoun en. In C. Georgopoulos and R. Ishihara, editors, *Interdisciplinary Approaches to Language. Essays in honour of Yuki Kuroda*, pages 417–433. Kluwer, Dordrecht.

[Lamiroy1991b] B. Lamiroy. 1991b. Coréférence et référence disjointe: les deux pronoms en. *Travaux de Linguistique*, 22:41–67.

[Lankhorst1994] M. Lankhorst. 1994. Automatic word categorization with genetic algorithms. In A. Eiben, B. Manderick, and Z. Ruttkay, editors, *Proceedings of the ECAI'94 Workshop on Applied Genetic and other Evolutionary Algorithms Amsterdam*, pages 103–107. Springer Verlag.

[Lappin and Leass1994] Shalom Lappin and H.J. Leass. 1994. An algorithm for Pronominal Anaphora Resolution. *Computational Linguistics*, 20(4):535 – 562.

[Losee1996] Robert M. Losee. 1996. Learning Syntactic Rules and Tags with Genetic Algorithms for Information Retrieval and Filtering: An Empirical Basis for Grammatical Rules. *Information Processing and Management*, 32(2):185–197.

[Manning and Schütze1999] Christopher D. Manning and Hinrich Schütze. 1999. *Foundations of statistical natural language processing*. The MIT Press.

[Marcus et al.1994] M. Marcus, G. Kim, M. Marcinkiewicz, R. MacIntyre, A. Bies, M. Ferguson, K. Katz, and B. Schasberger. 1994. The Penn Treebank: Annotating predicate argument structure.

[Matthews and Chodorow1988] A. Matthews and M.S. Chodorow. 1988. Pronoun resolution in two-clause sentences: Effects of ambiguity, antecedent location and depth of embedding. *Journal of Memory and language*, (27):245–260.

[Maynard et al.2001] D. Maynard, V. Tablan, C. Ursu, H. Cunningham, and Y. Wilks. 2001. Named Entity Recognition from Diverse Text Types. In *Proceedings of RANLP2001*, pages 159–165, Tzigov Chark, Bulgaria.

[McCarthy and W.Lehnert1995] J. McCarthy and W.Lehnert. 1995. Using decision trees for coreference resolution. In *Proceeding of IJCAI'95*, pages 1050–1055.

[McNemar1947] Q. McNemar. 1947. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, (12):153–157.

[Mehler and Dupoux1987] C. Mehler and J. Dupoux. 1987. De la psychologie à la science cognitive. *Le Débat*, (47):65–87.

[Meng et al.2001] Soon Wee Meng, Hwee Tou Ng, and Daniel Chung Yong Lim. 2001. A Machine Learning Approach to Coreference resolution of Noun Phrases. *Computational Linguistics*, 27(4):521–544, December.

[Milner1982] Jean-Claude Milner. 1982. *Ordres et raisons de langue*. Le Seuil, Paris.

[Mitchell1997] Tom M. Mitchell. 1997. *Machine learning*. McGraw-Hill.

[Mitkov and Barbu2000] R. Mitkov and C. Barbu. 2000. Improving pronoun resolution in two languages by means of bilingual corpora. In *Proceedings of the Discourse Anaphora and Reference Resolution Conference (DAARC2000)*, pages 133–137, Lancaster, UK.

[Mitkov and Stys1997] R. Mitkov and M. Stys. 1997. Robust reference resolution with limited knowledge: high precision genre-specific approach for English and Polish. In *Proceedings of RANLP'97*, pages 74–81, Tzigov Chark, Bulgaria.

[Mitkov et al.1998] Ruslan Mitkov, Lamia Belguith, and M. Stys. 1998. Multilingual Anaphora Resolution. In *The Third International Conference on Empirical Methods in Natural Language Processing*, pages 7–16, Granada, Spain.

[Mitkov et al.1999] R. Mitkov, C. Orăsan, and R. Evans. 1999. The importance of annotated corpora for NLP: the cases of anaphora resolution and clause splitting. In *Proceeding of "Corpora and NLP: Reflecting on Methodology Workshop", TALN'99*, pages 60–69.

[Mitkov et al.2000] Ruslan Mitkov, R. Evans, C. Orăsan, C. Barbu, L. Jones, and V. Sotirova. 2000. Coreference and anaphora: developing annotating tools, annotated resources and annotation strategies. In *Proceedings of the Discourse, Anaphora and Reference Resolution Conference (DAARC2000)*, pages 49 – 58, Lancaster, UK.

[Mitkov et al.2002] R. Mitkov, R. Evans, and C. Orăsan. 2002. A new, fully automatic version of Mitkov's knowledge-poor pronoun resolution method. In Al. Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing*, pages 169–187. Springer Verlag.

[Mitkov1998] Ruslan Mitkov. 1998. Robust pronoun resolution with limited knowledge. In *Proceedings of the 18th International Conference on Computational Linguistics (COLING'98/ACL'98*, pages 867 – 875. Morgan Kaufmann.

[Mitkov2000] Ruslan Mitkov. 2000. Towards more comprehensive evaluation in anaphora resolution. In *Proceedings of the Second International Conference on Language Resources and Evaluation*, volume III, pages 1309 – 1314, Athens, Greece.

[Mitkov2001] Ruslan Mitkov. 2001. Outstanding issues in anaphora resolution. In Al. Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing*, pages 110–125. Springer.

[Mitkov2002] R. Mitkov. 2002. *Anaphor resolution*. Studies in Language and Linguistics. Longman.

[Neale1990] S. Neale. 1990. *Descriptions*. MIT Press, Cambridge, Massachusets.

[Orăsan and Evans2001] C. Orăsan and R. Evans. 2001. Learning to identify animate references. *Proceedings of the Workshop on Computational Natural Language Learning (CoNLL-2001), ACL2001*, pages 129–136.

[Orăsan et al.2000] Constantin Orăsan, Richard Evans, and Ruslan Mitkov. 2000. Enhancing Preference-Based Anaphora Resolution with Genetic Algorithms. In *Proceedings of Natural Language Processing - NLP2000*, pages 185 – 195. Springer.

[Orăsan2000] Constantin Orăsan. 2000. CLinkA - A coreferential links annotator. In *Proceedings of LREC 2000*, pages 491 – 496, Athens, Greece.

[Paice and Husk1987] Chris D. Paice and G.D. Husk. 1987. Towards the automatic recognition of anaphoric features in English text: the impersonal pronoun *it*. *Computer Speech and Language*, 2:109 – 132.

[Passoneau and Litman1997] R.J. Passoneau and D.J. Litman. 1997. Discourse segmentation by human and automated means. *Computational Linguistics*, 23(1):103–139.

[Pinchon1972] J. Pinchon. 1972. *Les pronoms adverbiaux en et y*. Droz, Genève.

[Poesio and Vieira1998] Massimo Poesio and Renata Vieira. 1998. A Corpus-based Investigation of Definite Description Use. *Computational Linguistics*, 24(2):183–216.

[Popescu-Belis and Robba1997] A. Popescu-Belis and I. Robba. 1997. Cooperation between pronoun and reference resolution for unrestricted texts. In *Proceedings of ACL'97/EACL'97 Workshop "Operational Factors in Practical, Robust Anaphora Resolution for Unrestricted Texts"*, pages 94–99.

[Prince1982] E. Prince, 1982. *Discourse description: diverse analyses of a fund-raising text*, chapter The ZPG letter: subjects, definiteness, and information status, pages 295–325. John Benjamins.

[Quinlan1993] J.R. Quinlan. 1993. *C4.5: Programs for Machine Learning*. Morgan Kaufmann.

[Quirk et al.1985] Randolph Quirk, Sidney Greenbaum, Geoffrey Leech, and Jan Svartvik. 1985. *A Comprehensive Grammar of the English Language*. Longman.

[Reboul1997] A. Reboul. 1997. Le projet CERVICAL. Représentations mentales, référence aux objets et aux événements. Technical report, LORIA, Nancy.

[Roth2000] D. Roth. 2000. Learning in Natural Language: Theory and Algorithmic Approaches. In *Proceedings of CoNLL'00*, pages 1–6.

[Rouwet1990] N. Rouwet. 1990. En et y - deux clitiques pronominaux antilogophoriques. *Langages*, (97):51–81.

[Sanford and Lockhart1990] A.J. Sanford and F. Lockhart. 1990. Description types and methods of conjoining as factors influencing plural anaphora: A continuation study of focus. *Journal of Semantics*, (7):365–378.

[Sanford1965] S. Sanford. 1965. *The Phonological and Morphological Structure of French*. Ph.D. thesis, MIT.

[Sidner1979] C.L. Sidner. 1979. *Towards a computational theory of definite anaphora comprehension in English discourse*. PhD Thesis, MIT.

[Sidner1983] C.L. Sidner. 1983. Focusing in the interpretation of definite anaphora. In M.Brady and R.C.Berwick, editors, *Computational Models of Discourse*, pages 267–330. MIT Press.

[Siegel and Castellan1988] S. Siegel and N.J. Castellan. 1988. *Nonparametric statistics for the Behavioral Sciences*. McGraw-Hill, 2nd edition.

[Siegel1956] S. Siegel. 1956. *Non-parametric statistics for the behavioral sciences*. McGraw-Hill, NY.

[Sparck-Jones and Galliers1996] Karen Sparck-Jones and Julia R. Galliers. 1996. *Evaluating natural language processing systems: an analysis and review*. Number 1083 in Lecture Notes in Artificial Intelligence. Springer.

[Tanev and Mitkov2000] H. Tanev and R. Mitkov. 2000. LINGUA - a robust architecture for text processing and anaphora resolution in Bulgarian. In *Proceedings of MT2000*, pages 20.1–20.8, Exeter, UK.

[Tapanainen and Järvinen1997] P. Tapanainen and T. Järvinen. 1997. A Non-Projective Dependency Parser. In *Proceedings of the 5th Conference of Applied Natural Language Processing*, pages 64 – 71, Washington D.C., USA.

[Tetreault1999] Joel R. Tetreault. 1999. Analysis of Syntax-Based Pronoun Resolution Methods. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL '99)*, pages 602 – 605, Maryland, USA.

[Tetreault2001] J. Tetreault. 2001. A Corpus-based Evaluation of Centering and Pronoun Resolution. *Computational Linguistics*, 27(4):507–520, December.

[Tetreault2002] Joel R. Tetreault. 2002. Clausal Structure and Pronoun Resolution. In *Proceedings of DAARC'02*, pages 217 – 220, Lisbon, Portugal.

[Togeby1982] K. Togeby. 1982. *Grammaire française*. Akademisk Forlag, Copenhague.

[Trouilleux2001] F. Trouilleux. 2001. *Identification des reprises et interprétation automatique des expressions pronominales dans des textes en français*. Ph.D. thesis, Université Blaise Pascal - Clermont Ferrand.

[Tutin et al.2000] Agnés Tutin, François Trouilleux, Catherine Clouzot, Éric Gaussier, Annie Zaenen, Stéphanie Rayot, and Georges Antoniadis. 2000. Annotating a Large Corpus with Anaphoric Links. In *Proceedings of Discourse, Anaphora and Reference Resolution Conference, DAARC-2000*, pages 28 – 38, Lancaster, UK.

[Vieira and Poesio1999] Renata Vieira and Massimo Poesio. 1999. Processing definite descriptions in corpora. In S. Botley and A.M. McEnery, editors, *Corpus-based and Computational Approaches to Discourse Anaphora*, Studies in Corpus Linguistics, chapter 10, pages 189 – 212. John Benjamins Publishing Company.

[Vieira and Poesio2000] Renata Vieira and Massimo Poesio. 2000. An Empirically-Based System for Processing Definite Descriptions. *Computational Linguistics*, 26(4):525–579.

[Vieira1998] Renata Vieira. 1998. *Definite description processing in unrestricted text*. PhD Thesis, University of Edinburgh.

[Walker and Moore1997] Marilyn A. Walker and Johanna D. Moore. 1997. Empirical Studies in Discourse. *Computational Linguistics*, 23(1):1–12.

[Walker1989] Marilyn Walker. 1989. Evaluating discourse processing algorithms. In *Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics*, pages 251–261, Vancouver, Canada.

[Wilks1973] Y. Wilks. 1973. Preference Semantics. Stanford AI Laboratory Memo AIM-206, Stanford University.

[Winograd1972] T. Winograd. 1972. Understanding natural language. *Cognitive psychology*, 3:1–191.

[Yang1993] J.-J. Yang. 1993. *Use of Genetic Algorithms for Query Improvement in Information Retrieval Based on a Vector Space Model*. Ph.D. thesis, University of Pittsburgh, Pittsburgh, PA.

[Zhang1992] J. Zhang. 1992. Selecting typical instances in instance-based learning. In *Proceedings of the International Machine Learning Conference 1992*, pages 470–479.